From simple factors to artificial intelligence: evolution of prognosis prediction in childhood cancer: a systematic review and meta-analysis

Petra Varga,[a,b] Mahmoud Obeidat,[a] Vanda Máté,[a,c] Tamás Kói,[a,d] Szilvia Kiss-Dala,[a] Gréta Szilvia Major,[a,b] Ágnes Eszter Tímár,[a,b] Ximeng Li,[a] Ádám Szilágyi,[a] Zsófia Csáki,[a] Marie Anne Engh,[a] Miklós Garami,[a,c] Péter Hegyi,[a,e,f] Ibolya Túri,[a,g] and Eszter Tuboly[a,h,*]

[a]Centre for Translational Medicine, Semmelweis University, Budapest, Hungary
[b]Heim Pál National Pediatric Institute, Budapest, Hungary
[c]Pediatric Center, Semmelweis University, Budapest, Hungary
[d]Department of Stochastics, Institute of Mathematics, Budapest University of Technology and Economics, Budapest, Hungary
[e]Institute of Pancreatic Diseases, Semmelweis University, Budapest, Hungary
[f]Institute for Translational Medicine, Medical School, University of Pécs, Pécs, Hungary
[g]Pető András Faculty, Semmelweis University, Budapest, Hungary
[h]Hungarian Pediatric Oncology Network, Budapest, Hungary

eClinicalMedicine
2024;78: 102902

Published Online xxx
https://doi.org/10.1016/j.eclinm.2024.102902

Summary

Background Current paediatric cancer care requires innovative approaches to predict prognosis that facilitates personalised stratification, yet studies on the performance, composition and limitations of contemporary prognostic models are lacking. We aimed to compare the accuracy of traditional and advanced prognostic models.

Methods A systematic search for this systematic review and meta-analysis (CRTN42022370251) was conducted in PubMed, Embase, Scopus, and the Cochrane Library databases on 28 June 2024. Studies on the accuracy of prognostic markers or models used in paediatric haematological malignancies, central nervous system (CNS), or non-CNS solid tumours (NCNSST) were included. Three model categories were defined using: 1-clinical parameters, 2-genomic-transcriptomic data, and 3-artificial intelligence (AI). Primary outcomes were area under the receiver operating characteristic curve with a 95% confidence interval (CI) for various overall survival intervals and event-free survival. Two independent groups performed selection and data extraction. We used data published by the authors and publicly available databases.

Findings Of 12,982 studies, 358 were included in the meta-analysis and 27 in the systematic review, with limited data on AI-approaches. Most data were reported on NCNSST at 5-year OS, where a statistically significant difference was observed between Category-1 (0.75 CI: 0.72–0.79) and Category-2 (0.85 CI: 0.82–0.88) ($p < 0.001$), but not between Categories-2 and -3 ($p = 0.2834$) (0.82 CI: 0.77–0.88). Internal validation studies showed significantly better performance compared to those using external validation, highlighting the high risk of bias (ROB) inherent in internal validation. High ROB was most commonly experienced in the outcomes and statistical analysis domains, assessed using PROBAST and QUIPS.

Interpretation It is advisable to introduce Category-2 and -3 models in a clinical setting, especially for NCNSST prognostic for aiding risk-stratification. Although AI-supported predictions in paediatric oncology are at an early stage of development, it is imperative to further explore their potential. This requires structured data collection and ethical sharing from paediatric oncology patients in sufficient quantity and quality.

Funding None.

Abbreviations: AI, artificial intelligence; AUROC, area under the receiver operating characteristic curve; C, composite (validation); C-index, concordance index; CI, confidence interval; CNS, central nervous system; CNST, central nervous system tumours; CSS, cancer-specific survival; E, external (validation); EFS, event-free survival; I, internal (validation); HM, haematological malignancies; ML, machine learning; NB, neuroblastoma; NCNSST, non-CNS solid tumours; NGS, next-generation sequencing; OS, overall survival; pAUROC, pooled area under the receiver operating characteristics curve; PC, paediatric cancer; pC-index, pooled concordance index; PICO, patients, intervention, control, outcomes; PO, paediatric oncology; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses; PROBAST, Prediction model Risk Of Bias Assessment Tool; QUIPS, Quality In Prognosis Studies; RoB, risk of bias; SD, standard deviation; SE, standard error; SEER, Surveillance, Epidemiology, and End Results; TARGET, Therapeutically Applicable Research to Generate Effective Treatments

*Corresponding author. 1088-Budapest, József körút 69. fszt. 1, Hungary.
E-mail address: tuboly.eszter@gyerekklinika.com (E. Tuboly).

Keywords: Paediatric oncology; Prognosis prediction; Prognostic model; Next-generation sequencing; Artificial intelligence; Public database; Biobank data

Research in context

Evidence before this study

There is no available evidence on the accuracy of various prognostic approaches for predicting paediatric cancer outcomes. Our aim was to provide a comprehensive review of the available literature on predicting a wide range of outcomes of paediatric and adolescent (<21 years) patients with primary malignant tumours, using models classified into three categories (1-clinical data-based; 2-robust genomic-transcriptomic; 3-artificial intelligence-based models) assessed with direct statistics on the accuracy of the models. A systematic search of four databases (PubMed, Embase, Scopus, and Cochrane Library) was conducted on the 22nd of November 2022, and references of eligible full texts were checked as well, without language limitations. The search was updated on the 28th of June 2024. The search key consisted of four domains: paediatric, cancer, prognosis, and accuracy. The risk of bias was assessed using two tools: the Quality in Prognosis Studies (QUIPS) and the Prediction model Risk Of Bias Assessment Tool (PROBAST).

Added value of this study

Our extensive analysis of different outcomes and subgroups highlights the superior predictive power of genomic-transcriptomic and artificial intelligence-based models. Subgroup analysis showed that the method of validation can lead to significant differences in accuracy, and internal validation can lead to better but biased results.

Implications of all the available evidence

In clinical settings, more modern, genomic-transcriptomic data-based models are recommended. Artificial intelligence is a promising approach in model development, and external validation should be chosen to obtain objective and generalizable models.

Introduction

Paediatric cancers (PC) are characterised by diverse biological behaviour, aetiology, and clinical trajectories.[1–3] In high-income countries, more than 80% of patients can be cured[4]; however, survival rates drop following unfavourable events (e.g., 0.58 5-year survival rate for acute lymphoblastic leukaemia after relapse in Nordic countries),[5] which requires differentiation between patients with various clinical courses.[6] The scarcity of paediatric cohorts limits the availability of scientific reports and high-quality data for valid clinical decisions. Approximately 400,000 new PC cases are diagnosed worldwide each year.[7]

PC can be divided into three main categories by the origin of the tumour, which are characterised by very different therapies, prognostic tools, and outcomes. The most common are haematological malignancies (HM) and central nervous system (CNS) tumours (CNST). Non-CNS solid tumours (NCNSST) consist of a variety of rare histological types. The age-standardized incidence rates for more common leukaemias and brain tumours worldwide are 5.41 and 2.04 per 100,000, respectively, in sharp contrast with, e.g., 0.19 per 100,000 of liver cancer.[8] Despite being rare, NCNSSTs are relevant, as – based on a study from the United States written in 2020- the 5-year OS of, for example, bone tumours, soft tissue sarcomas, and hepatoblastoma is less than 70% in the US.[9]

Despite recent breakthroughs in therapeutic approaches, tailoring treatments to patients individually, and high-risk case management remain a challenge, raising awareness of clinical decision support. Often, burdensome multiple-arm therapeutic protocols are required, but with poor upfront risk stratification, some patients may receive excessive therapies, others may face disease progression.[10,11]

Over the last decade, prognostic tools in paediatric oncology (PO) have evolved from early models using observable traits[12–14] to integrating comprehensive molecular profiling via next-generation sequencing (NGS),[15–17] improving accurate risk stratification.[18] Artificial intelligence (AI)-driven prognostic approaches show promise in capturing the complex interactions between various data modalities and enabling adequate data integration.[19,20]

Despite recent milestones, without prior meta-analyses, objective evidence on the performance of these models in PO is lacking. The aim of this systematic review and meta-analysis was to provide a comprehensive overview of the evolution of prognostic modelling in PC, revealing novel insights into the construction of accurate prognostic models and their clinical impact, facilitating the development of more accurate and robust prognostic tools.

Methods

Our study followed the recommendations of the Preferred Reporting Items for Systematic Reviews and

Meta-Analyses (PRISMA) 2020 guidelines,[21] and the Cochrane Handbook.[22] The protocol was registered on PROSPERO (CRTN42022370251). After a systematic search, we deviated slightly from the protocol, extending our analysis to include additional outcomes beyond those initially defined.[23]

Eligibility criteria

To formulate our clinical question, we utilised the PICO framework (patients, intervention, control, outcomes).[24] Studies with PC patients were included, and various prognostic prediction approaches were considered as 'intervention' and 'control'. The area under the receiver operating characteristic curve (AUROC) of 1- (short-term), 2-, 3- (mid-term), 5-year (long-term) OS and event-free survival (EFS), 10-year OS, non-time dependent OS and EFS prediction were the primary outcomes. The concordance index (C-index) of OS, EFS, and cancer-specific survival (CSS) predictions were selected as secondary outcomes.

Peer-reviewed studies were included if patients were younger than 21 years (more than half of them/mean/median age) and were diagnosed with a primary malignant tumour. The AUROC or C-index had to be presented on the accuracy of a prognostic model or factor, with additional statistical data (standard error (SE), standard deviation (SD), 95% confidence interval (CI), sensitivity, specificity, positive/negative predictive value, true/false positive/negative cases, number of dead/alive or event/no event patients) either numerically or as figures. If a study analysed training and validation sets, it was included if statistical data were provided in the validation set. For more details see Supplementary Methods S3.

Information sources

Our systematic search was conducted in four main databases: Embase, MEDLINE (via PubMed), Cochrane Central Register of Controlled Trials (CENTRAL), and Scopus, on 22 November 2022. In addition, a backward citation search was performed using a reference-checking tool[25] on 7 June 2023 to identify all potential references of the originally included articles that met our eligibility criteria. We have updated our search on 28 June 2024 to find studies published after the original date of the search. No language restrictions were applied.

Search strategy

Our search key included four main domains: paediatric, cancer, prognosis, and accuracy (see the whole search key in Supplementary Table S1).

Data extraction

Relevant data from eligible studies were extracted independently by two groups (P.V. and T.K.+Sz.K.D.). Disagreements were resolved by the corresponding author (E.T.). All data were manually collected and entered into an Excel spreadsheet (Office365, Microsoft, Redmond, WA, USA). Prognostic factors and models were divided into distinct categories. In Category-1, models rely on conventional clinical and genetic factors, whereas those in Category-2 utilise NGS. Category-3 models can use both clinical and NGS-based genomic-transcriptomic factors but during the model building, machine-learning (ML) and AI (Supplementary Fig. S1) methods are implemented for choosing the most appropriate set of factors. In order to obtain a comprehensive overview of the evolution of childhood cancer prognostics, we added Category-0: factors with weak prognostic power present in certain studies as a comparison to the models developed by the authors without the intention of using these factors for prognosis prediction. Screening, selection, and data items are explained in Supplementary Methods S1.

Study risk of bias assessment

The risk of bias assessment (ROB) was independently conducted by two groups (P.V. and Á.T.+G.M.). For studies focusing on a single prognostic factor, the Quality in Prognosis Studies (QUIPS) tool was employed,[26] whereas for studies analysing complex prognostic models using several factors combined, the Prediction model Risk Of Bias Assessment Tool (PROBAST) tool was chosen.[27] In case of disagreement, consensus was reached after discussion with the corresponding author (E.T.).

Synthesis methods

Statistical analyses were performed using R statistical software (version 4.1.2.).[28] A p-value of less than 0.05 was considered significant for all statistical analyses. Due to the large number of performed statistical tests, false significant results can be present in the manuscript. The p-value has an important role in this respect: the smaller the significant p-value the less likely that the finding is false. We separately analysed the AUROC values of the predictions corresponding to different time points. We estimated the SDs of the AUROC values using the CIs. If no CI was available, based on published KM curves and scatter plots, we estimated the number of patients with and without the investigated event and using the formula published by Hanley and McNeil.[29] The results were visualised in forest plots.

We also meta-analysed C-index statistics analogously to AUROC, using the advice of Debray et al.[30] We assessed publication bias by creating funnel plots. Due to the complexity of the data the conventional heterogeneity analysis is not appropriate. Nevertheless, to get a glimpse into the heterogeneity, we calculated conventional heterogeneity statistics in a few cases. See details in the Supplementary Methods S4.

Studies were classified primarily by the disease of interest (HM, CNST, NCNSST) and the category of the

model (Categories-1-2-3). To address confounding bias caused by certain study characteristics, we sorted the articles into subgroups, considering validation type (internal/external/composite) and whether the proposed model included the success or failure of treatment as a prognostic factor (yes/no). To assess whether a specific dataset could provide advantage in model construction, we compared commonly used, publicly available training and validation dataset pairs containing NB[31–33] or osteosarcoma[31,34–37] patients with each other (within tumour type) and with other datasets (patients of the authors or less common public databases).

Role of the funding source

There was no funding source for this study. The corresponding author, M.O., T.K., Sz.K.D. and V.P. had access to all the data and had responsibility for the decision to submit the study for publication.

Results

Altogether, 10,870 applicable studies were identified by our systematic search of four databases, an additional 7986 were found eligible among the references and 1252 during the updated search. In total 385 articles were included, 358, 92 of which was included during the updated search, in the meta-analysis and 27 additional ones in the systematic search (PRISMA Flowchart Fig. 1).

Most (379) included studies were retrospective cohort studies and we had 5 eligible prospective cohort studies and 1 cross-sectional study, covering the period from 1991 to 2024. Studies from various regions were included, with the largest proportion (73.25%) from China. Of the included studies, most (272 studies) aimed to predict the prognosis of NCNSST, whereas only 81 and 32 focused on HM and CNST, respectively. In the articles included in the meta-analysis, the predominant prediction approach was Category-2 (168 studies), followed by Category-1 (145 studies) and Category-3 with significantly fewer, only 45 studies. The ratio of the included studies during the updated search was similar to the original one, both in the aspect of model categories and tumour types. The basic characteristics of the included studies are presented in Supplementary Table S2.

As for NCNSST, consistent pooled AUROC (pAUROC) values characterized 1-year OS predictions across all model categories: 0.8 (CI: 0.75–0.85) in Category-1, 0.85 (CI: 0.80–0.91) in Category-2 and 0.81 (CI: 0.74–0.88) in Category-3. No significant differences were shown between any of the categories (Category-1 VS -2 p = 0.169; Category-1 VS -3 p = 0.831; Category-2 VS -3 p = 0.245). At the 2-year mark, Category-2 models demonstrated robust performance with a pAUROC of 0.80 (CI: 0.72–0.89), whereas Category-3 models showed a decrease to 0.76 (CI: 0.64–0.88) but the difference was not significant (p = 0.659).

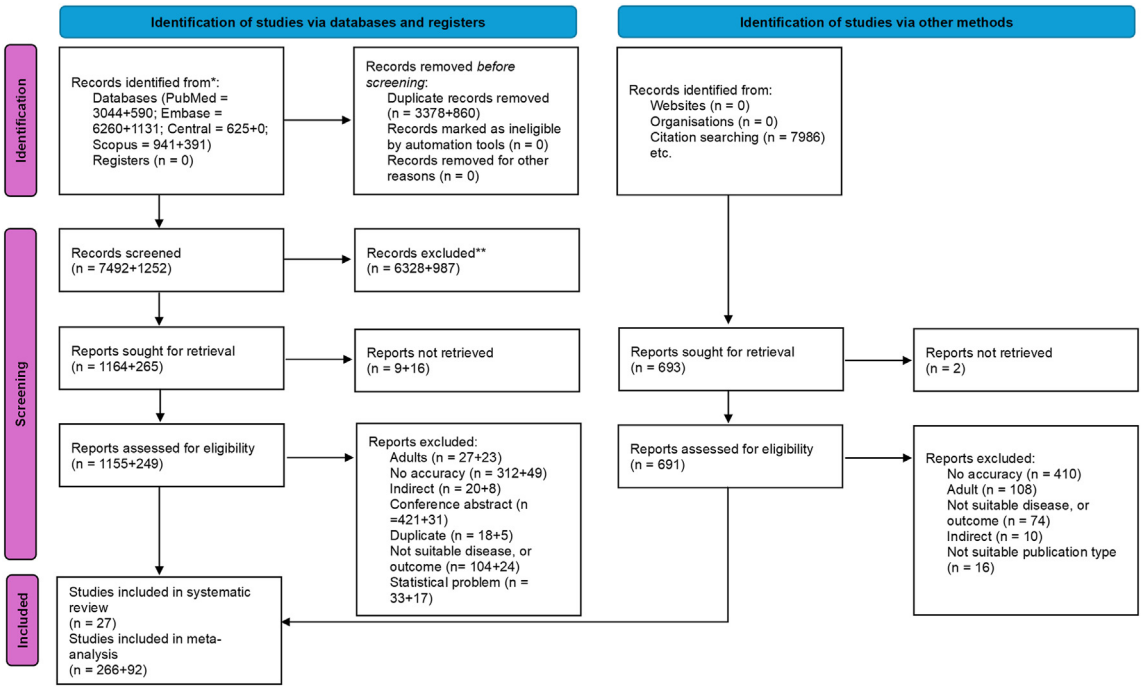

Fig. 1: PRISMA flowchart of the article selection process.

After 3 years, Category-1 decreased to 0.77 (CI: 0.73–0.80), whereas Category-2 maintained its high pAUROC of 0.84 (CI: 0.81–0.86), while Category-3 presented a pAUROC of 0.77 (CI: 0.73–0.80), respectively. The difference was statistically significant only between Categories- 1 and -2, favouring the latter (p = 0.035) (Category-1 VS -3 p = 0.552; Category-2 VS -3 p = 0.061).

For 5-year OS predictions, which can be considered a gold standard in PC prognosis prediction and was therefore chosen as a key Forest plot that we included in the main text (Fig. 2), a good performance was manifested in Category-2 with pAUROC values of 0.85 (CI: 0.82–0.88) and 0.82 (CI: 0.77–0.88) in Category-3, while Category-1 showed a decreased value, 0.75 (CI: 0.72–0.79). Similarly to 3-year OS, in the case of 5-year OS, we could show a significant difference between Category-1 and -2, also favouring the latter (p < 0.001). Considering this outcome, the performance of Category-3 models was nearly significantly better than Category-1 models (p = 0.0511). Category-2 models still showed remarkable efficacy over the 10-year horizon (pAUROC: 0.79 (CI: 0.74–0.85)).

The assessment of prediction model efficacy for HM was hindered by limited data. For 1-year OS, sufficient data could only be pooled from Category-2 articles, with a pAUROC value of 0.71 (CI: 0.63–0.79). In addition, predictions for 3-year OS revealed comparable performance between Category-1 and Category-2 models, with corresponding pAUROC values of 0.73 (CI: 0.69–0.77) and 0.75 (CI: 0.67–0.82) (Category-1 VS -2 p = 0.824). At the 5-year mark (Fig. 2), a pAUROC value for Category-1 was 0.76 (CI: 0.69–0.83), and for Category-2 0.74 (CI: 0.71–0.77) (Category-1 VS -2, p = 0.562).

For CNST, only Category-2 provided sufficient data to predict 1-year OS, yielding a high pAUROC of 0.8 (CI: 0.71–0.88). Both Category-1 and -2 predicting 3- and 5-year OS could be analysed, with similar pAUROC values predicting both outcomes, Category-1 yielded 0.74 (CI: 0.71–0.77) and 0.75 (CI: 0.72–0.78), while Category-2 presented worse pAUROC values, 0.69 (CI: 0.63–0.76) and 0.64 (CI: 0.63–0.65) (Category-1 VS -2, 3-year OS p = 0.282; the difference was significant in the case of 5-year OS, p = 0.0141) (Fig. 3A).

A statistically significant difference was also shown between predicting 1-year OS with Category-2 models in NCNSST and HM, favouring NCNSST (NCNSST VS HM p = 0.046). Category-2 models could predict 5-year OS significantly better in NCNSST than CNST (NCNSST VS HM p = 0.003).

In the subgroup analysis, in addition to model types, groups were further subdivided according to validation type (I-internal/E-external/C-composite) and treatment outcome as a prognostic factor (yes/no). For CNST, we could compare 2-no-I and 2-no-E subgroups but we found no significant differences (1-year OS, p = 0.691; 3-year OS, p = 0.694; 5-year OS, p = 0.556) (Fig. 4A), whereas for HM, no significant differences were detected between the subgroups, based on neither model nor validation type (3-year OS: 1-no-I VS 2-no-I, p = 0.501, 2-no-I VS 2-no-E, p = 0.520; 5-year OS: 1-no-I VS 2-no-I, p = 0.995, 2-no-I VS 2-no-E, p = 0.448) (Fig. 4B).

Considering NCNSST, for 10-year OS, we had insufficient data for meaningful comparisons. However, a statistically significant difference was found when predicting 3- and 5-year OS between the 2-no-I and 1-no-I groups (p < 0.001 for both outcomes), with the former performing better. A similar difference was observed between the 2-no-I and –C groups in predicting 3-year OS (p = 0.0074). Regarding the prediction of 3- and 5-year OS (Fig. 5), we found similar differences between 2-no-I and -E (p < 0.001 for both outcomes).

We have found no differences between 1-yes-I and 1-no-I subgroups that differed based on the inclusion of therapy details as prognostic factors (3-year OS, p = 0.479; 5-year OS, p = 0.276). We found statistically significant difference in the prediction of 2-year OS regarding both model categories and validation types (2-no-E VS 3-no-E, p = 0.031; 2-no-E VS 2-no-I, p < 0.001). No significant difference was observed between Categories-2 and -3 in the rest of the subgroups (1-year OS: 2-no-I VS 3-no-I, p = 0.983; 3-year OS: 2-no-E VS 3-no-E, p = 0.126; 2-no-I VS 3-no-I, p = 0.617; 5-year OS: 2-no-E VS 3-no-E, p = 0.2997; 2-no-I VS 3-no-I, p = 0.353) (Fig. 3B).

Amongst NCNSST, data for 1-year EFS Category-2 models could be pooled, resulting in a pAUROC of 0.70 (CI: 0.60–0.80), while for 3-year EFS, enough data was available for both Category-1 and -2 models yielding pAUROCs of 0.69 (CI: 0.66–0.73) and 0.76 (CI: 0.72–0.80), respectively (no statistically significant difference, p = 0.065. For 5-year EFS, both Category-1 and Category-2 models showed similar performance, with pAUROC values of 0.74 (CI: 0.68–0.80) and 0.77 (CI: 0.74–0.81), respectively (no statistically significant difference, p = 0.32). In terms of HM, a pAUROC of 0.73 (CI: 0.68–0.79) for Category-2 models predicting 1-year EFS, 0.67 (CI: 0.52–0.82) and 0.77 (CI: 0.71–0.82) was observed for Category-1 and -2 models (Category-1 VS -2 difference was not significant, p = 0.302), respectively, predicting 2-year EFS and 0.77 (CI: 0.66–0.88) for Category-1 predicting 5-year EFS (Fig. 6A).

We had markedly less data available for each subgroup for these outcomes and were unable to make any meaningful statistical comparisons. There is a slight difference in 2-no-E subgroup results in NCNSST patients, as the pAUROC value of 1-year EFS prediction (0.74 (CI: 0.54–0.95)) is smaller than 3- and 5-year EFS (0.77 (CI: 0.69–0.85) and 0.77 (CI: 0.73–0.81)) (Fig. 6B).

In terms of non-time dependent OS of NCNSST, Category-3 models emerged as frontrunners with a pAUROC value of 0.85 (CI: 0.83–0.87), whereas Category-1 and 2 performed similarly, resulting in pAUROC values of 0.76 (CI: 0.71–0.81) and 0.78 (CI:

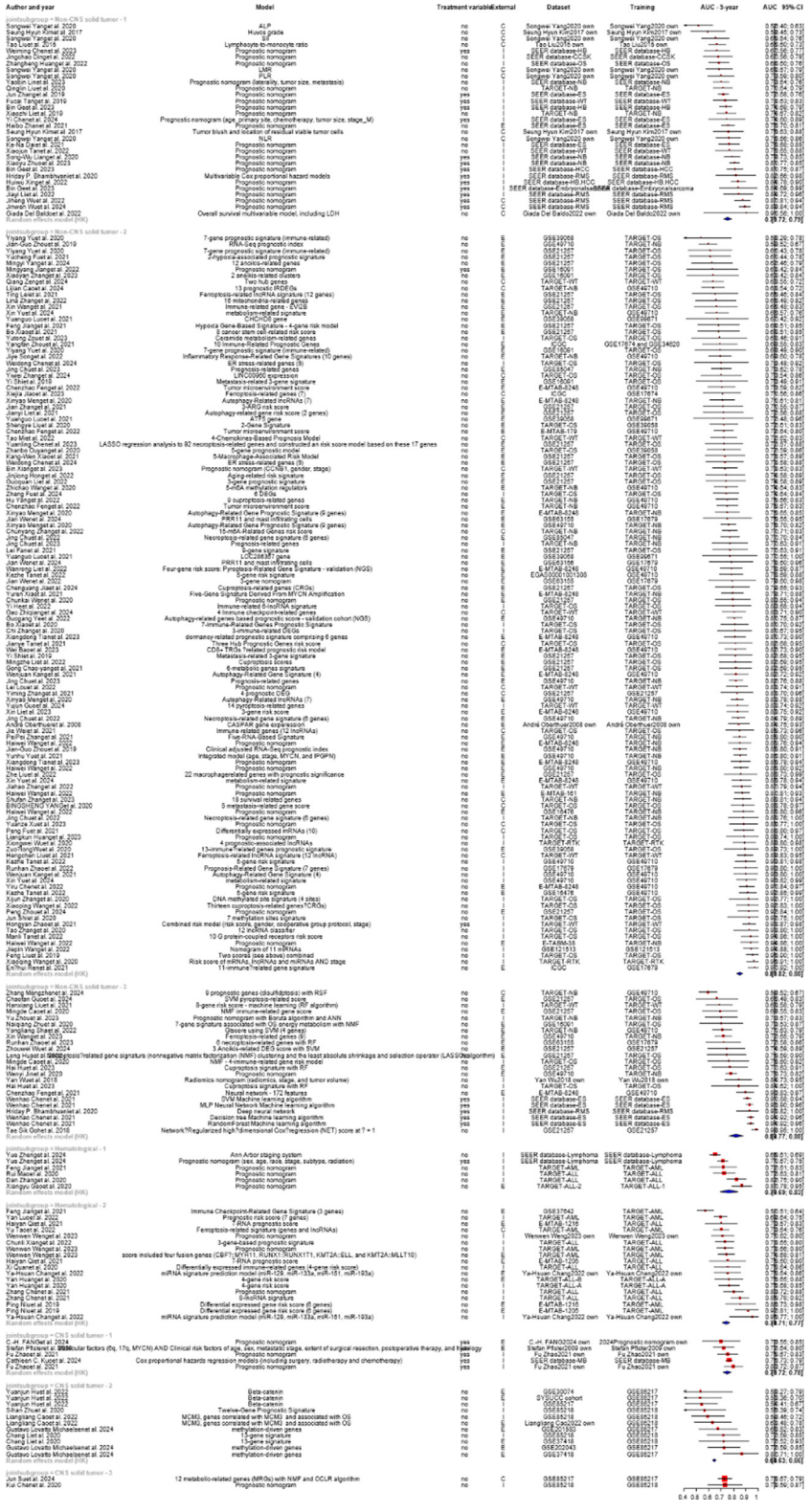

Fig. 2: Area under the receiver operating characteristic curve (AUC) with 95% confidence interval (CI) of predicting 5-year OS in paediatric patients with non-central nervous system (non-CNS), haematological malignancies, central nervous system (CNS) tumours with three model categories.[1–3] I-internal validation, E-external validation, C-composite validation.

0.73–0.83) (Category-1 VS -2, p = 0.578), respectively. The difference was statistically significant as well (Category-1 VS -3, p = 0.014; Category-2 VS -3, p = 0.035) Category-1 and Category-2 models performed comparably in HM, with pAUROC values of 0.74 (CI: 0.64–0.84) and 0.72 (CI: 0.67–0.77) (Category-1 VS -2, p = 0.805), respectively.

When we examined EFS, the NCNSST cohort provided data across all model categories. The results were similar, 0.77 (CI: 0.74–0.81), 0.81 (CI: 0.74–0.89), and 0.78 (CI: 0.75–0.82) for Categories-1 to –3 (Category-1 VS -2, p = 0.377; Category-1 VS -3, p = 0.715), respectively. In HM, pAUROC values of 0.81 (CI: 0.74–0.88) and 0.78 (CI: 0.74–0.81) were estimated in Categories-1 and -2 (p = 0.454), respectively (Fig. 7A).

For the non-time dependent EFS in the subgroup analysis, we could only make one meaningful statistical comparison, due to the scarcity of data, which resulted in no significant difference (1-no-I VS 1-no-C, p = 0.62). The prediction of OS showed no significant difference between the 2- and 3-no-E groups (p = 0.063), similar to time-dependent OS. However, we did see a notable, although not statistically significant, difference in the prediction of OS, between Categories-1 and -2 in both HM and NCNSST patients. Interestingly, for HM patients, Category-1 methods produced better results (1-no-C VS 2-no-C, p = 0.339) compared to NCNSST patients (1-no-C VS 2-no-C, p = 0.345) (Fig. 7B).

As a secondary outcome, the C-index provides further insights into predictive efficacy across OS, EFS, and CSS. In NCNSST patients, robust performance was observed across all model categories for OS predictions, with Category-3 models attaining a pooled C-index (pC-index) of 0.81 (CI: 0.74–0.80), compared to 0.74 (CI: 0.71–0.76) in Category-1 and 0.77 (CI: 0.74–0.80) in Category-2. For OS-predictions for HM and CNST, which were predominantly facilitated by Category-1 models, pC-indices of 0.73 (CI: 0.68–0.78, HM) and 0.69 (CI: 0.63–0.75, CNST) were observed. OS in HM was also predicted with Category-2, yielding a pAUROC of 0.71 (CI: 0.64–0.78, HM). The C-index of EFS prediction in NCNSST patients with Category-1 models was analysed, with pC-index values of 0.75 (CI: 0.68–0.80) In HM patients, both Category-1 and -2 models were analysed, with pC-indices of 0.67 (CI: 0.63–0.71) and a higher 0.74 (CI: 0.67–0.80), respectively. A pC-index of 0.76 (CI: 0.71–0.79) was observed in the prediction of CSS in NCNSST with Category-1 models (Fig. 8A).

For CSS, we could not make comparisons due to scarce data, whereas for EFS, we could make one comparison without statistical significance (NCNSST patients 1-no-C VS 1-no-I, p = 0.277).

In the subgroup analysis, the most common outcome measured with C-index was OS. We could not detect any statistical difference between Category-1 (1-no-E) and –2 (2-no-E) in NCNSST patients, the latter even showing a slightly worse performance

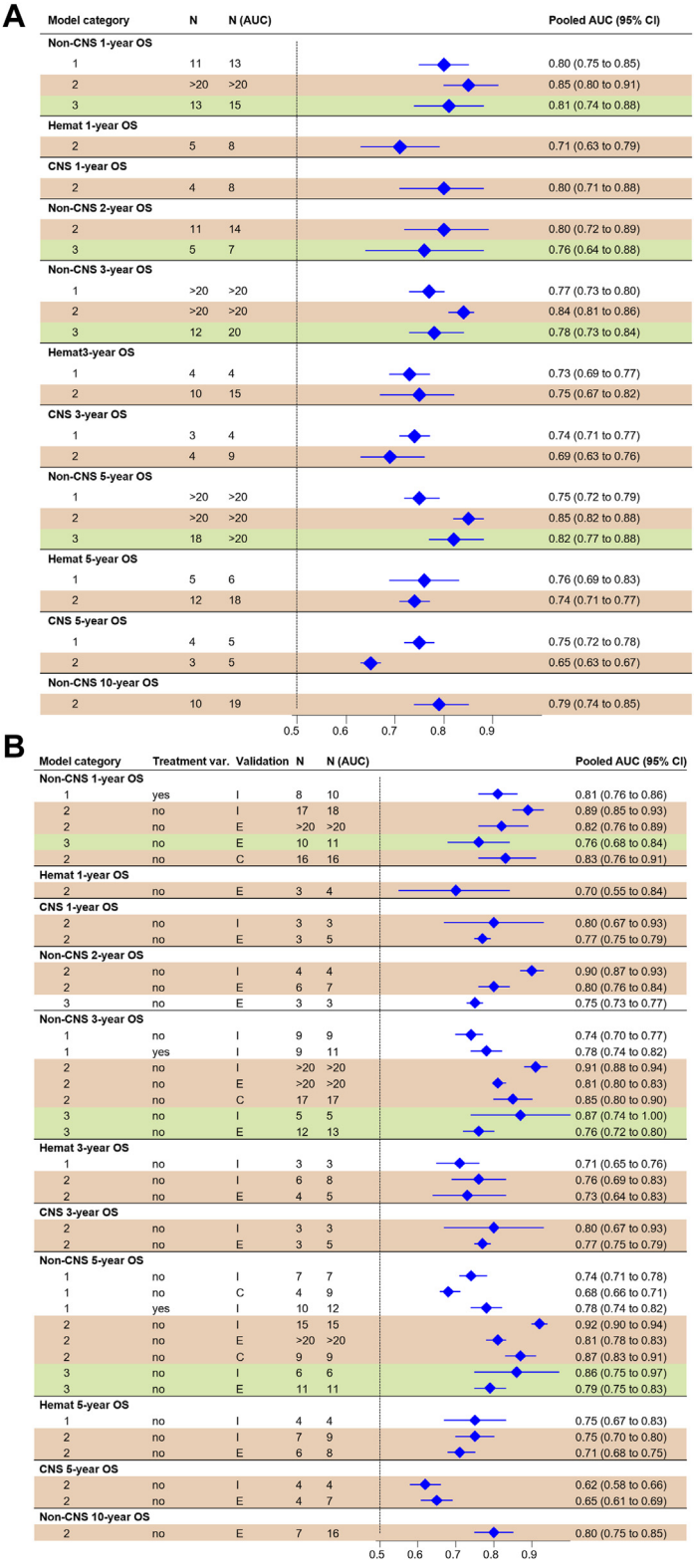

Fig. 3: Summary forest plot demonstrating the accuracy of three model categories[1–3] predicting 1-, 2-, 3-, 5- and 10-year overall survival of paediatric cancer patients, with haematological malignancies, non-central nervous system solid tumours or central nervous system tumours.

(pC-index of 0.81 (CI: 0.75–0.94) VS 0.74 (CI: 0.68–0.80), p = 0.1916). We also did not observe the marked differences between validation methods presented earlier, with studies using external validation performing better on Category-1 models than internal validation users (1-no-I VS 1-no-E pC-index 0.72 (CI: 0.69–0.76) VS 0.81 (CI: 0.75–0.94), p = 0.11). We examined the effect of using therapy details as prognostic factors, the inclusion of these factors provided better results, but we found no statistically significant difference (1-no-I VS 1-yes-I pC-index 0.72 (CI: 0.69–0.76) VS 0.75 (CI: 0.72–0.76), p = 0.059). Sufficient articles focusing on HM were identified with this outcome for statistical comparison between validation types, but the difference was not significant (1-no-E VS 1-no-I, p = 0.5553) (Fig. 8B).

The ROB assessment showed low (ROB) in the participants and outcome domains due to strict inclusion criteria. However, high ROB was identified in the Predictors and Analysis domains, mainly due to inappropriate variable handling, use of univariable analysis, and internal validation methods. Further details and the results of the ROB assessment are presented in Supplementary Fig. S41 (QUIPS) and Supplementary Fig. S42 (PROBAST).

See results for Category-0, subgroup analysis of specific training and validation dataset pairs, systematic review studies in the Supplementary Results.

Discussion

Our study provides the first Level II evidence[38] on the predictive accuracy of prognosis in PO. We examined 385 studies for short-, mid-, and -long-term survival and various unfavourable events. The superior performance of progressive, Category-2 and-3 models was shown in case of 5-years OS in NCNSST (Category-1 VS -2, p < 0.001; Category-1 VS -3 p = 0.0511), with Category 3 proven to be the best-performing predictor for non-time dependent OS (Category-1 VS -3, p = 0.014; Category-2 VS -3, p = 0.035). The strong signifying effect of the validation method was also demonstrated by comparing model performance after internal and external validation (3- and 5-year OS, 2-no-I VS 2-no-E (p < 0.001 for both outcomes).

In recent years, better biological understanding due to genetic and molecular data collection has improved histological classification, prognosis prediction and risk stratification.[39] Extensive research on this was observed in the literature, as 161 of our included studies used modern, Category-2 and -3 methods. Classic Category-1 modelling with robust statistics was already present before 2000, with the earliest eligible articles from the 1990s. After 2020, the number of articles has multiplied, with Category-2 being the most prevalent study type in this period (56.5% of 184 studies from the 2020s) as the use of NGS became more widespread. We note that Category-1 is still popular, with almost half of Category-1 articles (45.45%) written after 2020. Category-3 has not seen a significant increase over time. Although studies employing this approach date back to as early as 2004, we could identify only a total of 50 (45 for the meta-analysis and 5 for the systematic review) articles, harnessing AI.

We observed a rapid evolution in model building from Category-0 to −3, with next-generation methods such as NGS (present in both Categories-2 and -3) examining hundreds of factors, compared to traditional clinical models (Category-1) which consider significantly fewer factors. Starting from the origin, the authors of these papers also considered basic prognostic factors, which were typically employed as a comparison to the scoring systems developed by themselves (categorised as Category-0). These generally tend to perform significantly lower than the actual models (data not shown). However, the established clinical scoring systems continue to provide a solid foundation in paediatric cancer care for the upcoming decades[40] to date. We observed that Category-1 models that included pathological (e.g., histology, stage) or radiomic characteristics performed particularly well. Indeed, for non-time-dependent OS, Category-1 showed better results than Category-2 in HM. Given the relative prevalence of HMs[8] compared to other pediatric oncology complications, even Category-1 models in this subgroup are already well-developed and incorporate more accurate predictive factors. In our analysis, the majority of articles fell into Category-2, where models frequently derived prognostic scores from thousands of genes or RNA sequences, integrated with clinical factors or staging systems. This combination resulted in more accurate and comprehensive models. Nonetheless, Category-3 models, which integrate clinical characteristics, biomarkers, and genomic-transcriptomic data using an AI approach, were less common in the articles, they still showed comparable efficiency to Category-2 models.

The rarity of PC occurrences[7] and the limited availability of data are major obstacles in developing accurate prognostic models. Enriching data sources with the least prevalent paediatric tumours, such as those in the NCNSST[20,41] subclass focusing on predicting OS may be a deliberate strategy. Unlike HMs, where OS rates are relatively high,[5] the recommended prediction targets are EFSs, which correlates not only with survival but also

A: main analysis based on tumour type and model category. **B**: Subgroup analysis with the addition of model validation and the use of therapy outcome as prognostic factor. AUC: area under the receiver operating characteristic curve. N: number of articles in a certain group. N (AUC): number of AUC values in a certain group. CI: confidence interval. OS: overall survival. var.: variable. I: internal validation. E: external validation. C: composite validation. Non-CNS: non-central nervous system solid tumours. Hemat: haematological malignancies. CNS: central nervous system tumours. White: Category-1. Red: Category-2. Green: Category-3.

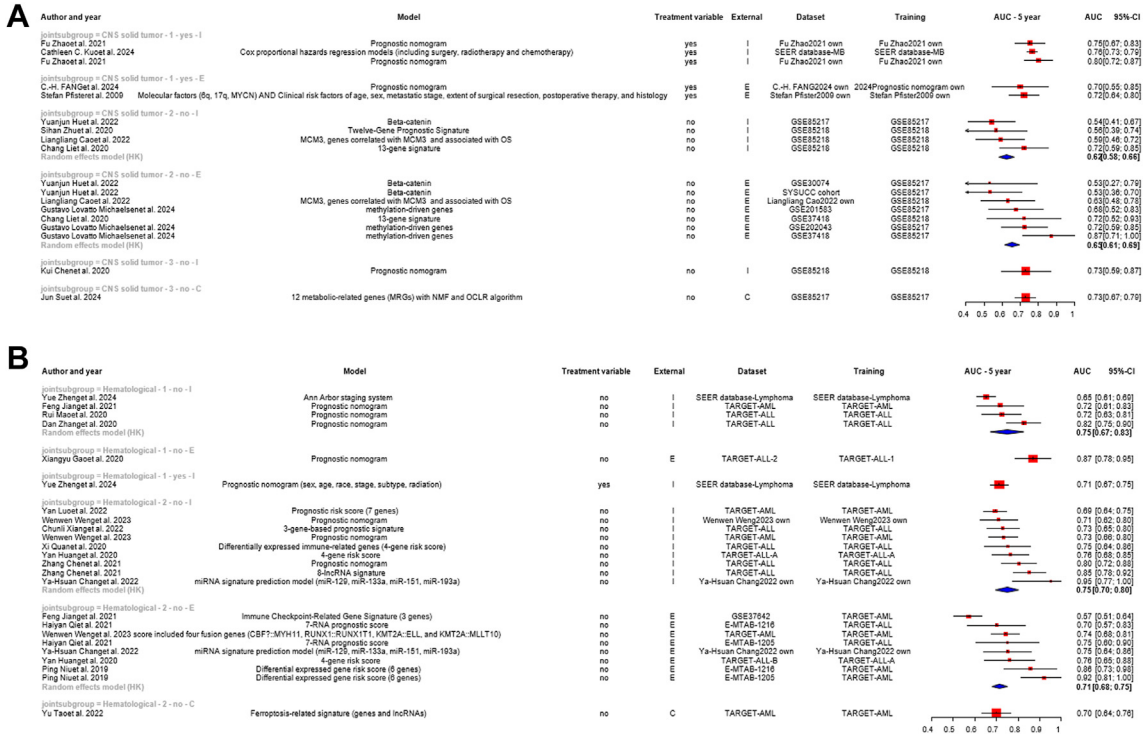

Fig. 4: **A**: Area under the receiver operating characteristic curve (AUC) with 95% confidence interval (CI) of predicting 5-year OS in paediatric patients with central nervous system (CNS) tumours. I-internal validation, E-external validation, C-composite validation. **B**: Area under the receiver operating characteristics curve (AUC) with 95% confidence interval (CI) of predicting 5-year OS in paediatric patients with haematological malignancies. I-internal validation, E-external validation, C-composite validation.

with quality of life. In terms of short-, mid- and long-term predictions for patients with NCNSST, all model categories accurately predicted OS after one year, but differences emerged after 3 and 5 years, with Category-1 models showing a decline. Interestingly, this difference was not observed for HM, although this finding should be treated with caution, based on the significant bias-effect of the type of validation. Category-1 models may have shown an advantage in HM for both OS and EFS-predictions, but in most of the articles these models were subject to internal validation. These results are therefore skewed towards appearing better or similarly good than the almost exclusively externally validated advanced models. The best practice in today's prognostic model building is the use of separate external validation sets, which was confirmed by our meta-analysis.

As already highlighted, advanced prognostic models have demonstrated greater effectiveness in NCNSST prognostics, especially in predicting OS, which remains the gold standard in oncology prognostic practice. Risk-stratification upon this outcome is essential for optimizing complex treatment protocols for each individual cases.[10,11] Our analysis demonstrated statistically significant superiority in 5-year OS prediction using both Category-2 and Category-3 models, with Category-3 outperforming Category-2 in non-time dependent OS predictions. However, our updated systematic search conducted in late June revealed that many studies published within the past 1.5 years still relied on Category-1 (30 new articles found) prognostics for OS prediction, even in NCNSST. Given these insights, we advocate for the inclusion of second- or third-generation sequencing advantages for improving paediatric cancer prognostics, whenever the resources allow for it. While AI-driven models leveraging comprehensive sequencing datasets hold the potential for even greater accuracy, the current meta-analysis lacks a sufficient number of studies to fully validate this hypothesis (Category-2 VS Category-3168 VS 45 studies in the meta-analysis altogether). Moreover, there has been a marked increase in the number of published articles on HM predictions utilising Category-2 models between 2022 and 2024 (10 articles added to the previous 17), underscoring this as a rapidly evolving research area. These models, however, statistically not proven, appeared to provide more accurate EFS estimates compared to OS predictions. However, as previously noted, no studies have yet assessed the performance of Category-3 models in this tumour subclass. Given the increasing body of evidence

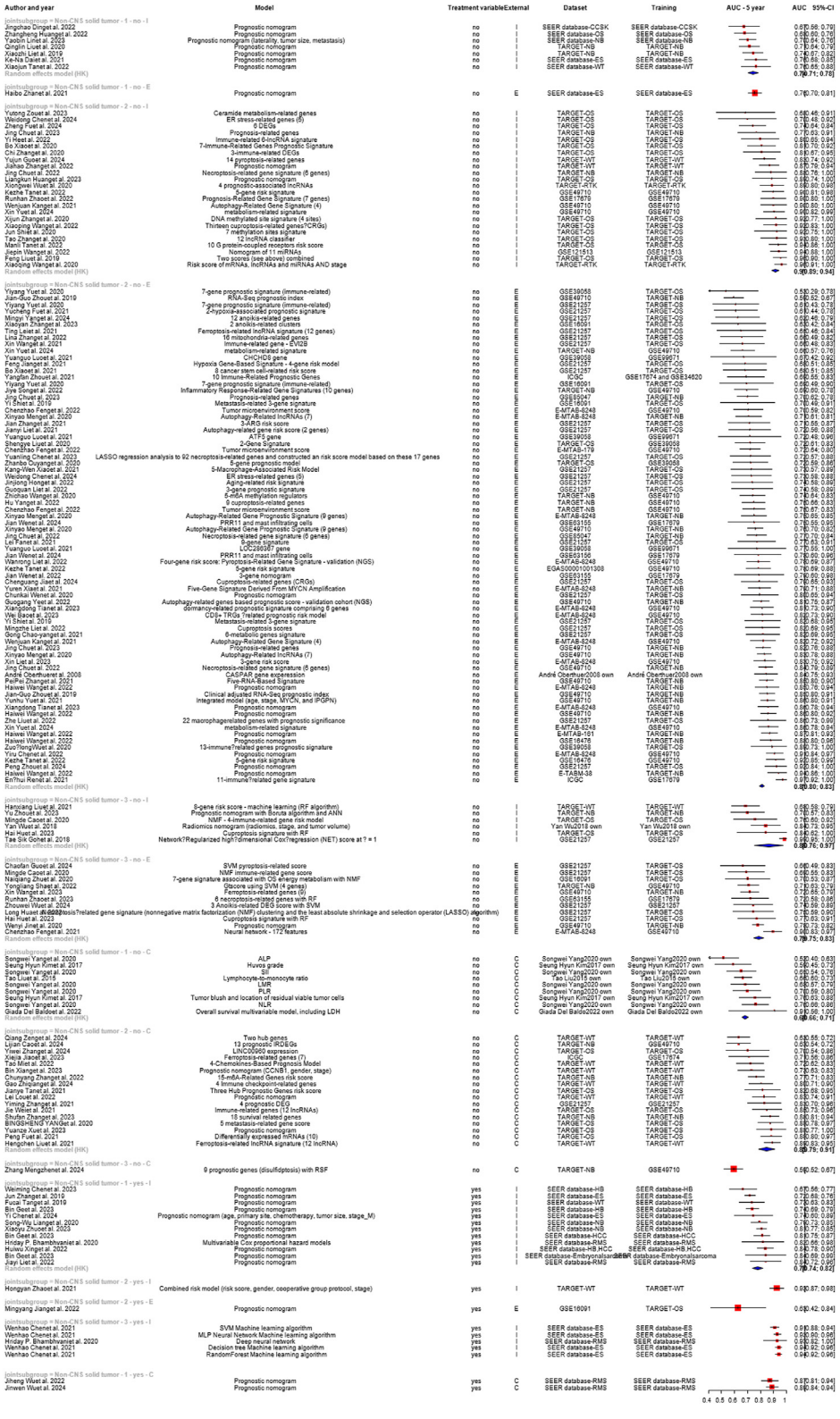

Fig. 5: Area under the receiver operating characteristic curve (AUC) with 95% confidence interval (CI) of predicting 5-year OS in paediatric patients with non-central nervous system (non-CNS) tumours. I-internal validation, E-external validation C-composite validation.

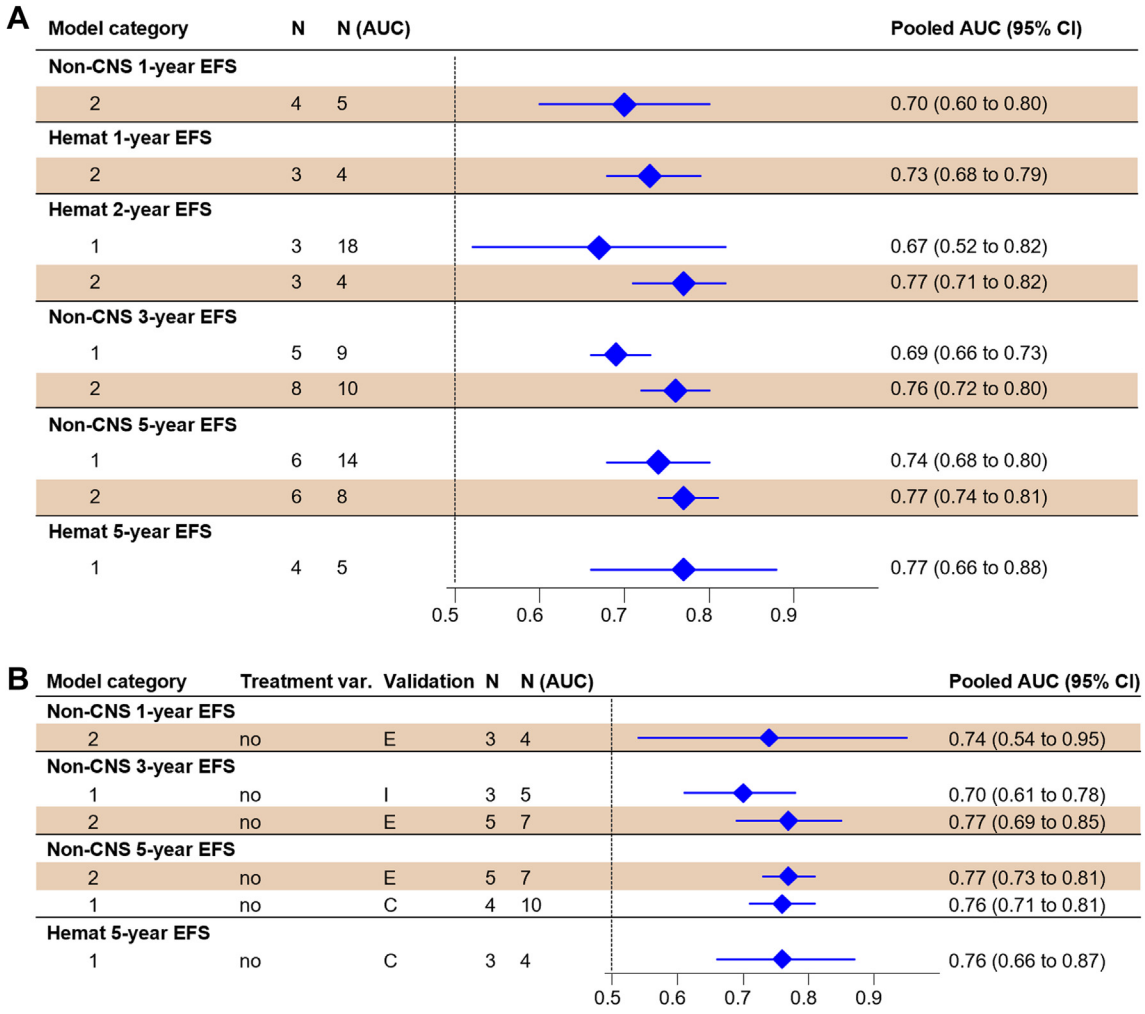

Fig. 6: Summary forest plot demonstrating the accuracy of three model categories[1–3] predicting 1-, 2-, 3-, 5-year event-free survival of paediatric cancer patients, with haematological malignancies or non-central nervous system solid tumours. A: main analysis based on tumour type and model category B: subgroup analysis with the addition of model validation and the use of therapy outcome as prognostic factor. AUC: area under the receiver operating characteristic curve. N: number of articles in a certain group. N (AUC): number of AUC values in a certain group. CI: confidence interval. EFS: event-free survival. var.: variable. I: internal validation. E: external validation. C: composite validation. Non-CNS: non-central nervous system solid tumours. Hemat: haematological malignancies. White: Category-1. Red: Category-2. Green: Category-3.

supporting EFS as a reliable surrogate endpoint in various HMs[42–44] it is strongly recommended that research efforts now focus on evaluating the potential of Category-3 models.

The method of data processing and factor selection is as important in model building as the factors chosen. Using univariate Cox regression and then multivariate Cox regressions with the chosen variables, although quite common, may introduce bias by ignoring relationships between factors. This finding supports the use of other mathematical methods to eliminate this bias. In addition, internally validated models outperform externally validated ones, indicating reduced accuracy when applied to different populations. A bias from therapy outcome as a prognostic factor was observed, but not statistically significant due to limited use in articles.

Systematic, quality-assured bio-resources are crucial for biomedical science and personalised therapies, yet they are under-utilised and lack data harmonisation, especially for vulnerable populations.[41,45] Our observations show that such databases (e.g., Therapeutically Applicable Research to Generate Effective Treatments (TARGET) and Surveillance, Epidemiology, and End Results (SEER) Database) are fundamental for developing accurate prediction models. Further international data collection and sharing is therefore necessary to establish large, quality-assured databases for training

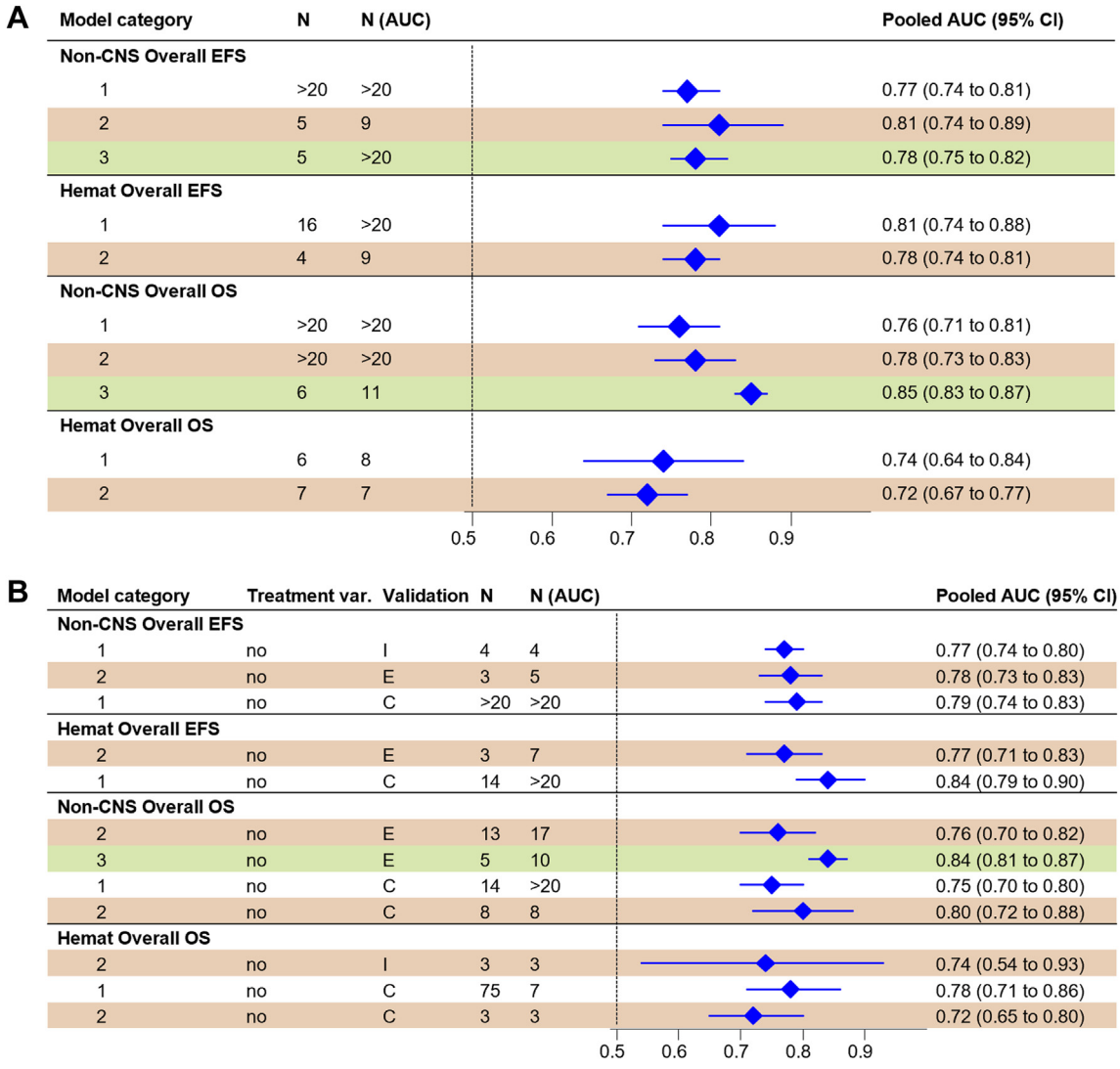

Fig. 7: Summary forest plot demonstrating the accuracy of three model categories[1–3] predicting overall and event-free survival of paediatric cancer patients with haematological malignancies or non-central nervous system solid tumours. **A**: Main analysis based on tumour type and model category. **B**: Subgroup analysis with the addition of model validation and the use of therapy outcome as prognostic factor AUC: area under the receiver operating characteristic curve. N: number of articles in a certain group. N (AUC): number of AUC values in a certain group. CI: confidence interval. OS: overall survival. EFS: event-free survival. var.: variable. I: internal validation. E: external validation. C: composite validation. Non-CNS: non-central nervous system solid tumours. Hemat: haematological malignancies. White: Category-1. Red: Category-2. Green: Category-3.

AI-driven algorithms. With emphasis on appropriate validation, modern models offer improved prediction accuracy in clinical settings.[46,47]

We included a large number of articles and outcomes to provide a thorough review of the available literature, including many up-to-date state-of-the-art studies. Several subgroups were considered when performing the analysis to address heterogeneity and bias. A rigorous methodology was applied following well-established guidelines. Most of the included studies used publicly available quality-assured datasets, making their work transparent and reliable.

We encountered a number of statistical challenges requiring adequate resolution. Standard errors of AUROC and C-index values were frequently missing, necessitating estimation from various available data types. The consistent presence of moderate-to-high ROB is another limitation. Stringent quality control was implemented throughout data collection and analysis, accompanied by transparent reporting of biases. The

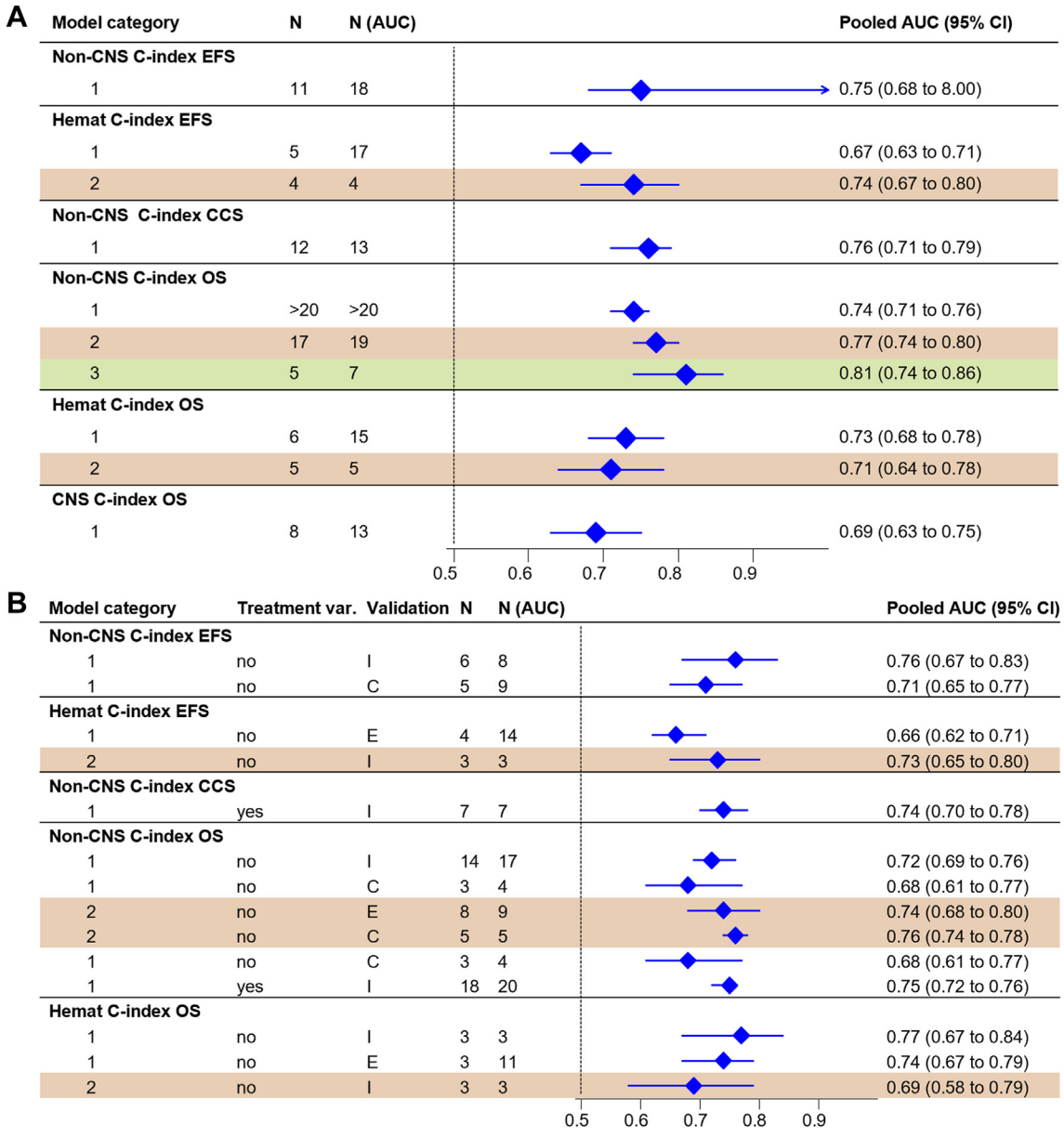

Fig. 8: Summary forest plot demonstrating the accuracy of three model categories[1–3] predicting overall, event-free, and cancer-specific survival of paediatric cancer patients with haematological malignancies, non-central nervous system solid tumours or central nervous system tumours. **A**: Main analysis based on tumour type and model category. **B**: Subgroup analysis with the addition of model validation and the use of therapy outcome as a prognostic factor. C-index: concordance index. N: number of articles in a certain group. N (C-index): number of C-index values in a certain group. CI: confidence interval. OS: overall survival. EFS: event-free survival. CSS: cancer-specific survival. var.: variable. I: internal validation. E: external validation. C: composite validation. Non-CNS: non-central nervous system solid tumours. Hemat: haematological malignancies. CNS: central nervous system tumours. White: Category-1. Red: Category-2. Green: Category-3.

absence of individual tumour type analysis introduces variability into the results, potentially masking associations existing in specific cancer subtypes, which can be uncovered by conducting subgroup analyses for individual tumour types.

Our results demonstrated the superior predictive power of genomic-transcriptomic and principally, AI-based prognostic models in the case of the NCNSST. Our subgroup analysis clearly demonstrated significant differences in accuracy based on validation type, favouring internal validation for improved but potentially biased results. While AI-based prognostic approaches have yet to be fully integrated into standard practice, our results suggest they can achieve similar

levels of accuracy to NGS-based prognostics, indicating the additional benefit of AI in paediatric cancer outcome prediction. This, however, urges the need for structured data collection and their ethical exchange to address data scarcity and limited availability, emphasizing the importance of high-quality paediatric cancer registries and biobanks.

Overall, while this systematic review and meta-analysis does not provide definitive conclusions for every tumour type, it remains the most extensive and up-to-date meta-analysis in the field, offering valuable insights and clear directions for future investigation.

Contributors

Petra Varga: conceptualisation, project administration, methodology, formal analysis, visualisation, accessing and verifying the data, writing – original draft; **Mahmoud Obeidat:** conceptualisation, methodology, visualisation, writing – review & editing; **Vanda Máté:** conceptualisation, methodology, writing - review & editing; **Tamás Kói:** conceptualisation, formal analysis, data curation, visualisation, accessing and verifying the data, writing – original draft; **Szilvia Kiss-Dala:** conceptualisation, formal analysis, data curation, visualisation, accessing and verifying the data, writing - review & editing; **Gréta Szilvia Major:** conceptualisation, data curation, writing - review & editing; **Ágnes Eszter Tímár:** conceptualisation, data curation, writing - review & editing; **Xinyi Li:** conceptualisation, data curation, writing - review & editing; **Ádám Szilágyi:** conceptualisation, data curation, writing - review & editing; **Zsófia Csáki:** conceptualisation, data curation, writing - review & editing; **Marie Engh:** conceptualisation, writing - review & editing; **Miklós Garami:** conceptualisation, writing - review & editing; **Péter Hegyi:** conceptualisation, writing - review & editing; **Ibolya Túri:** conceptualisation, writing - review & editing; **Eszter Tuboly:** conceptualisation, supervision, accessing and verifying the data, writing – original draft.

All authors certify that they have participated sufficiently in the work to take public responsibility for the content, including participation in the concept, design, analysis, writing, or revision of the manuscript.

Availability of data

All the data are available in the full text and Supplementary Materials of the included studies.

Data sharing statement

All extracted data supporting the findings of this specific systematic review and meta-analysis are available upon request after approval of a proposal from the corresponding author (E.T., tuboly.eszter@gyerekklinika.com).

Declaration of interests

All authors declare no competing interests.

Acknowledgements

None to declare. No funding.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2024.102902.

References

Citations of the included articles can be found in Supplementary material.

1 Blattner-Johnson M, Jones DTW, Pfaff E. Precision medicine in pediatric solid cancers. *Semin Cancer Biol*. 2022;84:214–227.
2 Zou H, Poore B, Broniscer A, Pollack IF, Hu B. Molecular heterogeneity and cellular diversity: implications for precision treatment in medulloblastoma. *Cancers*. 2020;12(3):643.
3 Quessada J, Cuccuini W, Saultier P, Loosveld M, Harrison CJ, Lafage-Pochitaloff M. Cytogenetics of pediatric acute myeloid leukemia: a review of the current knowledge. *Genes*. 2021;12(6):924.
4 Erdmann F, Frederiksen LE, Bonaventure A, et al. Childhood cancer: survival, treatment modalities, late effects and improvements over time. *Cancer Epidemiol*. 2021;71(Pt B):101733.
5 Oskarsson T, Soderhall S, Arvidson J, et al. Relapsed childhood acute lymphoblastic leukemia in the Nordic countries: prognostic factors, treatment and outcome. *Haematologica*. 2016;101(1):68–76.
6 Krzyszczyk P, Acevedo A, Davidoff EJ, et al. The growing role of precision and personalized medicine for cancer treatment. *Technology*. 2018;6(03n04):79–100.
7 Steliarova-Foucher E, Colombet M, Ries LAG, et al. International incidence of childhood cancer, 2001–10: a population-based registry study. *Lancet Oncol*. 2017;18(6):719–731.
8 Wu Y, Deng Y, Wei B, et al. Global, regional, and national childhood cancer burden, 1990–2019: an analysis based on the Global Burden of Disease Study 2019. *J Adv Res*. 2022;40:233–247.
9 Lupo PJ, Spector LG. Cancer progress and priorities: childhood cancer. *Cancer Epidemiol Biomarkers Prev*. 2020;29(6):1081–1094.
10 Chen K, Huang B, Yan S, et al. Two machine learning methods identify a metastasis-related prognostic model that predicts overall survival in medulloblastoma patients. *Aging*. 2020;12(21):21481–21503.
11 Liu Z, Liang M, Grant CN, Spiegelman VS, Wang H-G. Interpretable models for high-risk neuroblastoma stratification with multi-cohort copy number profiles. *Inform Med Unlocked*. 2021;25:100701.
12 Scrideli CA, Queiroz RGDP, Bernardes JE, Valera ET, Tone LG. PCR detection of clonal IgH and TCR gene rearrangements at the end of induction as a non-remission criterion in children with ALL: comparison with standard morphologic analysis and risk group classification. *Med Pediatr Oncol*. 2003;41(1):10–16.
13 Bulzico D, De Faria PAS, De Paula MP, et al. Recurrence and mortality prognostic factors in childhood adrenocortical tumors: analysis from the Brazilian National Institute of Cancer experience. *Pediatr Hematol Oncol*. 2016;33(4):248–258.
14 Morandi F, Corrias MV, Levreri I, et al. Serum levels of cytoplasmic melanoma-associated antigen at diagnosis may predict clinical relapse in neuroblastoma patients. *Cancer Immunol Immunother*. 2011;60(10):1485–1495.
15 Zhang WB, Han FM, Liu LM, Jin HB, Yuan XY, Shang HS. Characterizing the critical role of metabolism in osteosarcoma based on establishing novel molecular subtypes. *Eur Rev Med Pharmacol Sci*. 2022;26(8):2926–2943.
16 Qi W, Yan Q, Lv M, Song D, Wang X, Tian K. Prognostic signature of osteosarcoma based on 14 autophagy-related genes. *Pathol Oncol Res*. 2021;27:1609782.
17 Qian H, Lei T, Hu Y, Lei P. Expression of lipid-metabolism genes is correlated with immune microenvironment and predicts prognosis in osteosarcoma. *Front Cell Dev Biol*. 2021;9:673827.
18 O'Donohue T, Farouk Sait S, Glade Bender J. Progress in precision therapy in pediatric oncology. *Curr Opin Pediatr*. 2023;35(1):41–47.
19 Quintás G, Yáñez Y, Gargallo P, et al. Metabolomic profiling in neuroblastoma. *Pediatr Blood Cancer*. 2020;67(3):e28113.
20 Wang X, Wu X, Li T, et al. Identification of biomarkers to construct a competing endogenous RNA network and establishment of a genomic-clinicopathologic nomogram to predict survival for children with rhabdoid tumors of the kidney. *BioMed Res Int*. 2020;2020:1–27.
21 Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71.
22 Chandler J, Hopewell S. Cochrane methods - twenty years experience in developing systematic review methods. *Syst Rev*. 2013;2(1):76.
23 Booth A, Clarke M, Dooley G, et al. PROSPERO at one year: an evaluation of its utility. *Syst Rev*. 2013;2(1):4.
24 Munn Z, Stern C, Aromataris E, Lockwood C, Jordan Z. What kind of systematic review should I conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences. *BMC Med Res Methodol*. 2018;18(1):5.
25 Haddaway NR, Grainger MJ, Gray CT. Citationchaser: a tool for transparent and efficient forward and backward citation chasing in systematic searching. *Res Synth Methods*. 2022;13(4):533–545.
26 Hayden JA, Van Der Windt DA, Cartwright JL, Côté P, Bombardier C. Assessing bias in studies of prognostic factors. *Ann Intern Med*. 2013;158(4):280.

27 Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51.
28 Team RCR. *A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing; 2021.
29 Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29–36.
30 Debray TP, Damen JA, Riley RD, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. *Stat Methods Med Res*. 2019;28(9):2768–2786.
31 Genomic data commons, therapeutically applicable research to generate effective treatments 2007-2024. Available from: https://www.cancer.gov/ccg/research/genome-sequencing/target.
32 Zhang W, Shi L, Hertwig F, et al. *SuperSeries GSE47792; SubSeries GSE49710, GSE49711, GSE62564*. 2014.
33 Hammerschmidt W, Buschle A. *E-MTAB-8428*. 2021.
34 Davis S. *SuperSeries GSE16102; SubSeries GSE16091*. 2009.
35 Buddingh EPK, Marieke L, Duim RAJ, et al. *Anne-marie series GSE21257*. 2011.
36 Kelly A. *SuperSeries GSE39058*. 2013.
37 Ho D, Kõks S, Phung P, et al. *Series GSE99671*. 2017.
38 Burns PB, Rohrich RJ, Chung KC. The levels of evidence and their role in evidence-based medicine. *Plast Reconstr Surg*. 2011;128(1):305–310.
39 American Cancer S. *Cancer facts & figures 2023*. Atlanta: American Cancer Society; 2023.
40 Lehrnbecher T, Robinson PD, Ammann RA, et al. Guideline for the management of fever and neutropenia in pediatric patients with cancer and hematopoietic cell transplantation recipients: 2023 update. *J Clin Oncol*. 2023;41(9):1774–1785.
41 Joseph N, Roberts CK, Graham Kathryn, et al. Biobanking in the twenty-first century: driving population metrics into biobanking quality. In: Karimi-Busheri F, ed. *Biobanking in the 21st century. Advances in experimental medicine and biology*. vol. 864. Cham: Springer; 2015:95–114.
42 Maurer MJ, Ellin F, Srour L, et al. International assessment of event-free survival at 24 Months and subsequent survival in peripheral T-cell lymphoma. *J Clin Oncol*. 2017;35(36):4019–4026.
43 Zhu J, Yang Y, Tao J, et al. Association of progression-free or event-free survival with overall survival in diffuse large B-cell lymphoma after immunochemotherapy: a systematic review. *Leukemia*. 2020;34(10):2576–2591.
44 Norsworthy KJ, Gao X, Ko CW, et al. Response rate, event-free survival, and overall survival in newly diagnosed acute myeloid leukemia: US food and drug administration trial-level and patient-level analyses. *J Clin Oncol*. 2022;40(8):847–854.
45 Rush A, Byrne JA, Watson PH. *Applying findable, accessible, interoperable, and reusable principles to biospecimens and biobanks*. Biopreserv Biobank; 2024.
46 Hegyi P, Eross B, Izbeki F, Parniczky A, Szentesi A. Accelerating the translational medicine cycle: the Academia Europaea pilot. *Nat Med*. 2021;27(8):1317–1319.
47 Hegyi P, Petersen OH, Holgate S, et al. Academia europaea position paper on translational medicine: the cycle model for translating scientific results into community benefits. *J Clin Med*. 2020;9(5):1532.