Computer Methods and Programs in Biomedicine 250 (2024) 108166

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Predicting adverse long-term neurocognitive outcomes after pediatric intensive care unit admission

Felipe Kenji Nakano [a,b,*], Karolijn Dulfer [c], Ilse Vanhorebeek [d], Pieter J. Wouters [d], Sascha C. Verbruggen [c], Koen F. Joosten [c], Fabian Güiza Grandas [d,1], Celine Vens [a,b,1], Greet Van den Berghe [d,1]

[a] KU Leuven, Campus KULAK, Department of Public Health and Primary Care, Etienne Sabbelaan 53, Kortrijk, 8500, Belgium
[b] Itec, imec research group at KU Leuven, Etienne Sabbelaan 53, Kortrijk, 8500, Belgium
[c] Intensive Care Unit, Department of Paediatrics and Paediatric Surgery, Erasmus Medical Centre, Sophia Children's Hospital, Doctor Molewaterplein 40, Rotterdam, 3015 GD, the Netherlands
[d] Clinical Division and Laboratory of Intensive Care Medicine, Department of Cellular and Molecular Medicine, UZ Herestraat 49, Leuven, 3000, Belgium

ARTICLE INFO

Keywords:
Pediatric intensive care units
Label ranking
Personalized healthcare
Machine learning

ABSTRACT

Background and objective: Critically ill children may suffer from impaired neurocognitive functions years after ICU (intensive care unit) discharge. To assess neurocognitive functions, these children are subjected to a fixed sequence of tests. Undergoing all tests is, however, arduous for former pediatric ICU patients, resulting in interrupted evaluations where several neurocognitive deficiencies remain undetected. As a solution, we propose using machine learning to predict the optimal order of tests for each child, reducing the number of tests required to identify the most severe neurocognitive deficiencies.
Methods: We have compared the current clinical approach against several machine learning methods, mainly multi-target regression and label ranking methods. We have also proposed a new method that builds several multi-target predictive models and combines the outputs into a ranking that prioritizes the worse neurocognitive outcomes. We used data available at discharge, from children who participated in the PEPaNIC-RCT trial (ClinicalTrials.gov-NCT01536275), as well as data from a 2-year follow-up study. The institutional review boards at each participating site have also approved this follow-up study (ML8052; NL49708.078; Pro00038098).
Results: Our proposed method managed to outperform other machine learning methods and also the current clinical practice. Precisely, our method reaches approximately 80% precision when considering top-4 outcomes, in comparison to 65% and 78% obtained by the current clinical practice and the state-of-the-art method in label ranking, respectively.
Conclusions: Our experiments demonstrated that machine learning can be competitive or even superior to the current testing order employed in clinical practice, suggesting that our model can be used to severely reduce the number of tests necessary for each child. Moreover, the results indicate that possible long-term adverse outcomes are already predictable as early as at ICU discharge. Thus, our work can be seen as the first step to allow more personalized follow-up after ICU discharge leading to preventive care rather than curative.

1. Introduction

Critically ill children may suffer from impaired neurocognitive functions years after pediatric intensive care unit (PICU) discharge, which severely hinders their overall quality of life [1]. These impairments can be measured in a follow-up evaluation [2]. More specifically, the previously hospitalized patients are subjected to a sequence of clinically and internationally validated tests which evaluates outcomes related to intelligence, visual-motor integration, alertness, motor coordination and memory. The sequence of tests is fixed and the same for all patients.

Undergoing all tests is, however, arduous, time-consuming and expensive. The completion of all tests may require up to four hours. Furthermore, the tests are sometimes performed at the residence of the patients, if they are too fragile to commute to the hospital. Altogether

[*] Corresponding author.
E-mail address: felipekenji.nakano@kuleuven.be (F.K. Nakano).

[1] Fabian Güiza Grandas, Celine Vens and Greet Van den Berghe have equally contributed to this manuscript.

https://doi.org/10.1016/j.cmpb.2024.108166
Received 28 February 2023; Received in revised form 18 March 2024; Accepted 5 April 2024
Available online 10 April 2024

these factors often result in interrupted evaluations where approximately 30% of the patients do not complete all tests [2], which leads to undetected neurocognitive deficiencies. Hence, reducing the number of tests required to identify the most severe neurocognitive deficiencies is of paramount importance in follow-up evaluations.

To the best of our knowledge, we present the first study to address this problem using machine learning. Several studies have developed machine learning models to assist medical decision making in the PICU, including mortality prediction [3], sepsis [4], cardiac arrest [5] and acute kidney injury [6]. In our case, the objective consists of building a model that predicts a personalized sequence of tests for each patient. This sequence prioritizes tests associated to neurocognitive functions that are expected to be affected, and, differently from the approach currently employed in clinical practice, considers the features of each patient to perform individual predictions.

Our work provides a first step towards deciding who should be followed-up and for which outcomes. This will provide valuable clinical validation for the use of individualized follow-up programs of critically ill children at risk of worse neurocognitive functioning after PICU discharge.

We model this problem as a label ranking task, a field of machine learning that aims to build predictive models capable of predicting a ranking of outcomes according to a relevance criterion [7]. Label ranking has already been applied in medical applications, in the context of antibiotic treatment in primary care [8,9]. In the context of long-term outcome prediction, however, to the best of our knowledge, the literature still lacks studies on it.

More precisely, we propose a new label ranking approach that initially builds a multi-target prediction model per neurocognitive function that only predicts its outcomes, i.e. one separate model for: intelligence, visual-motor integration, alertness, motor coordination and memory, making a total of five models. Further, we combine the output of each model into a ranking where worse neurocognitive outcomes are placed in higher positions, and, thus, recommended to be evaluated first in the follow-up evaluation.

As input to our approach, we employ the data obtained in the 2 years follow-up of the PEPaNIC-RCT study (ClinicalTrials.gov-NCT01536275), a multicenter, randomized and controlled trial, which compared early and late parenteral nutrition in PICUs [2]. The institutional review boards at each participating site have approved this follow-up study (ML8052; NL49708.078; Pro00038098).

The outcomes of this dataset (i.e., the results of neurocognitive tests) are numerical values in different ranges, which means that they are not directly suitable for our task. In order to adapt the dataset, we propose a procedure that transforms the absolute outcome values into relative ones according to their deficiency in the patient and relative to similar healthy subjects. This way, new outcomes are associated to each patient which reflect how adverse their neurocognitive functions are in comparison to healthy individuals.

Using the transformed dataset, our experiments reveal that our proposed approach manages to perform better than the current standardized expert approach employed in clinical practice. Furthermore, we also show that, despite its simplicity, our approach outperforms several of its machine learning competitors, including the current state-of-art approach in label ranking, called *BoostLR* [10].

The main contributions of this work are:

- We present the first study that applies machine learning to predict adverse long-term neurocognitive outcomes after pediatric intensive care unit hospitalization;
- We propose a new label ranking approach which relies on building a separate multi-target model for each group of outcomes and combining the output of each model in a final ranking;
- We propose to transform the outcomes by incorporating information from healthy individuals, allowing for a more objective interpretation of the neurocognitive status of a patient;
- We perform a comparison between our proposed approach, the approach used in clinical practice and the state-of-art approach in label ranking using several evaluation measures and different feature subsets;
- We provide concrete examples of the benefits of our approach compared to the fixed sequence of tests currently employed in clinical practice;

The remainder of this paper is organized as follows: Section 2 presents a more detailed description of our methods, including dataset acquisition, our approach to the problem and evaluation measures; Section 3 discusses our experiments where we validate different parameters, compare our method against several competitors using different age groups and features and provide a concrete case analysis; Section 4 briefly summarizes our results and its implications in clinical practice; Finally, Section 5 brings our conclusions and future work directions.

2. Methods

2.1. Dataset acquisition

We employed the data from the PEPaNIC trial. PEPaNIC is a multicenter, randomized and controlled trial that compared early parenteral nutrition with withholding supplemental parenteral nutrition for 1 week in the PICU [2,11]. The PICU data was collected from June 18, 2012 to July 27, 2015 in three PICUs located in Leuven (Belgium), Rotterdam (The Netherlands) and Edmonton (Canada) [11], whereas the 2 year follow-up was performed between August 4, 2014 and January 19, 2019. Its main objective consisted of investigating the long-term impact on physical and neurocognitive outcomes of critical illness and of early versus late parenteral nutrition as compared to matched healthy controls [2].

Precisely, we made use of the data obtained in a pre-planned 2-year follow-up. All patients included in the trial were approached for assessment of physical and neurocognitive outcomes, in comparison with healthy children, matched according to age and sex, who had never been admitted in the PICU. To minimize genetic, socioeconomic and environmental background, siblings and relatives, which demographically matched the patients age and sex, were prioritized to form the control group. Exclusion criteria from the control group include previous admission to PICU or neonatal ICU and history of suspicious chronic metabolic diseases that require a special diet [2]. The dataset consists of 786 previously hospitalized patients and 405 control group individuals.

As for features, each patient is represented by 23 numerical features and 22 categorical features. These features, presented in Table 1, are separated into two groups: i) 35 features available at discharge from PICU and ii) 10 extra features available at the 2 years follow-up.

As features available at PICU discharge, this dataset presents demographics, such as: age, gender and socioeconomic status. It also contains features available at PICU, for instance, Pedriatric Index of Mortality 3 (*PIM3 score*), *STRONGkids risk score* (a screening tool for risk on nutritional status and growth), Pediatric Logistic Organ Dysfunction (*PELOD score*), and acute effects post randomization, namely *duration of ICU stay*, *duration of mechanical ventilatory support*, *duration of treatment with anti-biotics*, *duration of treatment with benzodiazepines*, a pre-randomization syndrome or a prior illness defined as affecting or possibly affecting neurocognitive development, among others.

Features available at the 2 years follow-up were obtained through validated and internationally recognized questionnaires. These questionnaires are related to several emotional, behavioral and executive functions which are completed by parents or caregivers. In the PEPaNIC follow-up study [2], the results of these questionnaires were considered outcomes. In this work, however, we use them as input features, since they can be obtained before assessment of the child and may contain relevant information.

Table 1
A detailed description of the features in our dataset. Numerical attributes (N) are further described by their mean and standard deviation in parentheses, whereas the number of occurrences and their percentage are used for categorical attributes (C).

Demographic characteristics	Controls (n = 405)	Patients (n = 786)
Age (N)	6.0 (4.7)	5.7 (4.5)
Sex (C)		
Female	186 (46%)	331 (42%)
Male	219 (54%)	455 (58%)
Known non-white race (C)	33 (8%)	63 (8%)
Known non-European origin (C)	54 (13%)	152 (19%)
Known not exclusive Dutch or English language (C)	76 (19%)	184 (23%)
Socioeconomic status: Parents educational (C)		
Level 1.0	13 (3%)	37 (5%)
Level 1.5	23 (6%)	54 (7%)
Level 2.0	55 (14%)	184 (23%)
Level 2.5	76 (19%)	131 (17%)
Level 3.0	215 (53%)	200 (25%)
Level unknown	23 (6%)	180 (23%)
Socioeconomic status: Parents occupational (C)		
Level 1.0	2 (1%)	10 (1%)
Level 1.5	25 (6%)	76 (10%)
Level 2.0	47 (12%)	127 (16%)
Level 2.5	26 (6%)	77 (10%)
Level 3.0	83 (20%)	121 (15%)
Level 3.5	40 (10%)	54 (7%)
Level 4.0	116 (29%)	108 (14%)
Level unknown	66 (16%)	213 (27%)
Test location(C)		
Hospital	NA	502 (64%)
Home	NA	279 (35%)
School	NA	4 (<1%)
COS	NA	1 (<1%)
Parental smoking between birth and PICU admission (C)		
Yes	NA	354 (45%)
No	NA	432 (55%)
Maternal smoking during pregnancy (C)		
Yes	16 (4%)	52 (7%)
No	389 (96%)	734 (93%)
Parental smoking during pregnancy (C)		
Yes	77 (19%)	315 (40%)
No	328 (81%)	471 (60%)
Maternal smoking pre-pregnancy (C)		
Yes	84 (21%)	278 (35%)
No	321 (79%)	508 (65%)
Parental smoking pre-pregnancy(C)		
Yes	378 (48%)	141 (35%)
No	408 (52%)	264 (65%)
Hand preference(C)		
Right	371 (92%)	706 (90%)
Left	34 (8%)	80 (10%)
Randomization to late vs. early initiation of PN (C)		
Early parenteral nutrition	NA	395 (50%)
Late parenteral nutrition	NA	391 (50%)
Centrum (C)		
Leuven	243 (60%)	465 (59%)
Rotterdam	162 (40%)	301 (38%)
Edmonton	0 (0%)	20 (3%)

More specifically, these questionnaires are related to executive functioning and emotional and behavioral problems of the children. Executive functioning is evaluated using the BRIEF (Behavior Rating Inventory of Executive Function) preschool version for children aged between 2.5 and 5 years, whereas BRIEF was used for patients older than 6 [12,13]. The overlapping scales and indexes of these questionnaires, (inhibition, flexibility, emotional control, working memory, planning and organization) and (meta-cognition index, comprising the scales working memory and planning and organization), respectively, and the total score were reported. Similarly, the Child Behavior Checklist [14,15] (CBCL 1.5–5 years or CBCL 6–18 years) questionnaires were used to assess emotional and behavioral problems.

As outcomes, this dataset presents neurocognitive functions evaluated using validated clinical tests. In our work, we focus on 5 groups of neurocognitive functions where each contains a different number of tests (represented in parentheses): general intelligence (3), visual-motor integration (1), alertness (4), motor coordination (4) and memory (11), resulting in 23 tests. The outcomes are further detailed in Table 2.

General intelligence was measured using the appropriate versions of the Wechsler intelligence according to the age of the patients. For younger children, aged between 2.5 years and 5 years 11 months [16], the Wechsler intelligence scale for children (WPPSI-III-NL) was used [17]. The WISC-III-NL version was employed for children aged between 6 years and 16 years 11 months. Lastly, the Wechsler adult intelligence

Table 1 (*continued*)

Patient characteristics on PICU admission		
STRONGkids risk level (C)		
Medium	NA	707 (90%)
High	NA	79 (10%)
PELOD score, first 24 h in PICU (N)	NA	20.0 (11.6)
PIM3 score (N)	NA	–3.5 (1.4)
Diagnostic category (C)		
Surgical: abdominal	NA	70 (9%)
Surgical: burns	NA	2 (<1%)
Surgical: cardiac	NA	339 (43%)
Surgical: neurosurgery or traumatic brain injury	NA	71 (9%)
Surgical: thoracic	NA	42 (5%)
Surgical: transplantation	NA	14 (2%)
Surgical: orthopedic surgery or trauma	NA	23 (3%)
Surgical: other	NA	27 (3%)
Medical: cardiac	NA	26 (3%)
Medical: gastrointestinal or hepatic	NA	3 (1%)
Medical: oncological or hematological	NA	8 (1%)
Medical: neurological	NA	44 (6%)
Medical: renal	NA	0 (0%)
Medical: respiratory	NA	83 (11%)
Medical: other	NA	34 (4%)
Malignancy (C)	0 (0%)	42 (5%)
Diabetes (C)	0 (0%)	1 (1%)
Syndrome (C)	5 (1%)	79 (10%)
Acute effects of randomization and post-randomization treatments in PICU		
Duration of stay in the PICU, days (N)	NA	7.4 (15.1)
Number of patients who acquired a new infection in PICU (N)	NA	105 (13%)
Duration of mechanical ventilatory support, days (N)	NA	4.7 (11.0)
Hypoglycemia 40 mg/dL (C)	NA	0.1 (0.5)
Hypoglycemia	NA	717 (0.91)
No Hypoglycemia	NA	69 (0.09)
Duration of antibiotic treatment, days (N)	NA	5.1 (13.4)
Duration of hemodynamic support, days (N)	NA	2.5 (7.2)
Duration of treatment with opioids, days (N)	NA	4.7 (8.8)
Duration of treatment with benzodiazepines, days (N)	NA	4.2 (9.8)
Duration of treatment with hypnotics, days (N)	NA	1.4 (5.6)
Duration of treatment with α2 agonists, days (N)	NA	1.0 (6.4)
Duration of treatment with corticosteroids, days (N)	NA	1.2 (3.7)
Features available at the 2 years follow-up		
Child Behavior Checklist (CBC)		
Internalizing problems (N)	46.7 (10.7)	51·1 (13·5)
Externalizing problems (N)	46.8 (10.1)	49·8 (13·2)
Overall problems (N)	46.1 (10.4)	50·9 (13·2)
Behavior Rating Inventory of Executive Function (BRIEF)		
Inhibition (N)	46.3 (11.5)	49·9 (15·2)
Flexibility (N)	46.7 (11.3)	49·9 (15·3)
Emotional (N)	47.7 (11.2)	49·7 (13·5)
Working memory (N)	46.7 (12.1)	51·4 (16·7)
Planning and organization (N)	46.9 (11.9)	50·3 (14·7)
Meta-cognition index (N)	46.8 (12.5)	50·2 (15·2)
Overall (N)	45.9 (11.6)	50·2 (15·4)

scale (WAIS-IV-NL) was computed for patients who were 17 years or older [18]. For all of these tests, total IQ, verbal IQ, and performance IQ scores were calculated.

As for visual–motor integration, the Beery developmental test [19] was used for children aged 2.5 years and older to compute the ability to integrate visual and motor functions.

The validated computerized Amsterdam neuropsychological tasks (ANT) program [20] was employed to measure alertness and motor coordination in children aged 4 years or older. More specifically, alertness was measured using the reaction time of both hands, and their respective standard deviation within person.

Similarly, ANT-Tapping was used to measure motor coordination [20]. In this case, the number of right hand, left hand, alternating and synchronous taps.

The assessment of memory is possible only in children aged between 5 years and 16 years 11 months, and it involves four tests from the Children's Memory Scale (CMS) [21]: CMS-Numbers (2), CMS-Word pairs (4), CMS-Picture locations (1), CMS-Dot locations (3) and CMS-Learning (1).

CMS-numbers measures short-term verbal memory span and verbal working memory load; CMS-Word Pairs measures short and long-term verbal memory; CMS-Picture Locations measures immediate visual memory; and CMS-Dot Locations measures immediate and delayed visual memory. Lastly, the CMS-Learning index corresponds to the learning abilities of the child. Subsequently, the CMS-Learning Index is presented (1).

Missing features and outcomes, due only to incomplete evaluations and not to age restrictions, were imputed using MICE [22] with 31 iterations as reported in [2]. All experiments were performed using a pooled version of the imputed datasets, which summarizes the 31 imputed versions.

We have created two versions of the dataset, using different sets of features described in Table 1.

Table 2
A detailed description of the outcomes in our dataset. All outcomes are numerical and are further described by their mean and standard deviation in parentheses.

Outcomes	Controls (n = 405)	Patients (n = 786)
Group: Intelligence (range, 45–155)		
Total IQ	100.7 (13.0)	90.6 (16.5)
Verbal IQ	100.8 (14.1)	92.0 (18.2)
Performance IQ	100.7 (13.8)	91.5 (16.4)
Group: Visual–motor integration (range, 0.9–20)		
Visual-motor integration	9.6 (2.4)	8.2 (3.5)
Group: Alertness		
Reaction time right hand, ms	480.8 (290.2)	561.1 (700.4)
Within-person SD of repeated tests	219.3 (276.0)	278.8 (715.0)
Reaction time left hand, ms	459.7 (239.2)	536.2 (538.1)
Within-person SD of repeated tests	217.3 (222.4)	287.4 (542.7)
Group: Motor coordination (number of taps in 10 s)		
Number of right hand taps	41.4 (16.1)	37.9 (41.1)
Number of left hand taps	36.3 (14.4)	34.9 (36.6)
Number of valid alternating taps	18.3 (23.2)	18.6 (63.8)
Number of valid synchronous taps	23.9 (15.1)	21.9 (35.8)
Group: Memory		
Verbal–auditory		
Numbers (range 1–19)		
Memory span (forward)	10.2 (2.9)	8.6 (5.7)
Working memory (backward)	10.3 (3.0)	8.7 (4.5)
Word pairs (proportion of correct responses)		
Learning	0.50 (0.2)	0.43 (0.8)
Immediate memory	0.47 (0.2)	0.33 (0.6)
Delayed memory	0·40 (0.3)	0.31 (0.8)
Recognition	0.95 (0.2)	0.87 (0.5)
Non-verbal, visual–spatial		
Pictures (proportion of correct responses)	0.85 (0.1)	0.78 (0.3)
Dots (proportion of correct responses)		
Learning	0.86 (0.2)	0.78 (0.5)
Immediate memory	0.87 (0.2)	0.80 (0.8)
Delayed memory	0.87 (0.2)	0.80 (0.8)
Learning index (range 50–150)	100.2 (22.5)	92.2 (85.5)

- **Discharge features**: In this version, we include only features obtainable at PICU discharge, which results in a total of 25 features;
- **Discharge and 2 years follow-up features**: In this second version, the discharge features were augmented with the CBCL and BRIEF values. As aforementioned, both CBCL and BRIEF values are provided by caregivers and are related to emotional and behavioral problems and executive functioning, respectively. This version includes a total of 35 features;

Both dataset versions were used as input to the data transformation procedure described below (for all data driven comparison approaches). All numerical features were normalized.

2.2. Our proposed approach

First, we propose a data transformation procedure to the dataset. Second, we introduce label ranking as a predictive task and we present our approach.

2.2.1. Data transformation

In its original version, all outcomes were presented in a numerical format where each outcome has its own range of values. For this study, however, we transformed the absolute outcome values into relative ones where worse outcomes are associated with higher numbers. The motivation for this step is that we are not interested in predicting the exact numeric outcomes, but rather in predicting a ranking of the outcomes according to their deficiency in the patient and relative to similar healthy subjects. In order to do so, we have adopted the procedure described in Fig. 1.

This procedure consists of two steps (shown in the upper and bottom part of the figure, respectively). In the first step, we aim to find the most similar control subjects to a given patient. To achieve this, we have employed the K-NN algorithm[2] [23]: a well-established machine learning algorithm that, given an instance (patient), identifies its most similar instances in the dataset. As input, we used its descriptive features and all control group individuals. As output, in this particular case, K-NN provides the K most similar control group individuals to the patient in question.

In the second step, we seek to build the ranking of outcomes. This is accomplished by using the outcomes of the control group (output from the first step) as normalizing factors. That is, all outcomes of the patient are scaled, generating, thus, rankings that represent a relative comparison of the patients with their most similar control subjects. Hence, the control group individuals are used exclusively to transform the dataset.

A similar normalization procedure, without the notion of control group, was also adopted by Cheng [24] to generate benchmark datasets for label ranking.

2.2.2. Label ranking as a local multi-target problem

Label ranking can be defined as a predictive task that aims to predict rankings of labels according to a relevance criterion, for instance, the severity of a neurocognitive function. More formally, given an instance space X and a set of labels Y, the goal is to map X to a set of permutations of Y, such that the relevance criterion is maximized [7,24].

Label ranking problems can be tackled in different manners. Traditional classification algorithms, such as K-NN and decision trees, can be

[2] We have employed the Euclidean distance for numerical attributes. As for the categorical ones, the hamming distance was used. Features not available in the control group, e.g. ICU related features, were not considered.

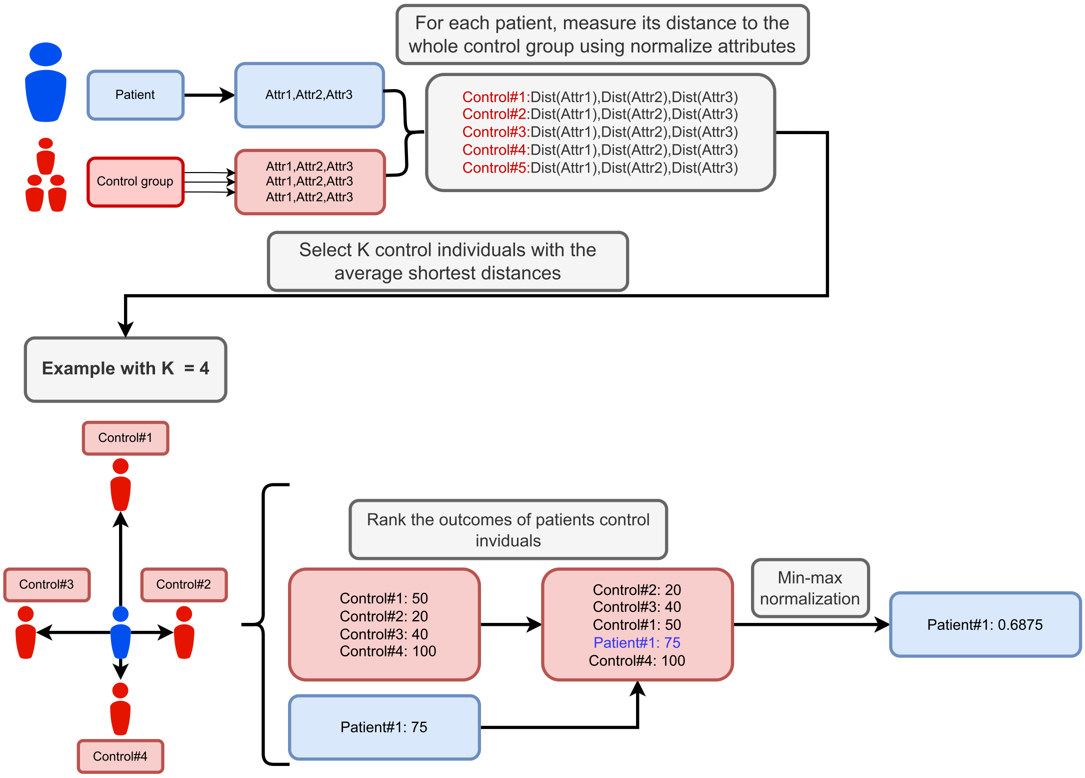
For each patient, measure its distance to the whole control group using normalize attributes
Patient
Attr1,Attr2,Attr3
Control group
Attr1,Attr2,Attr3
Attr1,Attr2,Attr3
Attr1,Attr2,Attr3
Control#1:Dist(Attr1),Dist(Attr2),Dist(Attr3)
Control#2:Dist(Attr1),Dist(Attr2),Dist(Attr3)
Control#3:Dist(Attr1),Dist(Attr2),Dist(Attr3)
Control#4:Dist(Attr1),Dist(Attr2),Dist(Attr3)
Control#5:Dist(Attr1),Dist(Attr2),Dist(Attr3)
Select K control individuals with the average shortest distances
Example with K = 4
Control#1
Control#3
Control#2
Control#4
Rank the outcomes of patients control inviduals
Control#1: 50
Control#2: 20
Control#3: 40
Control#4: 100
Control#2: 20
Control#3: 40
Control#1: 50
Patient#1: 75
Control#4: 100
Min-max normalization
Patient#1: 0.6875
Patient#1: 75

Fig. 1. Example of procedure to generate the dataset using only one outcome where the outcome of Patient 1 is transformed from 75 to 0.6875. The values obtained are considered the gold standard where higher values are always associated to worst outcomes.

adapted to consider the ranking aspect of the labels [24]. From a different perspective, label ranking tasks can also be decomposed into simpler prediction tasks. At first, the labels are divided into several groups, and a model is trained for each of them. Afterwards, the predictions of each model are combined into a final ranking. Despite well-fundamented [25], current local approaches in label ranking can result in an undesirable number of models, especially if the number of labels is too high.

The current state-of-the-art approach in label ranking, BoostLR [10], employs boosting with decision trees as base models where the loss function is adapted to label ranking. More specifically, the loss function is measured based on the number of swaps needed to obtain the correct label ranking, which is formalized using Kendall's Tau-b coefficient [26]. Despite being the state-of-the-art, this approach might struggle if the number of labels is too high.

Motivated by the relatively high number of labels of our application and also by the fact that our task is naturally divided in groups (Table 2), we propose a novel approach to label ranking. More specifically, we build multi-target models responsible to rank only the labels within a specific group, addressing, thus, each group separately. The rationale behind this is that labels within a specific group are expected to be more correlated, and superior performance can be achieved by making them explicit to the model. The final predictions are obtained by combining and ranking the output of each model in the solution.

Given the groups described in Table 2, we build models responsible for predicting the ranking of the labels from: i) intelligence, ii) visual-motor integration, iii) alertness, iv) motor coordination and v) memory. As for the underlying model, any off-the-shelf multi-target prediction algorithm is applicable. In this work, we employ multi-target random forests [27]. Random forests is an ensemble method that relies on multiple decision trees built on data sampled using bagging and random feature sampling per split. Further, they are robust, efficient, and often considered the state-of-the-art in problems with tabular datasets [28,29]. We call the resulting algorithm Label Ranking per Group of Outcomes (LaRGO). A representation of our approach is provided in Fig. 2.

2.3. Evaluation setup

In this section, we present the comparison approaches and the evaluation measures employed. All approaches were implemented using Python version 3.7.4 and the libraries Pandas 0.25.1, Numpy 1.17.2 and Scikit-learn 0.21.3. Graphs were generated using the Plotly library version 4.5.

2.3.1. Comparison approaches

We compared the performance of our proposed *LaRGO* machine learning approach to the following approaches, which are grouped into data driven and expert approaches.

Data driven approaches

- *LaRGO* machine learning approach: Our proposed approach (Fig. 2) which contains one random forest with 50 trees per group of outcomes shown in Table 2: i) Intelligence, ii) visual-motor integration, iii) alertness, iv) motor coordination and v) memory. Thus, our approach directly exploits the correlations within the groups of labels;
- *Global* machine learning approach: A single random forest with 50 trees that concurrently predicts all labels [27]. Thus, it is used to validate the idea of grouping labels related to the same neurocognitive outcome;
- *BoostLR* [10]: A recently proposed label ranking approach which relies on a single ensemble with 50 trees built using boosting. This approach is thus also a global approach and it is currently referred to as the state-of-art in label ranking. *BoostLR* adapts the loss function by using Kendall's Tau-b coefficient [26];

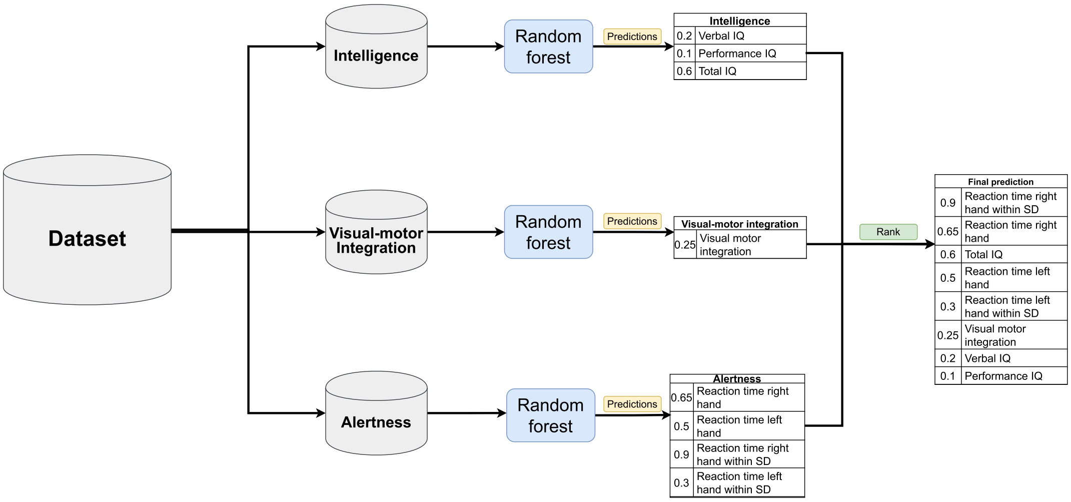
Dataset
Intelligence
Visual-motor Integration
Alertness
Random forest
Random forest
Random forest
Predictions
Predictions
Predictions
Intelligence
0.2 Verbal IQ
0.1 Performance IQ
0.6 Total IQ
Visual-motor integration
0.25 Visual motor integration
Alertness
0.65 Reaction time right hand
0.5 Reaction time left hand
0.9 Reaction time right hand within SD
0.3 Reaction time left hand within SD
Rank
Final prediction
0.9 Reaction time right hand within SD
0.65 Reaction time right hand
0.6 Total IQ
0.5 Reaction time left hand
0.3 Reaction time left hand within SD
0.25 Visual motor integration
0.2 Verbal IQ
0.1 Performance IQ

Fig. 2. Representation of our proposed approach using multi-target random forests as base models. In this example, we are exemplifying a case with only 3 groups of outcomes: Intelligence, visual-motor integration and alertness.

- Linear regression (*LinReg*) [30]: A baseline approach that employs one linear regression for each outcome, resulting in 23 models. The final prediction consists of sorting the outcomes provided by each model;
- *Mean* prediction: A baseline approach which predicts the mean ranking of the data (fixed sequence) for all patients;

Expert approaches These are the current approaches employed in neurocognitive outcome assessment which consistently follow a predefined order of tests. We have included two orders of tests.

- *Expert 1*: An *expert* approach which adopts the order of the PEPaNIC follow-up [2]: <*intelligence, visual motor integration, memory, alertness, and motor coordination*>;
- *Expert 2*: An *expert* approach which adopts the neurocognitive hierarchy order: <*intelligence, visual motor integration, alertness, motor coordination and memory*>;

For the expert approaches, the tests within each group of outcomes follow the same sequence presented in Table 2.

2.3.2. Evaluation measures

Since the evaluation of a patient might be interrupted at any given point, it is necessary to employ a metric capable of estimating performance regardless of the number of tests performed by the patient.

Hence, we propose to use Precision@N [31].[3] Given a value of N, Precision@N provides the proportion of correct predictions within the subset of top N predictions. Formally, Precision@N is defined in Equation (1) where TP stands for true positives, N is an integer related to the number of labels to be considered and X is the number of instances being evaluated.

$$Precision@N = \frac{1}{X}\sum_{1}^{X}\frac{TP_N}{N} \tag{1}$$

We provide an example in Table 3 considering a single instance and three outcomes, where the true ranking of outcomes is as follows: [Intelligence, Alertness, Memory], and the predicted ranking is [Alertness, Memory, Intelligence]. When N equals to 1, only the label predicted with highest priority (Alertness) is compared against the true label with highest priority (Intelligence). Since the prediction does not match, the value of the Precision@1 equals 0. Likewise, when $N = 2$, Precision@N evaluates the subset of (Alertness and Memory) against (IQ and Alertness). In this case, Precision@2 equals 0.5 since half of the labels evaluated were correctly predicted. Lastly, the Precision@3 = 1 since all three labels were correctly predicted.

The Precision@N measure is plotted in graphs where N ranges from 1 to 23 (the number of outcomes). The graphs allow to visualize the evolution of the performance for different values of N and also allow to compare the approaches for a fixed value of N (i.e., if we would perform 5 tests on the patients, how accurate would our test selection be?)

In a complementary manner, we can summarize these performance values in the Mean Average Precision (MAP, Equation (2)) value which averages the Precision@N values for all values of N.

$$MAP = \frac{1}{X}\sum_{1}^{X}\sum_{1}^{N} Precision@N \tag{2}$$

Lastly, inspired by the related field of information retrieval [32], we also propose two new evaluation measures: **A**verage **T**rue **R**anking (ATR) of the worst predicted outcome and **A**verage **P**redicted **R**anking (APR) of the worst true outcome. They both provide a value between 1 and N, where N is the number of labels. In both cases, lower values are associated to superior performance.

As its name suggests, ATR computes the true ranking of the outcome that was predicted as first. That is, it reflects the performance in scenarios where a patient is only able to perform one neurocognitive test. Following the example presented in Table 3, its ATR would be 2 since Alertness was predicted as the worst outcome and its true ranking is 2. Formally, ATR is presented in Equation (3) where $TrueRank1_x$ is the true ranking of the outcome predicted in the first position.

$$ATR = \frac{1}{X}\sum_{i}^{X} TrueRank1_x \tag{3}$$

Likewise, APR (Equation (4), where $PredictedRank1_x$ stands for the predicted ranking of the first relevant label for instance x), reports the average predicted ranking of the worst true outcome. More specifically, it measures the number of neurocognitive tests that would be sufficient to guarantee that the worse neurocognitive function is diagnosed.

$$APR = \frac{1}{X}\sum_{i}^{X} PredictedRank1_x \tag{4}$$

[3] In the literature, Precision@N is normally referred to as Precision@K. For the sake of readability, we have replaced N by K since K is a parameter in K-NN.

Table 3
Example of measurement of Precision@N up to $N = 3$.

N	Predicted	True Labels	Precision@N
1	Alertness	Intelligence	0
2	Alertness,Memory	Intelligence,Alertness	0.5
3	Alertness,Memory,Intelligence	Intelligence,Alertness,Memory	1

Still following the example from Table 3, the APR would be 3 since Intelligence is ranked in the third position in the predicted ranking and it is the worst outcome in the true ranking. Obtaining, for instance, an APR of 2.5 means that on average 2.5 tests are necessary to diagnose the worst outcome.

3. Results

We present our results in five separate sections:

- **Determining the optimal value of K:** We first find the optimal value of K to generate the dataset using the procedure described in Fig. 1;
- **Comparison of the different approaches:** Using the optimal K value determined in the previous step, we provide a comparison between the performance of machine learning approaches and expert approaches;
- **Age restricted analysis:** Due to some tests being age restricted, we also evaluate the performance of our approach on a subset of patients whose outcomes are completely known;
- **Inclusion of features from the 2 years follow-up:** We evaluate whether the inclusion of data available at the 2 years follow-up is beneficial;
- **Concrete case analysis:** We illustrate some concrete patient profiles where the approach employed in clinical practice would fail to detect the most severe neurocognitive problems if the assessment was ended early, while our approach would successfully identify them;

All experiments were repeated following a 10 × 5-fold cross validation procedure, resulting in 50 folds in total. We reported the Precision@N (Equation (1)) for all patients and possible values of N (1 to 23), the MAP (Equation (2)), the ATR (Equation (3)) and APR (Equation (4)), averaged over the 50 repetitions. All figures in this section are best viewed in colors.

Experiments evaluating two other subsets of patients: i) patients older than 4 and no memory related neurocognitive functions and ii) neurocognitive functions related only to intelligence and visual-motor integration, and the feature importance of all features are available in Appendix A and Appendix B, respectively.

3.1. Determining the optimal value of K

First, we investigated which value of K should be employed in this application. In order to do so, we have compared the following values of K {20, 100, 200, 300 and 405 (entire control group)} using separate validation sets (20% of each training fold) and our proposed *LaRGO* approach. Precisely, for each of the 50 training folds (10 × 5-fold cross validation), a separate subset of the training dataset is employed as validation set. We report the average MAP and Precision@N on these validation subsets.

As shown in Fig. 3, a rapid increase of performance can be noticed for increasing N, using all values of K, reaching precision values above 0.7 with just 4 outcomes. This is followed by a constant but less pronounced improvement as the number of evaluated outcomes raises.

Higher values of K (200, 300 and 405) are mostly associated with superior performance. A visible difference is already perceived with four outcomes, where $K = 405$ reaches approximately 80% of precision, whereas $K = 20$ and $K = 100$ are slightly over 75%. This behavior is maintained as the number of evaluated outcomes increases until the performance of all values of K converges. This is further reinforced by the MAP values presented in Table 4 where $K = 405$ always led to better results. This difference is, however, rather minimal, specially compared to $K = 200$ and $K = 300$.

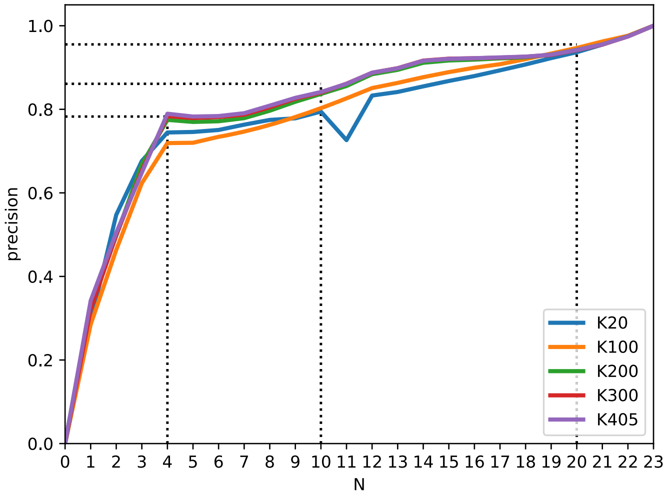

Fig. 3. Precision@N for different values of N obtained using our proposed approach with varying K values on validation sets. Each curve corresponds to the average obtained using the 10 × 5-fold cross validation.

Table 4
MAP obtained using our proposed approach and a separate validation dataset. Each value corresponds to the average obtained using the 10 × 5-fold cross validation. The symbol ⋆ indicates statistically significantly difference between $K = 405$ and $K = 300$ measured using the Wilcoxon signed-rank test (p-value < 0.05).

K	20	100	200	300	405
MAP	0.799	0.804	0.829	0.832	**0.834**⋆

Intuitively, we could expect that using a small number of controls would provide superior results since fewer and more similar control individuals are being used as comparison, nonetheless considering all the data available from the group consistently leads to superior results.

Hence, we suggest the use of higher values of K. More specifically, we recommend to set K equal to 405, which was used in the rest of the experiments, since the results slightly favor its usage over other values of K.

3.2. Comparison of the different approaches

Most of the data driven approaches are associated with superior results, as seen in Fig. 4. Similarly to the previous experiments, the performance of most of the machine learning approaches increases very swiftly with just a few labels.

This is related to the number of labels available per patient. As aforementioned, intelligence (3 tests) and visual motor-coordination (1 test) are the only neurocognitive functions available for children younger than 4, which comprises a considerable part of the data, as seen in Ta-

Table 5
Patients grouped according to their age, and the neurocognitive functions available per group.

Age	Neurocognitive functions available	Number of patients
Younger than 4	Intelligence and Visual-Motor Integration **(4)**	546
Between 4 and 5	Intelligence, Visual-Motor Integration, Alertness and Motor coordination **(12)**	37
Older than 5	Intelligence, Visual-Motor Integration, Alertness, Motor coordination and Memory **(23, all)**	203

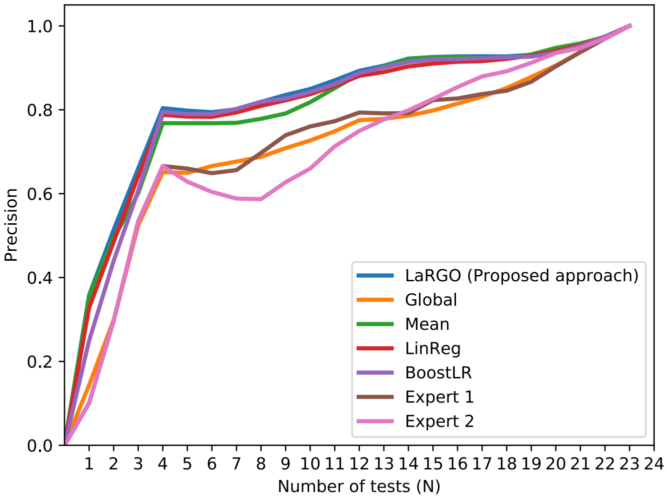

Fig. 4. Precision@N for different values of N obtained using all comparison approaches. Each curve corresponds to the average obtained using the 10 × 5-fold cross validation.

ble 5. Hence, when $N = 4$, all outcomes associated to those children are correctly predicted, leading to such rapid increase in performance.

Our *LaRGO* proposed approach is capable of providing the best results overcoming the current state-of-art in label ranking (*BoostLR*), its *Global* counterpart and also the baseline *LinReg*. Since it was specifically designed for this task, LaRGO exploits the correlations within each group of outcomes, leading to better performance.

Surprisingly, the baseline, *LinReg*, which was not designed for this type of predictive task, is also competitive. Despite being often employed as a simple comparison, it was only slightly outperformed in almost all cases. This small difference is further highlighted by the MAP values presented in Table 6, where *LaRGO* obtains 84% and *LinReg* 82%. However, *LaRGO* requires only 5 models, whereas *LinReg* relies on 23 models, one per outcome, making it more complex and its interpretation cumbersome, since combining the coefficients associated with each regressor may not be straightforward. Further, *LinReg* overlooks the correlation among the outcomes as models are totally independent from each other.

Additionally, the mean prediction should not be perceived as a deployable solution, since the same ranking is predicted for all patients, respecting the age restrictions of the tests, regardless of their features. Although reasonably competitive, its performance is associated with the distribution of the data. Tables C.20 and C.21 show that the *Mean* predicts labels that are more frequent. For instance, the top-1 outcome predicted by *Mean* prediction, *Intelligence: Total IQ*, is the first expected outcome in 41 patients which leads to Precision@1 = 26%, as the evaluated fold contains 158 patients. This value is further increased, since the *Mean* predicts *Motor coordination: Number of valid alternating taps* as the top-1 outcome for patients who are capable of undergoing this test, resulting into 23 more correct predictions and a total of Precision@1 = 41% (with some variation due to *Motor coordination* tests being age-restricted). This behavior persists when a higher number of outcomes is considered, leading to such competitive performance. It, however, fails to detect outcomes that are less frequent. It is noticeable that *LaRGO* provides a more diverse, an even personalized, prediction since outcomes as: *Intelligence: Performance IQ* and *Visual-motor integration*, are also often predicted.

Table 6
Mean average precision obtained using the values from Fig. 4. The symbol ⋆ indicates statistically significantly difference between *LaRGO* and *LinReg*, *Mean* and *Expert 1* measured using the Wilcoxon signed-rank test (p-value < 0.05).

LaRGO	Global	Mean	LinReg	BoostLR	Expert 1	Expert 2
0.84⋆	0.73	0.82	0.82	0.80	0.73	0.71

The state-of-art approach in label ranking, *BoostLR* [10], presented a competitive performance overall, but it was slightly outperformed by the three approaches discussed above. The *Global* approach was outperformed by all other machine learning approaches. Although this approach does have the advantage of requiring only a single model, concurrently addressing all 23 outcomes is seemingly challenging. We perceive this finding as an indication that local approaches are preferable in our application, and possibly in other label ranking applications where the number of outcomes is similar to ours.

The expert approaches struggled to be competitive. A sharp increase of performance with increasing N is also noticed in both approaches, nonetheless they are still underwhelming when compared to all the data driven approaches, except *Global*. Expert 1 outperforms Expert 2 in a handful of cases ($4 <= N <= 13$), nonetheless Expert 2 has the upperhand when a larger number of outcomes is considered. Similarly, according to MAP, both experts approaches performed poorly where Expert 1 scored 73% and Expert 2 (the approach currently employed in clinical practice) scored the lowest performance of 71%. The *Global* approach is the only machine learning approach which performs identically to Expert 2.

The main contrast in performance between experts and machine learning is noticed in ATR and APR. According to these measures, presented in Table 7, the *LaRGO* approach requires on average 3 tests to guarantee that the worst outcome is predicted (APR), whereas it would require approximately twice the number of outcomes, 6, for both expert approaches. Likewise, the ATR of the *LaRGO* and both expert approaches are also 3 and 6, respectively. Surprisingly, *BoostLR* and *LinReg* were outperformed by the *Mean* prediction.

We consider this is an indication that machine learning approaches not only tend to present a more personalized solution, but also a more effective one in terms of precision and also in terms of performance in comparison to the worst expected outcome.

Furthermore, the feature importance analysis (Table 8) reveals that the attribute *Age At Randomization* has a pivotal role. Similar findings were also reported by Verlinden et al. [33] where an interaction between age and deficiencies in several neurocognitive functions, such as memory, were identified, highlighting the age-dependent nature of the vulnerability to critical illness and its long-term impact. Additionally, the occupational and educational level of the parents seemed to be as relevant as the PIM3 which evaluates the severity of the illness at PICU admission. Also relevant are features describing the progression of the

Table 7
APR and ATR measured on all comparison approaches. We have reported the average obtained in each of the 10 × 5-fold cross validation. The symbol ⋆ indicates statistically significantly difference between *LaRGO* and *LinReg*, *Mean* and *Expert 1* measured using the Wilcoxon signed-rank test (p-value < 0.05).

	LaRGO	Global	Mean	LinReg	BoostLR	Expert 1	Expert 2
ATR	**3.02**⋆	5.24	3.11	3.16	3.47	6.23	6.23
APR	**3.29**⋆	6.03	3.79	3.58	3.58	6.41	6.41

Table 8
Feature importance averaged on the random forests used in the LaRGO approach. Here, we present only the top 10 features. A complete list is available in Appendix B.18.

Features	Importance
Age at randomization	83.48
Diagnostic category	23.50
Occupational level parents	18.40
Educational level parents	17.37
PIM3 score	17.17
Center	12.15
Duration of stay in the PICU	9.76
PeLOD score first 24 hrs	9.59
Duration of treatment with benzodiazepines	8.20
Duration of treatment with antibiotics	8.19

illness, such as the PELOD score for the first 24 hours and the overall duration of the stay in the PICU. Other important features pertain to the duration and need of treatments such as antibiotics for new infections or benzodiazepines, which were previously found to be associated with poorer neurological outcomes [2].

3.3. Age restricted analysis

Considering that 3 groups of tests (alertness, motor coordination and memory) are only available for children who are older than 5, we have evaluated the Precision@N considering only this cohort. In this case, we have reported the results for patients older than 5 at follow-up in each of the 50 test folds.

When compared to the previous experiments, this task is more challenging since all 23 targets are evaluated for all patients, whereas the results reported before contain children with only 4 and 12 known outcomes (younger than 5 years) due to age-restricted tests, as shown in Table 5.

As can be seen in Fig. 5, the previously observed rapid improvement is less accentuated in this case. More specifically, the precision of all approaches is considerably inferior. In the previous experiments, a precision of 80% is reached within 4 labels, whereas, in this case, it requires about 10 labels. On a similar fashion, the performance of *BoostLR* deteriorates significantly, becoming comparable to both Expert approaches.

Consequently, the values of the MAP measure (Table 9) also decrease, where *LaRGO* presented 74%, whereas Expert 1 and 2 achieved merely 58% and 54%, respectively.

According to ATR and APR (Table 10), both Expert 1 and 2 present poor performance (11 for both evaluation measures). These values are rather surprising, since it means that the experts seldom recommend the outcome expected to be most affected to be evaluated first. The *LaRGO* approach, however, manages to provide more desirable results with 4.56 and 4.98 for ATR and APR, respectively.

Despite these differences, the performance of all compared approaches is analogous to the previous experiments. That is, machine learning approaches tend to provide better performance, whereas human approaches achieve less prominent results. Hence, machine learning approaches are capable of performing satisfactorily in this subset of the cohort as well. Similar conclusions can be drawn when analyzing other subsets (children older than 4 years excluding outcomes related to memory and excluding outcomes related to alertness, motor coordination and memory), as shown in Appendix A.

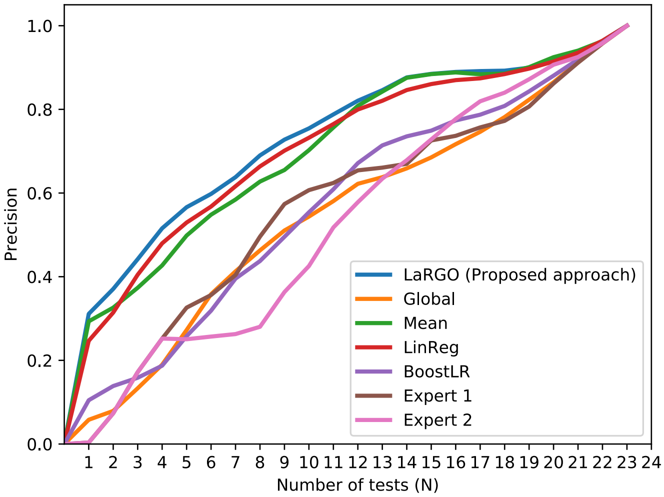

Fig. 5. Precision@N for different values of N obtained on patients older than 5 at follow-up, using all comparison approaches. Each curve corresponds to the average obtained using the 10 × 5-fold cross validation.

Table 9
Mean average precision obtained using the values from Fig. 5 (children older than 5). The symbol ⋆ indicates statistically significantly difference between *LaRGO* and *LinReg*, *Mean* and *Expert 1* measured using the Wilcoxon signed-rank test (p-value < 0.05).

LaRGO	Global	Mean	LinReg	BoostLR	Expert 1	Expert 2
0.74⋆	0.56	0.72	0.72	0.58	0.58	0.54

3.4. Inclusion of features from the 2 years follow-up

Since the *LaRGO* approach yielded superior performance in the previous experiments, we have used it to evaluate whether the inclusion of the features available at the 2 years follow-up improves performance.

More precisely, we have compared the Precision@N of the *LaRGO* approach using the two versions of the datasets discussed in Section 2.1: i) Discharge features and ii) Discharge features and 2 years follow-up features. We report results using all children in the datasets and using 10 × 5-fold cross-validation.

Despite containing relevant information, the inclusion of the CBCL and BRIEF values as features does not lead to a distinguishable improvement in performance, as shown in Fig. 6. Regardless of the number of outcomes considered, both solutions present overlapping results. These identical performances reinforce our results by showing that, the features available at discharge are exceptionally representative in this task to the point that they match the predictive power of features available at 2 years follow-up.

Although we observe equivalent performance, the feature importance (Table 11) reveals that the most important features differ from the previous experiments (Table 8). More specifically, the CBCL and BRIEF outcomes replace the attributes: *Center*, *duration of stay in the*

Table 10
APR and ATR measured on children older than 5 and all comparison approaches. We have reported the average obtained in each of the 10 × 5-fold cross validation. The symbol ⋆ indicates statistically significantly difference between *LaRGO* and*LinReg*, *Mean* and *Expert 1* measured using the Wilcoxon signed-rank test (p-value < 0.05).

	LaRGO	Global	Mean	LinReg	BoostLR	Expert 1	Expert 2
ATR	**4.56***	8.90	4.78	4.79	11.71	11.15	11.15
APR	**4.98***	11.07	6.31	5.70	9.84	11.74	11.74

Table 11
Feature importance averaged on the random forests used in the LaRGO approach built using features available at discharge and features from the 2 years follow-up. A complete list is available in Table B.19.

Feature	Importance
Age at randomization	77.28
Diagnostic category	21.21
Occupational level parents	16.61
Executive functioning: Working memory as reported by parents/caregivers	16.26
Executive functioning: Flexibility as reported by parents/caregivers	14.57
Educational level parents	13.64
PIM3 score	13.41
Executive functioning: Inhibition as reported by parents/caregivers	12.63
Emotional/Behavioral: Internalizing problems as reported by parents/caregivers	11.94

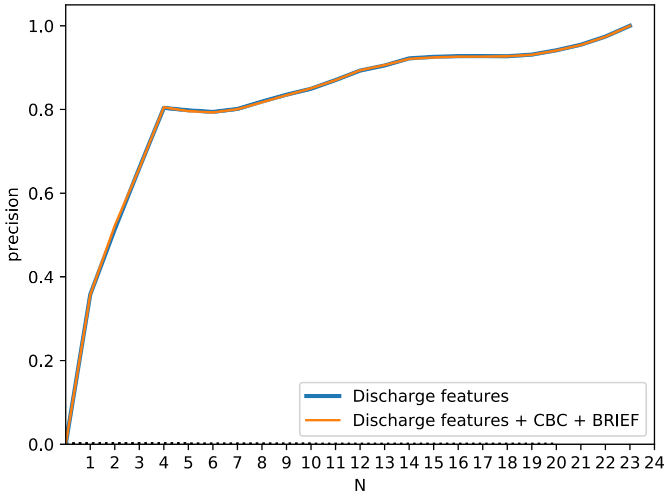

Fig. 6. Precision@N for different values of N obtained on the *LaRGO* approach using two different versions of the dataset. Each curve corresponds to the average obtained using the 10 × 5-fold cross validation.

PICU, *PeLOD score in the first 24 hours* and *duration of treatment with benzodiazepines*. On a practical level, this indicates that the CBCL and BRIEF outcomes may be used as a replacement for these features in the very exceptional cases where they are not available, since their inclusion is not associated with a deterioration in performance.

3.5. Concrete case analysis

In this section, we present concrete cases where clinical practice (Expert 1) would fail to detect the most severe neurocognitive problems if the assessment ended prematurely. Concretely, we have analyzed cases where Expert 1 would require more than 20 tests to predict the top-5 worst outcomes in patients older than 5 (60 patients).

When performing this analysis, we could only identify 8 (13%) cases where the *LaRGO* approach would fail to predict the top-5 outcomes in its top-20 predictions. As opposed to that, when employing Expert 1 approach, our analysis led to 43 patients. That means that, in approximately 71% of analyzed cases, the outcomes expected to be worst would plausibly not be diagnosed by Expert 1, since the patients would have to undergo at least 20 tests. For better assessment, we investigate 3 of these patients further in Tables 12 and 13.

For Patient#1, we could notice several outcomes related to Motor coordination. More specifically, there are 4 outcomes related to the number of hands taps. Similarly, Patient #2 also presents several outcomes related to Motor coordination. Lastly, the intelligence of Patient #3 is the most affected one, since all three of its tests are expected to be in the worst 5.

Additionally, Table 13 reveals that features of the patients also present differences where their *Age at randomization* corresponds to 7.23, 4.38 and 9.58, respectively. Patient #1 was admitted for abdominal surgery while Patients #2 and #3 were admitted for cardiac surgery. Furthermore, *Occupational level parents* and *Educational levels parents* also differ and it is the lowest in Patient #3. Of the three examples, Patient #3 was the sickest, requiring prolonged critical care after cardiac surgery with a PICU stay of 7 days, further needing treatment with benzodiazepines in addition to antibiotics for the majority of the stay. The severity of illness and the intensity and duration of the required therapy are features that in part explain the prediction for low overall intelligence scores for this patient at follow-up

3.6. Limitations

A limitation of the study is that despite a very high follow-up rate, many children were too young at the time of follow-up and thus not amenable to evaluation of all neurocognitve tests. The statistical power for the tests only possible in the older children was therefore reduced.

4. Discussion

We have presented the first step in developing a more focused ICU follow-up model for neurocognitive assessment for children, where unnecessary tests are omitted while important developmental problems can still be detected early. If such assessment is to be implemented in clinical practice, practical challenges would include establishing and maintaining direct communication between the different healthcare disciplines involved such as: intensivists, ICU nurses, pediatricians, and psychologists, among others. Besides these, introducing any software in the ICU requires many IT resources, a.o. to guarantee safety of patients' data and to adhere to legislation (e.g. GDPR and the AI Act).

Table 12
Top-5 outcomes of 3 patients where the Expert 1 approach would require more than 20 tests to identify the top-5 worst outcomes. In this analysis, we could observe that 8 (13%) patients were identified using the *LaRGO* approach, whereas 43 (71%) were detected using the Expert 1 approach. In the first column, #LaRGO represents the number of tests required by the LaRGO approach to identify the top-5 worst outcomes and #Expert 1 corresponds the number of tests required by Expert 1 to identify the top-5 worst outcomes. The second column (Top 5 Worst Outcomes) contains the top-5 worst outcomes. Lastly, the third column (Top 5 Predicted Outcomes) contains the top-5 predicted by the LaRGO approach.

	Top 5 Worst Outcomes	Top 5 Predicted Outcomes
Patient#1 #LaRGO = 5 #Expert 1 = 23	Visual-motor integration Motor coordination: Number of valid alternating taps Motor coordination: Number of right hand taps Motor coordination: Number of left hand taps Motor coordination: Number of valid synchronous tap	Motor coordination: Number of valid alternating taps Motor coordination: Number of right hand taps Visual-motor integration Motor coordination: Number of left hand taps Motor coordination: Number of valid synchronous taps
Patient#2 #LaRGO = 7 #Expert 1 = 22	Motor coordination: Number of valid alternating taps Memory: Verbal-auditory, word pairs immediate memory Motor coordination: Number of right hand taps Motor coordination: Number of left hand taps Memory: Verbal-auditory, word pairs delayed memory	Motor coordination: Number of valid alternating taps Memory: Verbal-auditory, word pairs immediate memory Motor coordination: Number of right hand taps Memory: Learning index Motor coordination: Number of left hand taps
Patient#3 #LaRGO = 6 #Expert 1 = 22	Intelligence: Total IQ Intelligence: Performance IQ Intelligence: Verbal IQ Visual-motor integration Motor coordination: Number of valid alternating taps	Motor coordination: Number of valid alternating taps Visual-motor integration Intelligence: Performance IQ Intelligence: Total IQ Motor coordination: Number of right hand taps

Table 13
Most relevant features, according to the feature importance (Table 8), of three patients evaluated in the concrete case analysis.

Feature	Patient#1	Patient#2	Patient#3
Age at randomization	7.23	4.38	9.58
Diagnostic category	Surgical Abdominal	Surgical Cardiac	Surgical Cardiac
Occupational level parents	3,5	2,5	2
Educational level parents	3	3,5	1,5
PIM3 score	0	32	31
Center	Leuven	Leuven	Leuven
Duration of stay in the PICU	1	2	7
PeLOD score first 24 hrs	-4.55	-3.98	-3.12
Duration of treatment with benzodiazepines	0	0	2
Duration of treatment with antibiotics	1	1	6

Despite of that, it is clear that any reduction in the number of tests, would reduce the psychological and socio-economic burden on the families of the patients and on the healthcare system, provided that the detection of the adverse outcomes is not compromised. Additionally, patients and caregivers would experience less stress during the evaluation procedure.

In this work, we have addressed this challenge, by predicting a personalized ranking of tests, such that if the tests are performed in the predicted order, the worst expected outcomes will be evaluated first. More specifically, we have used a real dataset with neurocognitive outcomes measured 2 years after hospitalization, for participants from the PEPaNIC trial. This trial consisted of a large international study carried out in Belgium, The Netherlands and Canada. Thus, increasing the likelihood that our findings will generalize to other centers as well. To model this problem, we have first transformed the dataset to change absolute outcome values into relative measures, by using information from a healthy control set. Second, we have proposed a new label ranking approach that builds a multi-target prediction model per group of tests, and combines their outputs by ranking them into a final prediction.

Our results have revealed that our proposed approach is able to reach excellent performance, even outperforming current clinical practice as recommended by experts. We have also shown that our approach slightly outperforms other machine learning based approaches, with higher differences visible in age-restricted subset analyses. Such better predictive results could lead to a better assessment of the patient condition. In practice, this means that, in the likely case that the evaluation procedure must be interrupted prematurely, more adverse outcomes would be tested, since our proposed method is able to identify a higher number of adverse outcomes in comparison to the current clinical practice.

We could also notice that features available at discharge are already highly informative in this context, despite the 2 years gap between features and outcome measurement. This allows the practice of preventive care, starting immediately at discharge, rather than curative after possible follow-up. Consequently, healthcare related costs will be reduced as curative care is substantially more expensive than preventive. Additionally, specific patients can be prioritized to receive better, necessary and more personalized care. Further, this tailored care will develop the ongoing follow-up care programs available for some specific populations [34] to a broader individualized follow-up program for all critically ill children. This will ultimately have a positive effect on societal participation of critically ill children at risk.

5. Conclusion

We have presented the first work to address adverse long-term neurocognitive outcomes assessment after pediatric intensive care unit admission using machine learning. More specifically, we have shown that data driven methods can surpass the current clinical practice employed by experts, which could possibly lead to more personalized care and consequently better life quality after discharge.

Further, we also aim to incorporate data from the 4 years follow-up study [35]. More specifically, we will employ the data from the 2 years follow-up to enhance the predictions for the outcomes at 4 years. This will require the adaptation of the LaRGO method introduced here. Such future studies will enable a comprehensive and realistic assessment of the progress of the neurocognitive function outcomes, possibly leading to the prediction of the whole trajectory of the patient after discharge. For the purpose of further external validation we will additionally explore the use data from similar studies outside the scope of PEPaNIC trial, such as [36].

Furthermore, developing a new heuristic for the random forests, which explicitly focuses on the ranking of the data, could improve the results further. We believe that other applications, which also require the prediction of a ranking of outcomes, could benefit from the methodology developed in this work. Lastly, we would like to implement our model at discharge and at follow-up. Our ultimate aim is to develop an application that will provide, at time of discharge from the PICU, a proposal for a focused patient-centered follow-up. This personalized follow-up should constantly be monitored during the growth and development of children since interventions and new life-events during follow-up might influence the outcomes. Consequently, specific patients could be selected for additional dedicated care, tutoring and psychological help. Future studies should also clearly delineate the practical challenges of implementing our approach.

CRediT authorship contribution statement

Felipe Kenji Nakano: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Conceptualization. **Karolijn Dulfer:** Conceptualization, Formal analysis, Funding acquisition, Supervision, Writing – original draft. **Ilse Vanhorebeek:** Formal analysis, Funding acquisition, Supervision, Writing – original draft. **Pieter J. Wouters:** Data curation, Formal analysis, Funding acquisition, Writing – original draft. **Sascha C. Verbruggen:** Writing – original draft, Funding acquisition, Formal analysis, Conceptualization. **Koen F. Joosten:** Writing – original draft, Formal analysis, Data curation, Conceptualization. **Fabian Güiza Grandas:** Writing – review & editing, Writing – original draft, Supervision, Resources, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Celine Vens:** Writing – review & editing, Writing – original draft, Supervision, Software, Resources, Methodology, Investigation, Funding acquisition, Conceptualization. **Greet Van den Berghe:** Writing – original draft, Supervision, Resources, Project administration, Conceptualization.

Declaration of competing interest

All authors declare that there is no conflict of interest.

Acknowledgements

This work was supported by the Research Fund Flanders (through research projects G080118N, 1235924N and G046024N to FKN and CV); by the Flemish Government (AI Research Program) to FKN and CV; by European Research Council Advanced Grants (AdvG-2012-321670 from the Ideas Program of the European Union 7th framework program and AdvG-2017-785809 from the Horizon 2020 Program to GVdB); by the Methusalem program of the Flemish Government (through the University of Leuven to GVdB, METH/08/07 and to GVdB and IV, METH14/06); by the Institute for Science and Technology, Flanders, Belgium (through the University of Leuven to GVdB, IWT-TBM150181 and IWT-TBM110685); by the Sophia Research Foundation (SSWO to SV); by the Stichting Agis Zorginnovatie (to SV); Nutricia Research B.V (to SV); by the Erasmus Trustfonds (to SV); and by an European Society for Clinical Nutrition and Metabolism (ESPEN) research grant (to SV).

Appendix A. Results different subsets

See Tables A.14–A.17 and Figs. A.7, A.8.

Table A.14
Mean average precision obtained using the values from Fig. A.7 (children older than 4 excluding outcomes related to memory).

LaRGO	Global	Mean	LinReg	BoostLR	Expert 1	Expert 2
0.72	0.63	0.67	**0.72**	0.65	0.59	0.59

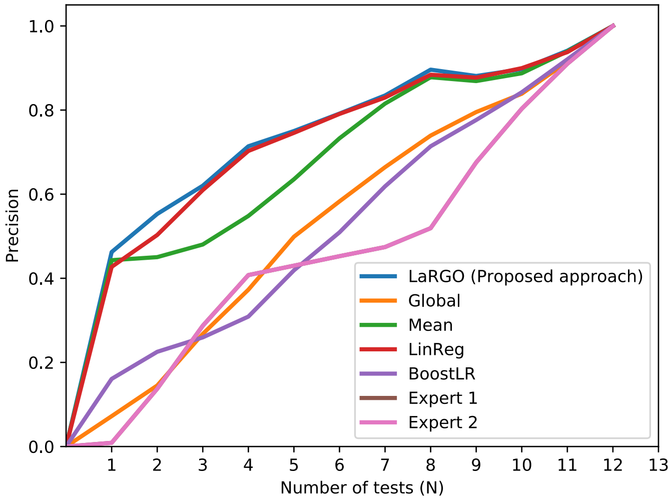

Fig. A.7. Precision@N obtained on children older than 4 and excluding outcomes related to memory (not available), using all comparison approaches and $N = (1,12)$. Each curve corresponds to the average obtained in each of the 10 × 5-fold cross validation. In this case, Experts 1 and 2 predict the same order since outcomes related to memory are not present.

Table A.15
APR and ATR measured on children older than 4, excluding outcomes related to memory. We have reported the average obtained in each of the 10 × 5-fold cross validation.

	LaRGO	Global	Mean	LinReg	BoostLR	Expert 1	Expert 2
ATR	**2.63**	5.69	2.78	2.66	11.06	6.83	6.83
APR	**2.70**	5.49	3.63	2.76	9.01	7.80	7.80

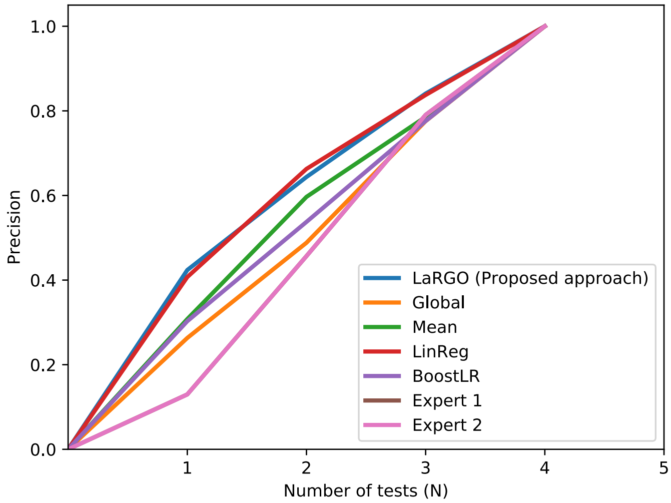

Fig. A.8. Precision@N obtained when outcomes related to alertness, motor coordination and memory are excluded, using all comparison approaches and $N = (1,4)$. Each curve corresponds to the average obtained in each of the 10 × 5-fold cross validation. In this case, Experts 1 and 2 predict the same order since outcomes related to alertness, motor coordination and memory are not present.

Table A.16
APR and ATR measured on all comparison approaches excluding outcomes related to alertness, motor coordination and memory. We have reported the average obtained in each of the 10 × 5-fold cross validation.

	LaRGO	Global	Mean	LinReg	BoostLR	Expert 1	Expert 2
ATR	**2.03**	2.36	2.15	2.08	3.47	2.87	2.87
APR	**2.09**	2.57	2.26	2.11	3.58	2.81	2.81

Table A.17
Mean average precision using the values obtained from Fig. A.8 (excluding targets related to alertness, motor coordination and memory).

LaRGO	Global	Mean	LinReg	BoostLR	Expert 1	Expert 2
0.72	0.63	0.67	**0.72**	0.65	0.59	0.59

Table B.18
Complete feature importance list using only features available at discharge.

Features	Importance
Age at randomization	83.48
Diagnostic category	23.5
Occupational level parents	18.4
Educational level parents	17.37
PIM3 score	17.17
Center	12.15
Duration of stay in the PICU	9.76
PeLOD score first 24 hrs	9.59
Duration of treatment with benzodiazepines	8.2
Duration of treatment with antibiotics	8.19
Duration of treatment with opioids	8.07
Duration of mechanical ventilatory support	7.33
Duration of treatment with hypnotics	6.87
Syndrome vs. no syndrome	6.27
Duration of hemodynamic support	5.6
Known non-European origin	5.58
Duration of treatment with corticosteroids	5.56
Parental smoking during pregnancy	3.79
Hand preference	3.58
Test location	3.27
Parental smoking between birth and PICU admission	3.18
Duration of treatment with alpha-2-agonists	3.15
Sex	2.91
Maternal smoking pre-pregnancy	2.52
Randomization to late vs. early initiation of PN	2.3
Known not exclusive Dutch or English language	2.29
Parental smoking pre-pregnancy	1.84
Malignancy vs. no malignancy	1.44
STRONGkids risk level	1.37
Known non-Caucasian	1.2
New infection	1.17
Maternal smoking during pregnancy	1.05
Days with hypoglycemic event	0.81
Hypoglycemia < 40 mg/dl	0.68
Diabetes vs. no diabetes	0.02

Appendix B. Feature importance

See Tables B.18 and B.19.

Appendix C. Comparing the predictions obtained the *LaRGO* and *Mean* approach

See Tables C.20 and C.21.

Table B.19
Complete feature importance list using features available at discharge and features available at the 2 years follow-up.

Feature	Importance
Age at randomization	77.28
Diagnostic category	21.21
Occupational level parents	16.61
Executive functioning: Working memory as reported by parents/caregivers	16.26
BRIEFallagesTFlexibility	14.57
Executive functioning: Flexibility as reported by parents/caregivers	13.64
PIM3 score	13.41
Executive functioning: Inhibition as reported by parents/caregivers	12.63
Emotional/Behavioral: Internalizing problems as reported by parents/caregivers	11.94
Executive functioning: Overall as reported by parents/caregivers	11.89
Executive functioning: Emotional control as reported by parents/caregivers	11.76
Emotional/Behavioral: Overall problems as reported by parents/caregivers	11.64
Executive functioning: Meta-cognition index as reported by parents/caregivers	11.36
Center	10.91
Executive functioning: Planning and organization as reported by parents/caregivers	10.42
Emotional/Behavioral: Externalizing problems as reported by parents/caregivers	9.54
Duration of stay in the PICU	7.16
PeLOD score first 24 hrs	7.11
Duration of treatment with benzodiazepines	6.26
Duration of mechanical ventilatory support	6.07
Duration of treatment with antibiotics	5.76
Duration of treatment with opioids	5.58
Duration of treatment with hypnotics	4.92
Known non-European origin	4.68
Duration of treatment with corticosteroids	4.41
Duration of hemodynamic support	4.09
Syndrome vs. no syndrome	3.29
Hand preference	2.67
Parental smoking during pregnancy	2.43
Test location	2.34
Parental smoking between birth and PICU admission	2.27
Duration of treatment with alpha-2-agonists	2.12
Sex	1.99
Maternal smoking pre-pregnancy	1.86
Randomization to late vs. early initiation of PN	1.85
Known not exclusive Dutch or English language	1.61
Parental smoking pre-pregnancy	1.32
Malignancy vs. no malignancy	1.22
Known non-Caucasian	0.98
New infection	0.86
STRONGkids risk level	0.85
Maternal smoking during pregnancy	0.66
Days with hypoglycemic event	0.58
Hypoglaemica < 40 mg/dl	0.48
Diabetes vs. no diabetes	0.01

References

[1] J.A. Hordijk, S.C. Verbruggen, C.M. Buysse, E.M. Utens, K.F. Joosten, K. Dulfer, Correction to: Neurocognitive functioning and health-related quality of life of children after pediatric intensive care admission: a systematic review, Qual. Life Res. (May 2022), https://doi.org/10.1007/s11136-022-03144-9.

[2] S. Verstraete, S.C. Verbruggen, J.A. Hordijk, I. Vanhorebeek, K. Dulfer, F. Güiza, E. van Puffelen, A. Jacobs, S. Leys, A. Durt, et al., Long-term developmental effects of withholding parenteral nutrition for 1 week in the paediatric intensive care unit: a 2-year follow-up of the pepanic international, randomised, controlled trial, Lancet Respir. Med. 7 (2) (2019) 141–153.

[3] S.Y. Kim, S. Kim, J. Cho, Y.S. Kim, I.S. Sol, Y. Sung, I. Cho, M. Park, H. Jang, Y.H. Kim, et al., A deep learning model for real-time mortality prediction in critically ill children, J. Crit. Care 23 (1) (2019) 1–10.

[4] R. Kamaleswaran, O. Akbilgic, M.A. Hallman, A.N. West, R.L. Davis, S.H. Shah, Applying artificial intelligence to identify physiomarkers predicting severe sepsis in the PICU, Pediatr. Crit. Care Med. 19 (10) (2018) e495–e503.

[5] B. Matam, H. Duncan, D. Lowe, Machine learning based framework to predict cardiac arrests in a paediatric intensive care unit, J. Clin. Monit. Comput. 33 (4) (2019) 713–724.

[6] J. Dong, T. Feng, B. Thapa-Chhetry, B.G. Cho, T. Shum, D.P. Inwald, C.J. Newth, V.U. Vaidya, Machine learning model for early prediction of acute kidney injury (AKI) in pediatric critical care, J. Crit. Care 25 (1) (2021) 1–8.

[7] S. Vembu, T. Gärtner, Label ranking algorithms: A survey, in: Preference Learning, Springer, 2010, pp. 45–64.

Table C.20
Count of the most frequent true outcomes (Labels), predictions provided by the *Mean* prediction and predictions provided by the *LaRGO* approach, considering the Top 1 and 2 outcomes in a single fold, presented in a non-accumulative manner. Despite always predicting a fixed order, the *Mean* prediction presents more than one outcome because it respects the age restrictions of the outcomes.

Labels			
Top#1	Count	Top#2	Count
Intelligence: Total IQ	41	Intelligence: Verbal IQ	34
Visual-motor integration	29	Intelligence: Total IQ	25
Motor coordination: Number of valid alternating taps	23	Intelligence: Performance IQ	24
Intelligence: Performance IQ	16	Visual-motor integration	17
Mean Prediction			
Intelligence: Total IQ	85	Intelligence: Performance IQ	85
Motor coordination: Number of valid alternating taps	73	Memory: Verbal-auditory, word pairs immediate memory	52
		Motor coordination: Number of right hand taps	21
LaRGO			
Motor coordination: Number of valid alternating taps	67	Intelligence: Performance IQ	58
Intelligence: Total IQ	52	Memory: Verbal-auditory, word pairs immediate memory	28
Intelligence: Performance IQ	20	Motor coordination: Number of right hand taps	23
Visual-motor integration	16	Intelligence: Total IQ	18

Table C.21
Count of the most frequent true outcomes (Labels), predictions provided by the *Mean* prediction and predictions provided by the *LaRGO* approach, considering the Top 3 and 4 outcomes in a single fold, presented in a non-accumulative manner. Despite always predicting a fixed order, the *Mean* prediction presents more than one outcome because it respects the age restrictions of the outcomes.

Labels			
Top#3	Count	Top#4	Count
Intelligence: Performance IQ	43	Visual-motor integration	43
Intelligence: Total IQ	26	Intelligence: Verbal IQ	25
Intelligence: Verbal IQ	19	Intelligence: Performance IQ	18
Visual-motor integration	16	Intelligence: Total IQ	12
Mean			
Visual-motor integration	85	Intelligence: Verbal IQ	85
Memory: Learning index	52	Motor coordination: Number of right hand taps	52
Motor coordination: Number of left hand taps	21	Motor coordination: Number of valid synchronous taps	21
LaRGO			
Intelligence: Verbal IQ	38	Intelligence: Verbal IQ	44
Memory: Learning index	26	Visual-motor integration	42
Visual-motor integration	23	Intelligence: Total IQ	13
Intelligence: Total IQ	23	Memory: Learning index	12

[8] J. Mantas, et al., An approach based on preference learning for identifying experts reasoning in antibiotic treatment, in: The Importance of Health Informatics in Public Health During a Pandemic, vol. 272, 2020, p. 115.

[9] R. Tsopra, J.-B. Lamy, K. Sedki, Using preference learning for detecting inconsistencies in clinical practice guidelines: methods and application to antibiotherapy, Artif. Intell. Med. 89 (2018) 24–33.

[10] L. Dery, E. Shmueli, BoostLR: a boosting-based learning ensemble for label ranking tasks, IEEE Access 8 (2020) 176023–176032.

[11] T. Fivez, D. Kerklaan, D. Mesotten, S. Verbruggen, P.J. Wouters, I. Vanhorebeek, Y. Debaveye, D. Vlasselaers, L. Desmet, M.P. Casaer, et al., Early versus late parenteral nutrition in critically ill children, N. Engl. J. Med. 374 (12) (2016) 1111–1122.

[12] K. Van der Heijden, J. Suurland, L. De Sonneville, H.B.-P. Swaab, Vragenlijst voor executieve functies voor 2-tot 5-jarigen: Handleiding, Hogrefe, Amsterdam, 2013.

[13] M. Huizinga, D. Smidts, BRIEF Vragenlijst executieve functies voor 5-tot 18-jarigen: Handleiding, Hogrefe Uitgevers, 2012.

[14] T.M. Achenbach, L.A. Rescorla, Manual for the ASEBA Preschool Forms and Profiles, vol. 30, University of Vermont, Research Center for Children, Youth ..., Burlington, VT, 2000.

[15] F.C. Verhulst, J. van der Ende, Handleiding ASEBA: vragenlijsten voor leeftijden 6 t/m 18 jaar: CBCL/6-18, YSR, TRF, ASEBA Nederland, 2013.

[16] J. Hendriksen, P. Hurks, WPPSI-III-NL Wechsler Preschool and Primary Scale of Intelligence; Nederlandse bewerking, Pearson, 2009.

[17] D. Wechsler III, WISC-III Nederlandstalige bewerking, Handleiding, 2005.

[18] D. Wechsler, WAIS-IV-NL: Wechsler Adult Intelligence Scale–Nederlandstalige bewerking, Pearson, 2012.

[19] K.E. Beery, V.MI. Beery, The Beery-Buktenica Developmental Test of Visual-Motor Integration with Supplemental Developmental Tests of Visual Perception and Motor Coordination: AND, Stepping Stones Age Norms from Birth to Age Six. Administration, Scoring, and Teaching Manual, PsychCorp, New York, NY, 2010.

[20] L.d. Sonneville, Handboek Amsterdamse Neuropsychologische Taken (ANT), 2014.

[21] M. Cohen, Children's Memory Scale (CMS), The Psychological Corporation, San Antonio, 1999.

[22] S. Van Buuren, K. Groothuis-Oudshoorn, MICE: Multivariate imputation by chained equations in r, J. Stat. Softw. 45 (2011) 1–67.

[23] S. Zhang, X. Li, M. Zong, X. Zhu, R. Wang, Efficient KNN classification with different numbers of nearest neighbors, IEEE Trans. Neural Netw. Learn. Syst. 29 (5) (2017) 1774–1785.

[24] W. Cheng, J. Hühn, E. Hüllermeier, Decision tree and instance-based learning for label ranking, in: Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 161–168.

[25] E. Hüllermeier, J. Fürnkranz, Comparison of ranking procedures in pairwise preference learning, in: Proceedings of the 10th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU-04), Perugia, Italy, vol. 21, 2004.

[26] M.G. Kendall, The treatment of ties in ranking problems, Biometrika 33 (3) (1945) 239–251.

[27] D. Kocev, C. Vens, J. Struyf, S. Džeroski, Tree ensembles for predicting structured outputs, Pattern Recognit. 46 (3) (2013) 817–833.

[28] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[29] L. Grinsztajn, E. Oyallon, G. Varoquaux, Why do tree-based models still outperform deep learning on typical tabular data?, Adv. Neural Inf. Process. Syst. 35 (2022) 507–520.

[30] S. Weisberg, Applied Linear Regression, vol. 528, John Wiley & Sons, 2005.

[31] W. Krichene, S. Rendle, On sampled metrics for item recommendation, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 1748–1757.

[32] D.R. Radev, H. Qi, H. Wu, W. Fan, Evaluating web-based question answering systems, in: LREC, Citeseer, 2002.

[33] I. Verlinden, K. Dulfer, I. Vanhorebeek, F. Güiza, J.A. Hordijk, P.J. Wouters, G.G. Guerra, K.F. Joosten, S.C. Verbruggen, G. Van den Berghe, Role of age of critically ill children at time of exposure to early or late parenteral nutrition in determining the impact hereof on long-term neurocognitive development: A secondary analysis of the pepanic-RCT, Clin. Nutr. 40 (3) (2021) 1005–1012.

[34] Nederlandse vereniging voor kindergeneeskunde: Follow-up van kinderen na opname op een intensive care, https://www.nvk.nl/themas/kwaliteit/richtlijnen/richtlijn?componentid=6881283&tagtitles=Intensive%252bCare. (Accessed 26 January 2022), 2021.

[35] A. Jacobs, K. Dulfer, R.D. Eveleens, J. Hordijk, H. Van Cleemput, I. Verlinden, P.J. Wouters, L. Mebis, G.G. Guerra, K. Joosten, et al., Long-term developmental effect of withholding parenteral nutrition in paediatric intensive care units: a 4-year follow-up of the pepanic randomised controlled trial, Lancet Child Adolesc. Health 4 (7) (2020) 503–514.

[36] D. Mesotten, M. Gielen, C. Sterken, K. Claessens, G. Hermans, D. Vlasselaers, J. Lemiere, L. Lagae, M. Gewillig, B. Eyskens, et al., Neurocognitive development of children 4 years after critical illness and treatment with tight glucose control: a randomized controlled trial, JAMA 308 (16) (2012) 1641–1650.