Predictability and Stability Testing to Assess Clinical Decision Instrument Performance for Children After Blunt Torso Trauma

Aaron E. Kornblith, MD [*, 1, 2]

Chandan Singh, BS [*, 3]

Gabriel Devlin, MD/CM [4]

Newton Addo, BS [1]

Christian J. Streck, MD [5]

James F. Holmes, MD, MPH [6]

Nathan Kuppermann, MD, MPH [6, 7]

Jacqueline Grupp-Phelan, MD, MPH [1, 2]

Jeffrey Fineman, MD [2]

Atul J. Butte, MD, Ph.D [8]

Bin Yu, Ph.D [3, 9, †]

[1] Department of Emergency Medicine, University of California, San Francisco

[2] Department of Pediatrics, University of California, San Francisco

[3] Department of Electrical Engineering & Computer Science, University of California, Berkeley

[4] Department of Pediatrics, University of California, San Francisco

[5] Department of Surgery, Medical University of South Carolina, Children's Hospital

[6] Department of Emergency Medicine, University of California, Davis

[7] Department of Pediatrics, University of California, Davis

[8] Bakar Computational Health Sciences Institute, University of California, San Francisco

[9] Departments of Statistics, University of California, Berkeley

[*] *Equal contribution*
[†] *Corresponding author*: binyu@berkeley.edu

Keywords: abdominal injuries/diagnostic imaging, abdominal injuries/surgery, child, pediatric trauma, clinical decision rules, interpretability

Author Contributions Statement: AK, CS, and BY conceived the study and designed the trial. AK, CS, CS, JH, NK, and BY supervised the conduct of the trial and data collection. AK, CS, NA, and BY provided statistical advice on study design and analyzed the data; all authors interpreted the data. AK, CS, and BY drafted the manuscript, and all authors contributed substantially to its revision. AK, CS, and BY take responsibility for the paper as a whole.

ABSTRACT

Objective The Pediatric Emergency Care Applied Research Network (PECARN) has developed a clinical-decision instrument (CDI) to identify children at very low risk of intra-abdominal injury. However, the CDI has not been externally validated. We sought to vet the PECARN CDI with the Predictability Computability Stability (PCS) data science framework, potentially increasing its chance of a successful external validation.

Materials & Methods We performed a secondary analysis of two prospectively collected datasets: PECARN (12,044 children from 20 emergency departments) and an independent external validation dataset from the Pediatric Surgical Research Collaborative (PedSRC; 2,188 children from 14 emergency departments). We used PCS to reanalyze the original PECARN CDI along with new interpretable PCS CDIs we developed using the PECARN dataset. External validation was then measured on the PedSRC dataset.

Results Three predictor variables (abdominal wall trauma, Glasgow Coma Scale Score <14, and abdominal tenderness) were found to be stable. Using only these variables, we developed a PCS CDI which had a lower sensitivity than the original PECARN CDI on internal PECARN validation but performed the same on external PedSRC validation (sensitivity 96.8% and specificity 44%).

Conclusion The PCS data science framework vetted the PECARN CDI and its constituent predictor variables prior to external validation. In this case, the PECARN CDI with 7 predictors, and our PCS-based CDI with 3 stable predictors, had identical performance on independent external validation. This suggests that both CDIs will generalize well to new populations, offering a potential strategy to increase the chance of a successful (costly) prospective validation.

1 INTRODUCTION

Background Blunt intra-abdominal injury is a leading cause of preventable death and disability in children in the U.S.[1] Computed tomography scans (CT) are the reference standard to diagnose intra-abdominal injury. In the last 30 years, CT use in children has increased without proportional improvements in clinical outcomes.[2] Indiscriminate use of CT is associated with an increased risk of radiation-induced malignancy.[3] Uncertainty and the lack of evidence in emergency department risk-stratification strategies lead to wide variation in CT use.[4] Furthermore, variability in practice increases cost and reduces effectiveness, efficiency, and quality of pediatric trauma care.[5] The Pediatric Emergency Care Applied Research Network (PECARN) prospectively developed a clinical decision instrument (CDI) to identify children after blunt torso trauma at very low risk for intra-abdominal injury undergoing acute intervention to decrease indiscriminate CT use.[6]

Importance Emergency care requires rapid and accurate decisions across a diverse group of patients and practices. CDIs reduce variability for high-prevalence conditions by offering the potential for more accurate and reliable diagnostic strategies than clinician judgment alone.[7] However, before widespread use, CDIs require external validation. External

validation is considered a more robust test of diagnostic performance than internal validation, and is critical to understanding the reliability of CDIs as they are generalized to new populations.[8,9] If the CDI performs poorly during external validation, it may be refined, reconsidered, or even abandoned.[10] However, prospective external validation may be expensive and cumbersome. Therefore, introducing a step to assess a CDI before external validation can ensure that it is developed and modeled to be as predictive and stable as possible, to increase the chance of successful external validation.

Recent progress in data science has led to innovative frameworks to assess the prediction performance and stability of healthcare-related diagnostic models, such as CDIs. The Predictability-Computability-Stability (PCS) framework is a unified approach to data science that protects against instability induced by subjective decisions made during the data science lifecycle.[11,12] PCS has improved drug-response prediction[12], gene-interaction search[13], and drug subgroup discovery in clinical trials[14]; these case-studies suggest that PCS may improve the CDI development and validation process before further investment into external validation. In addition to predictability as a reality check, two critical aspects of PCS are interpretability and stability analysis. To undergo PCS vetting, a CDI must be developed using interpretable methods, ensuring reproducibility.[15] Stability measures how much a CDI varies as choices made during the data science life cycle (including data cleaning and modeling), such as reasonable data alterations or different modeling techniques.[16] Here, multiple CDIs are developed by subsampling the original PECARN dataset. First, they are screened based on their test characteristics (predictability) and interpretability before assessing the variability of the importances of different predictor variables across high-performing CDIs (variable-level stability).

Goals of This Investigation The main objective of this study was to demonstrate the use of the PCS data science framework in vetting clinical decision instrument development (methods in Section 2.1 and results in Section 3.1). The secondary objective was to assess and externally validate the original PECARN clinical decision instrument for identifying children at very low risk of intra-abdominal injuries undergoing acute intervention after blunt torso trauma (methods in Section 2.2 and results in Section 3.2). Section 4 provides a discussion before Section 5 provides a conclusion.

2. METHODS

We analyzed two independent prospectively collected datasets from two large pediatric research networks, PECARN and the Pediatric Surgical Research Collaborative (PedSRC). This secondary analysis of anonymized data was deemed exempt from review by the University of California, San Francisco, and Medical University of South Carolina institutional review boards. There were two objectives of this study. The first (Section 2.1) was to demonstrate the PCS framework for improving CDI development. The second (Section 2.2) was to assess prediction performance and stability of the original PECARN CDI on external validation.

2.1 Objective 1: Demonstrate Predictability-Computability-Stability (PCS) Data Science Framework for Improving CDI Development

We followed the PCS framework, which goes beyond traditional reporting guidelines to assess the impact of reasonable human judgment calls by conducting reasonable data/model perturbations across the entire data science lifecycle.[7,16] PCS offers a framework to assess a CDI for diagnostic performance based on predictive performance (i.e. sensitivity and specificity) and computational needs, putting weight on stability. During the development of a CDI, investigators make

many "judgment calls", i.e. subjective decisions which may lead to variability in the final developed CDI. PCS recommends that investigators ensure that study conclusions are stable to any such judgment calls. These judgment calls can be checked by measuring the stability of conclusions when alternative "reasonable" judgment calls are made. Reasonable judgment calls are those solicited through direct engagement between clinicians and data scientists (see the Discussion section for a more detailed look at PCS in the context of CDIs).

In this study, the PCS framework was applied to CDI development (Figure A1), including all CDI development and validation stages (it could also be applied to the data cleaning stage, but was not done here). First, the PCS framework (1) defines the clinical problem, then reviews all aspects of (2) collecting and preprocessing data, and (3) develops CDIs using interpretable and rule-based models. Next, these CDIs are vetted for their (4) predictive performance (predictability) and the importance of predictor variables. Last, PCS (5) supports the interpretation of results by identifying variability in all the PCS steps (stability), ensuring CDIs are developed to be supported by both data and domain knowledge (provider input). In addition, PCS guided all aspects of data documentation and analysis; code is available on Github (https://github.com/csinva/iai-clinical-decision-rule).[10]

Development and Validation Dataset The PECARN dataset is a prospective cohort of 12,044 children after blunt torso trauma between May 2007 and January 2010 in 20 emergency departments.[6] Predictor variables were collected prospectively using a standard data collection tool. We used the PECARN definition for the a priori outcome of interest of intra-abdominal injury undergoing acute intervention.[6]

Following the original PECARN methods, we excluded any variable that was missing more than 5%, and used predictor variables with at least moderate inter-rater agreement, with the lower bound of the 95% confidence interval (CI) of the k measurements being at least 0.4.[17] Missing values for a predictor variable were imputed via its median, and we manually combined predictors that conveyed redundant information based on their correlations (Figure A2).

Original PECARN CDI Development. Redevelopment of the PECARN CDI ensures the replicability of the original trial. We followed the original PECARN development and internal cross-validation process to redevelop the PECARN CDI to identify children at very low risk for intra-abdominal injuries undergoing acute intervention.[6] We used a Classification and Regression Trees (CART) rule list,[18] which involves binary recursive partitioning using the Gini criterion.[19]

PCS CDI Development. We developed several alternative CDIs (corresponding to different judgment calls during modeling) to compare the predictive performance and perform stability analysis of the PECARN CDI. The following models were used to develop CDIs: logistic regression, CART decision trees, rule lists,[18] Bayesian Rule Lists,[20] iterative Random Forests,[13] RuleFit[21], Optimal sparse decision trees[22], Fast interpretable greedy-tree sums[23] and manual subgroup analysis. Each rule-based predictive model was chosen for its interpretability, taking the form of either a parsimonious list, tree, or set of binary rules. We used a stratified splitting technique to divide the PECARN dataset into a development set (i.e. a training set), 7,985 children (66%), and a validation set, 4,059 children (34%). Predictive models were fit using the imodels python package[24] (version 0.2.5). Hyperparameters were selected via manual tuning using only the development dataset.

CDI Predictive Performance. We calculated standard diagnostic statistics to report CDI performance. We used sensitivity and specificity curves to compare the diagnostic test characteristics of each CDI in the PECARN development and internal validation datasets. Furthermore, many more test characteristics were reported for each CDI, including their positive predictive value and Brier score (which helps evaluate the calibration of a CDI).[25] The CDIs were ranked heuristically from the sensitivity-specificity curves by weighting (threshold-dependent) sensitivity five times more than specificity. CDIs with poor predictive performance (i.e., achieving a sensitivity below 90%) were eliminated before further analysis.

CDI Stability. We assessed CDI stability by performing side-by-side comparisons of the PECARN CDI and alternative CDIs. To assess predictor-variable stability, we report the frequency and non-zero permutation-importance score of each predictor variable for each CDI.[26] The permutation importance measures the effect a predictor variable has on the overall prediction model's error. If a predictor variable is important, permuting or shuffling the value increases the model's error. The predictor variables with high permutation importance, especially across many different CDIs have greater stability.

We also compared the variability of diagnostic test characteristics between the PECARN development and internal validation datasets to assess the generalization of the model (i.e. stability of the predictive performance). Large changes in test characteristics suggest that the model is unstable in generalizing to new data. Moreover, a CDI can be unstable even when being re-developed to the same data. This is because many models contain some randomness in fitting, which can produce a different result when a model is re-developed. Therefore, we also measure randomness when each model is re-developed as a marker of stability. Prediction models were then ranked based on predictive performance (sensitivity and specificity), and then on variable-level stability.

2.2 Objective 2: Predictability and Stability of the Original PECARN CDI on External Validation

External Validation Dataset The PedSRC dataset is based on a prospective cohort of 2,188 children with blunt trauma at 14 non-PECARN Level I pediatric trauma centers.[27] Predictor variables were collected prospectively using a standard data collection tool. The PedSRC study defined intra-abdominal injury as any injury to an intra-abdominal structure identified on abdominal CT or at laparotomy. We matched the a priori PedSRC outcome of intra-abdominal injury undergoing intervention to the PECARN outcome.

We matched predictor and outcome variables between the datasets through distribution assessment and expert review. To ensure consistent matching, all variable linkages between datasets were reviewed by domain experts, including PECARN and PedSRC study principal investigators, to ensure biologic plausibility and ensure original data definition was congruent between the respective datasets. Variables with subjectivity were further screened, original documentation reviewed, and expert authorship team consensus was used to match variables. The same missing data strategy was used on the PedSRC and PECARN datasets.

PCS External Validation. To externally validate each CDI, we calculated threshold-bound and threshold-free standard diagnostic statistics. We calculated sensitivity, specificity, negative and positive predictive values, positive and negative likelihood ratios. We also included false positives, false negatives, accuracy, and F1 score. The F1 score, an accuracy indicator, emphasizes the clinical relevance of sensitivity over specificity and ranges from 1 (best value) to 0 (worst value).

We used sensitivity-specificity curves to compare the test characteristics of each candidate CDI on the external test dataset. We ranked predictor variable importance by assessing each variable's redundancy and weighted predictive power on the external validation dataset. Finally, we assessed overall CDI performance by evaluating the diagnostics test characteristics and variable importance.

We considered clinical context, predictive performance, computational speed, and stability to assign each CDI a rank. To compare the PCS framework to external validation, we first ranked predictive performance and stability. As the goal of the CDI is to limit unnecessary CT use in children after blunt torso trauma, we set a comparison threshold for predictive performance as a sensitivity five times more than specificity with a lower bound sensitivity of at least 95%. We calculated standard diagnostic statistics to report CDI performance, including sensitivity, specificity, negative and positive predictive values, positive, negative likelihood ratios, false positives, false negatives, accuracy, and F1 score. We also ranked predictor variable importance by assessing each variable's redundancy and weighted predictive power on the external validation dataset. Similarly, we measured overall stability as the proportion of the CDI's predictive performance assigned to predictor variables with the highest and lowest variable-level stability.

3 RESULTS

3.1 Results for Objective 1: Demonstrating the PCS Framework in CDI Development

Characteristics of Study Patients. The PECARN dataset included 12,044 children (Table 1). In PECARN, the mean (SD) age was 10.3 (5.4) years (1,167 patients <2 years), ranging from 0 to 18 years. The PedSRC external validation dataset included 2,188 children. The mean (SD) age was 7.8 (4.6) years (216 patients <2 years), ranging from 0 to 15 years. The PedSRC had a higher prevalence of motor vehicle collisions, compared to the PECARN development and validation datasets, 46.3% vs. 31.8% and 31.4%, and children with intra-abdominal injuries undergoing acute intervention, 2.8% vs. 1.7% and 1.7%, respectively (Table 1).

Table 1. Patient demographics and outcomes of the PECARN dataset split into development and validation (80:20), and the PedSRC external validation dataset.

	PECARN			**PedSRC**
	Total (N=12,044)	**Development** (n=7,985)	**Internal Validation** (n=4,059)	**External Validation** (N=2,188)
Age <2 years (%)	1167 (9.7%)	761 (9.5%)	406 (10%)	216 (9.9%)
Sex Male (%)	7384 (61.3%)	4887 (61.2%)	2497 (61.5%)	N/A
MVC (%)	3832 (31.8%)	2505 (31.4%)	1327 (32.7%)	1014 (46.3%)
CT scan (%)	5,179 (43.0%)	3,393 (42.5%)	1,786 (44.0%)	967 (44.2%)
IAI (%)	761 (6.3%)	485 (6.1%)	276 (6.8%)	261 (11.9%)
IAI-I (%)	203 (1.7%)	133 (1.7%)	70 (1.7%)	62 (2.8%)

PECARN: Pediatric Emergency Care Applied Research Network; PedSRC: Pediatric Surgery Research Collaborative; MVC: motor vehicle collision; CT scan: computed tomography; IAI: intra-abdominal injury; IAI-I: intra-abdominal injury undergoing acute intervention

Clinical Decision Instrument Development. We replicated the original PECARN CDI development using the PECARN dataset and redeveloped the identical seven ordered decision predictor variables in the PECARN rule list. The potential alternative CDIs, including Bayesian rule lists, CART Decision Trees, CART Rule Lists, Iterative Random Forests, and Rulefit are in Figures A3-7. The randomness for all re-developed models had no effect on any of the final CDIs performances.

Clinical Decision Instrument Internal Validation. Each CDI had a decline in performance between the development and internal validation PECARN datasets (Figure 1); however, the magnitude of the performance drop differed between different CDIs. The greater the magnitude in reduction suggests a less stable model. For example, the Iterative Random Forest CDI (red) and CART decision tree (orange) had the largest decline in performance between the development and validation datasets, suggesting that the prediction model was overfitting to the development dataset. In contrast, a fitted Bayesian rule list (blue), CART rule list (green), and Rule fit (purple) all retained similar predictive accuracy between development and validation. Table 2 summarizes the results of threshold-specific weights in which the sensitivity is weighted five times more heavily as specificity.

Figure 1. Sensitivity-specificity curves for clinical decision instruments to evaluate children after blunt torso trauma on the PECARN (a) development dataset (b) internal validation dataset, and (c) external validation on the PedSRC. The clinical decision instruments were then ranked by predictability from best to worst (top to bottom).

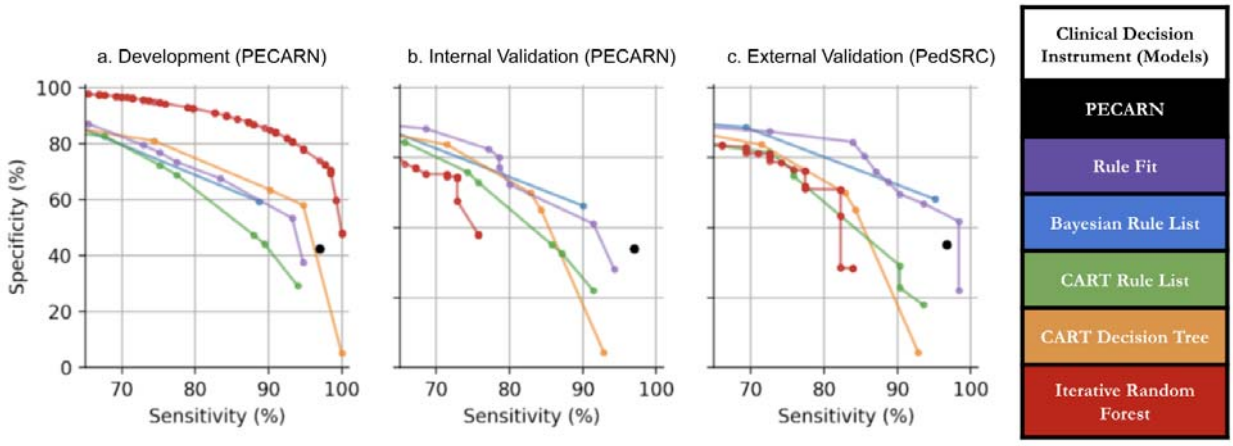


Table 2. Predictive performance of the clinical decision instruments with sensitivity weighted five times more heavily as specificity. (a) PECARN Development dataset, (b) PECARN Internal Validation Dataset, (c) PedSRC External Validation Dataset.

(a) PECARN Development Dataset	PECARN	Bayesian Rule List	CART Decision Tree	CART Rule List	Iterative Random Forest	Rule Fit
Sensitivity	98%	89%	95%	94%	98%	95%
Specificity	43%	59%	58%	29%	70%	47%
Negative predictive value	99.9%	100%	100%	100%	100%	100%
Positive predictive value	2.8%	4%	4%	2%	5%	3%
Negative likelihood ratio	0.035	0.19	0.09	0.21	0.02	0.1
Positive likelihood ratio	1.74	2.18	2.26	1.33	3.33	1.80
F1 score	0.056	0.07	0.07	0.04	0.10	0.06
Brier score	0.016	0.02	0.58	0.02	0.01	0.08

(b) PECARN Internal Validation Dataset	PECARN	Bayesian Rule List	CART Decision Tree	CART Rule List	Iterative Random Forest	Rule Fit
Sensitivity	94%	90%	84%	91%	71%	97%
Specificity	41%	58%	56%	28%	68%	33%
Negative predictive value	99.8%	100%	100%	99%	99%	100%
Positive predictive value	2.7%	4%	3%	2%	4%	2%
Negative likelihood ratio	0.14	0.17	0.28	0.31	0.42	0.09
Positive likelihood ratio	1.60	2.13	1.93	1.26	2.24	1.45
F1 score	0.053	0.07	0.06	0.04	0.07	0.04
Brier score	0.016	0.02	0.59	0.02	0.02	0.08

(c) PedSRC External Validation Dataset	PECARN	Bayesian Rule List	CART Decision Tree	CART Rule List	Iterative Random Forest	Rule Fit
Sensitivity	96.8%	95%	94%	90%	81%	97%
Specificity	44.0%	60%	60%	39%	63%	55%
Negative predictive value	99.8%	100%	100%	99%	99%	100%
Positive predictive value	4.8%	7%	6%	4%	6%	6%
Negative likelihood ratio	0.073	0.08	0.11	0.25	0.30	0.06
Positive likelihood ratio	1.73	2.39	2.33	1.47	2.21	2.13
F1 score	0.091	0.12	0.12	0.07	0.11	0.11
Brier score	0.026	0.03	0.56	0.03	0.03	0.09

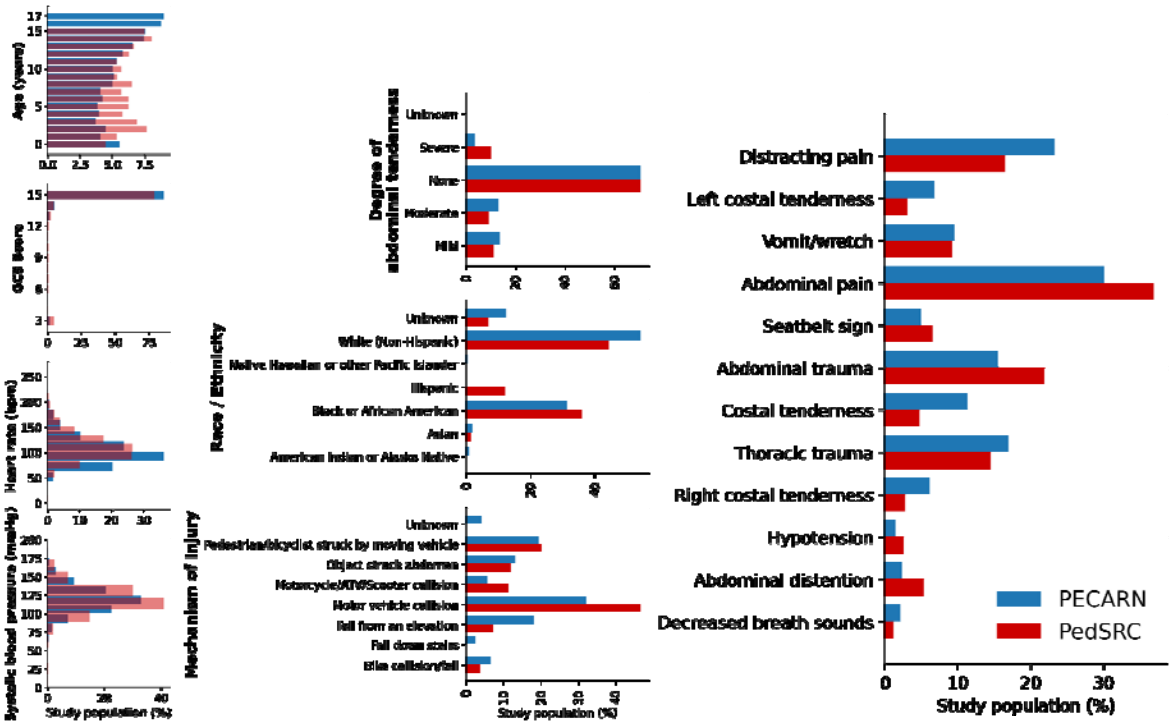
Predictability The original PECARN, Rule Fit, and Bayesian Rule List had minimal changes in performance between the development and internal validation datasets, suggesting relatively high predictability for these CDIs (Figure 1b). In contrast, CART Rule List, CART Decision Tree, and Iterative Random Forest had greater proportional declines in performance, suggesting lower predictability when heterogeneity in datasets was introduced.

Predictor-variable stability The most stable predictor variables were *abdominal trauma/seat belt sign, Glasgow Coma Scale Score < 14*, and *abdominal tenderness*. These three variables were the most frequent recurring predictor variables between CDIs. These three variables also had the highest non-zero permutation scores between the different CDIs (Figure A8). Therefore, it was recognized that the top three performing predictor variables were selected in the PECARN CDI and the four top-performing CDIs.

Computability Computability assesses the computational needs (e.g., hardware and demand for specialized equipment) of the project to understand the efficiency and feasibility of repeating the task. We evaluated the computational needs for CDI development and validation by timing each epoch and run time. In this case, all modeling and data analysis was performed on a standard laptop computer: CDI development took less than 10 minutes and validation less than 1 second.

3.2. Results for Objective 2: External Validation of Original PECARN Clinical Decision Instrument

Figure 2. Matched demographic and predictor variables from PECARN and PedSRC datasets visually represented for overall distributions.



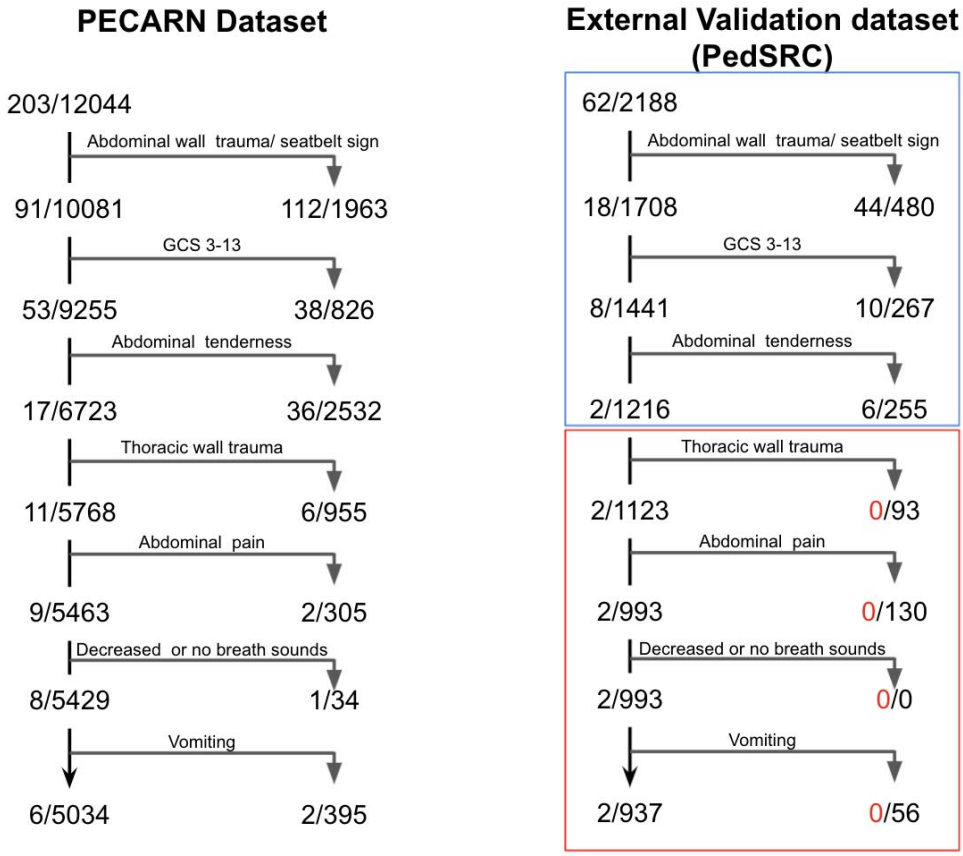
**GCS: Glasgow Coma Scale score; bpm: beats per minute; ATV: all-terrain vehicle; PECARN: Pediatric Emergency Care Applied Research Network; PedSRC: Pediatric Surgical Research Collaboration*

Distributions and Variable Matching for the external validation dataset. Predictor and outcome variables between the PECARN and PedSRC datasets were matched and evaluated for variable-level distributions (Figure 2). Most variables had direct matches between datasets (Table A1-3**)**. Predictor variables were assessed for redundancy and independent associations (Figure A2). The distribution of variables was well-matched except for the PECARN dataset inclusion of patients 15-17 years, and the lower frequency of children presenting after motor vehicle collisions (MVC) (Figure 2)**.**

External validation predictive performance. The original PECARN CDI successfully identified all but six children with intra-abdominal injuries undergoing acute interventions (sensitivity 97%, specificity 42.5%) on the PECARN dataset (Table 2b). On external validation using the PedSRC dataset, the original PECARN CDI maintained high prediction performance with an external validation sensitivity of 97.0% and specificity 44.0% (Table 2c). However, the original PECARN CDI missed two children with intra-abdominal injuries undergoing acute interventions; the clinical characteristics of these two children are presented in Table A4.

The top-performing three predictor variables of the PECARN CDI identified 60/62 patients with intra-abdominal injuries undergoing acute interventions in the PedSRC dataset, corresponding to 100% of the CDI's predictive power **(**blue box**,** Figure 3). The remaining four predictor variables (red box, Figure 3) did not add to the predictive performance of the PECARN CDI on the PedSRC dataset. The Brier score was 0.026, suggesting the predicted risk is well-calibrated when using the original PECARN CDI. The same three predictor variables also captured the majority of the predictive power in the original PECARN validation, identifying 186 of 201 outcomes. The additional four predictor variables substantially reduce the CDI's specificity. However, without these variables, the CDI misses 11 IAI-I patients in the PECARN dataset, resulting in an unacceptably low sensitivity.

Figure 3. Prediction tree for the original PECARN clinical decision instrument on (a) PECARN internal validation dataset, and (b) PedSRC external validation dataset. The blue box shows that the top three predictor variables retained all the predictive power for the clinical decision instrument on external validation. The red box shows the predictor variables without prediction power on external validation. From the top of the rule to the bottom, risks for the identified subgroups monotonically decrease, although risks are systematically higher on the PedSRC data.



3.3: Tying together Objective 1 and Objective 2: Comparing PCS Framework Predictions to External Validation of Clinical Decision Instruments

The ranked overall performance of the CDI on external validation matched that of the PCS framework prediction rankings (Figure 1c). This suggests that the results obtained from the PCS framework yielded useful information about the CDI's external validation performance, prior to collecting or analyzing the external validation dataset. In addition, the predictive performance was similar between internal validation and external validation (using the PedSRC dataset). However, most CDIs slightly improved their performances, suggesting that the CDIs are not overfitting to the PECARN dataset (Table 2c). The original PECARN CDI, Bayesian rule list, and Rule Fit had similar performances as in the PECARN datasets. In contrast, Iterative Random Forest, CART decision tree, and CART rule list had large declines in predictive performance (Figure 1c).

4. DISCUSSION

In the discussion, first we seek to describe PCS in the context of CDI development and vetting focusing on three key topics: predictability, stability, and interpretability. Next, we exemplify these three topics and their implications for the PECARN CDI.

4.1 Contextualizing PCS in the context of CDI development

Predictability The predictive performance of a CDI serves as the benchmark in the clinical literature. The concept of diagnostic test characteristics, such as sensitivity and specificity, are well-described and clinically used metrics for predictability. For example, previous literature has found that the PECARN CDI has a higher sensitivity than clinical judgment alone.[17] This study sought to evaluate the predictability of a CDI using threshold-dependent discriminative metrics (i.e., sensitivity) and threshold-free metrics (i.e. sensitivity-specificity curves). We found that the PECARN, Bayesian, and Rule Fit CDIs were the most predictable on external validation (PedSRC). However, CDIs used in clinical practice are designed to make predictions on varying populations, over time, and within differing conditions. Therefore, before using a CDI in clinical practice, investigators should validate how well a CDI will perform under varying conditions.

Stability Stability should be checked for all aspects of the data science lifecycle. Here, we largely focus on predictor-level stability, estimating how the feature importance of each predictor variable changes as a result of different judgment calls made during modeling. We also examine the stability of both the predictive performance and individual predictors to different calls made during data preprocessing. For example, we tried using GCS as a continuous predictor variable compared to different binary thresholds. The effect of this and many other judgment calls were found to be minimal and are omitted here (but can be found on our github).

Interpretability Interpretability enables the integration of domain expertise for the development and implementation of a CDI.[28–30] In contrast, black-box machine-learning models lack interpretability and may fail for unknown reasons when externally validated.[31] Post-hoc interpretations, such as permutation importance used here, can offer some interpretability[32–35], but are not a substitute for developing an interpretable model.[15,23,24,36] Therefore, we only consider parsimonious rule-based models. Each CDI is represented as a straightforward set or list of logical rules (IF:THEN statements), which can then be visualized. We restrict each model to a reasonable number of logical steps (fewer than 10), so each CDI can be assessed in real-time. We additionally fit logistic regression and optimal decision tree models, but found that they had poor; we find that fast interpretable greedy-tree sums learn precisely the same rules as CART so we omit this model here. PCS offers clear documentation guidelines to ensure the process is replicable, reproducible, and interpretable.[11]

As stated, black-box machine-learning models lack interpretability and may fail for unknown reasons when tested on new populations.[29] Examples of such complex models are neural networks, random forests, and support vector machines. However, even seemingly simple models such as logistic regression or decision trees can become uninterpretable if they are large enough and have too many steps.[15] Pennell (2020) utilized such models to re-evaluate the PECARN dataset.[37] The authors concluded that they had developed and validated a novel risk model using modern machine learning techniques. However, these complex machine-learning models lack the interpretability to integrate judgment, thus not allowing review nor the recognition of bias, which may build mistrust in the user.[30] Therefore, we use interpretable models with visual representation to allow stability analysis and ensure the integration of clinical judgment within the CDI.[28]

4.2 Implications for the PECARN CDI

As the second aim of this paper, we assessed the prediction performance and the stability of the original PECARN CDI for identifying children at very low risk of intra-abdominal injuries undergoing acute intervention after blunt torso trauma on external validation. Clinically, there is no standard, generalizable, validated strategy to identify children after blunt torso trauma in whom CT scans can safely be avoided. Instead, providers use ad hoc strategies that are inaccurate, and may fail to identify life-threatening injuries, leading to over-reliance on diagnostic imaging.[38–41] In 2013, PECARN sought to address the variability in accuracy and consistency by prospectively developing a CDI for children after blunt torso trauma.[6]

We used two uniquely matched prospectively collected but independent datasets to assess the CDI predictions and stability on external validation. Through this process, we reexamined the original PECARN findings using alternative reasonable statistical models and found the original PECARN CDI to be high performing. The PECARN CDI was highly predictive across the development, internal validation, and external validation datasets. Therefore, PECARN has strong predictive performance, which measures how well a CDI predicts in heterogeneous cohorts. We also found that three predictor variables made up the entirety of the predictive power on external validation: abdominal wall trauma, Glasgow Coma Scale Score <14, and abdominal tenderness. This is not surprising, as these three variables were also the most stable based on the PCS framework and made up the majority of the predictive power on the PECARN dataset (identifying 94.4% of the correctly predicted IAI-I patients).

Through the PCS framework, we found that the predictability, and stability of the original PECARN CDI warrants further investment and investigation, including prospective external validation. In contrast, if we found that the model or predictor variables were unstable in the original study, we would recommend against further validation. Our study can serve as an example for how investigators may evaluate the predictability and stability of a CDI for inherent weakness, prior to investing in a prospective external validation.

We found that if PCS could be successfully integrated as a novel step into prediction and diagnostic model development before external validation, there is a potential to streamline and evaluate CDIs to improve performance or expose weaknesses and avoid further investment in CDIs with poor stability. This is important because many CDIs have reduced accuracy during external validation.[42] Introducing a PCS step between CDI development and external validation, or using PCS directly for CDI development before external validation, will allow researchers, funders, and clinicians to understand better how CDIs may perform on future populations before external validation, impact analysis, or implementation into clinical practice. However, PCS is not able to replace external validation.

There are limitations to this study. First, we sought to develop high performing but interpretable CDIs. Therefore, we chose only rule-based models, including simple regression-based and complex machine learning models with interpretable visual outputs. The inclusion of less interpretable models may have improved diagnostic accuracy but interfered with conducting stability analysis, introducing domain expertise, and more easily recognizing bias. Second, the PECARN and PedSRC datasets were collected from different research groups. There is a potential for partial verification bias on external validation because the PedSRC dataset was not based on consecutive patient enrollment, and follow-up was limited to medical record review. Third, three predictor variables did not match between datasets. Two variables could not be matched because they were present in only one of the datasets: *gender* (PECARN only) and *femur fracture* (PedSRC only). The third

predictor variable was *distracting injury* (prospectively collected in PECARN but retrospectively aggregated in PedSRC). Given the limitations of this study, we believe prospective external validation is required before implementing the CDI.

5. CONCLUSION

In conclusion, the PCS data science framework helped vet CDI predictive performance and stability before external validation. The PCS framework offers a computational and less resource-intensive method than external validation. Even though it does not replace prospective external validation, PCS offers a method to vet for unstable CDIs to avoid further investment. We found that the predictive performance and stability of the PECARN CDI warranted further investigation. We used the external PSRC dataset to carry out this investigation, validating the PECARN CDI and a similar but simpler PCS-driven CDI.

CONFLICT OF INTERESTS

None.

FUNDING

This work was supported in part from NSF TRIPODS Grant 1740855, DMS-1613002, 1953191, 2015341, IIS 1741340, ONR grant N00014-17-1-2176. Moreover, this work is supported in part by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370. This project was supported in part by the National Center for Advancing Translational Sciences, National Institutes of Health, through UCSF-CTSI Grant Number UL1 TR001872.

REFERENCES

1. Kenefake ME, Swarm M, Walthall J. Nuances in pediatric trauma. *Emerg Med Clin North Am*. 2013;31(3):627-652. doi:10.1016/j.emc.2013.04.004

2. Meltzer JA, Stone ME Jr, Reddy SH, Silver EJ. Association of Whole-Body Computed Tomography With Mortality Risk in Children With Blunt Trauma. *JAMA Pediatr*. 2018;172(6):542-549. doi:10.1001/jamapediatrics.2018.0109

3. Miglioretti DL, Johnson E, Williams A, et al. The use of computed tomography in pediatrics and the associated radiation exposure and estimated cancer risk. *JAMA Pediatr*. 2013;167(8):700-707. doi:10.1001/jamapediatrics.2013.311

4. Marin JR, Wang L, Winger DG, Mannix RC. Variation in Computed Tomography Imaging for Pediatric Injury-Related Emergency Visits. *J Pediatr*. 2015;167(4):897-904 e3. doi:10.1016/j.jpeds.2015.06.052

5. Vogel AM, Zhang J, Mauldin PD, et al. Variability in the evalution of pediatric blunt abdominal trauma. *Pediatr Surg Int*. 2019;35(4):479-485. doi:10.1007/s00383-018-4417-z

6. Holmes JF, Lillis K, Monroe D, et al. Identifying children at very low risk of clinically important blunt abdominal injuries. *Ann Emerg Med*. 2013;62(2):107-116.e2. doi:10.1016/j.annemergmed.2012.11.009

7. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The tripod statement. *J Clin Epidemiol*. 2015;68(2):112-121. doi:10.1016/j.jclinepi.2014.11.010

8. Riley RD, Ensor J, Snell KIE, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ*. 2016;353:i3140. doi:10.1136/bmj.i3140

9. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med*. 1999;130(6):515-524. doi:10.7326/0003-4819-130-6-199903160-00016

10. Green SM, Schriger DL, Yealy DM. Methodologic standards for interpreting clinical decision rules in emergency medicine: 2014 update. *Ann Emerg Med*. 2014;64(3):286-291. doi:10.1016/j.annemergmed.2014.01.016

11. Yu B, Kumbier K. Veridical data science. *Proc Natl Acad Sci U S A*. 2020;117(8):3920-3929. doi:10.1073/pnas.1901326117

12. Li X, Tang TM, Wang X, Kocher JPA, Yu B. A stability-driven protocol for drug response interpretable prediction (staDRIP). *ArXiv201106593 Q-Bio Stat*. Published online November 16, 2020. Accessed May 11, 2021. http://arxiv.org/abs/2011.06593

13. Basu S, Kumbier K, Brown JB, Yu B. Iterative random forests to discover predictive and stable high-order interactions. *Proc Natl Acad Sci*. 2018;115(8):1943-1948. doi:10.1073/pnas.1711236115

14. Dwivedi R, Tan YS, Park B, et al. Stable Discovery of Interpretable Subgroups via Calibration in Causal Studies. *Int Stat Rev*. 2020;88(S1):S135-S178. doi:10.1111/insr.12427

15. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206-215. doi:10.1038/s42256-019-0048-x

16. Yu B. Three principles of data science: predictability, computability, and stability (PCS). In: *2018 IEEE International Conference on Big Data (Big Data)*. ; 2018:4-4. doi:10.1109/BigData.2018.8622080

17. Yen K, Kuppermann N, Lillis K, et al. Interobserver agreement in the clinical assessment of children with blunt abdominal trauma. *Acad Emerg Med*. 2013;20(5):426-432. doi:10.1111/acem.12132

18. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Wadsworth and Brooks; 1984.

19. Stiell IG, Wells GA. Methodologic standards for the development of clinical decision rules in emergency medicine. *Ann Emerg Med*. 1999;33(4):437-447. doi:10.1016/s0196-0644(99)70309-4

20. Letham B, Rudin C, McCormick TH, Madigan D. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction mode. *Ann Appl Stat*. 2015;9(3):1350-1371.

21. Friedman JH, Popescu BE. Predictive Learning Via Rule Ensembles. *Ann Appl Stat*. 2008;2(3):916-954.

22. Lin J, Zhong C, Hu D, Rudin C, Seltzer M. Generalized and Scalable Optimal Sparse Decision Trees. :11.

23. Tan YS, Singh C, Nasseri K, Agarwal A, Yu B. Fast Interpretable Greedy-Tree Sums (FIGS). *ArXiv220111931 Cs Stat*. Published online January 27, 2022. Accessed February 10, 2022. http://arxiv.org/abs/2201.11931

24. Singh C, Nasseri K, Tan Y, Tang T, Yu B. imodels: a python package for fitting interpretable models. *J Open Source Softw*. 2021;6(61):3192. doi:10.21105/joss.03192

25. Rufibach K. Use of Brier score to assess binary predictions. *J Clin Epidemiol*. 2010;63(8):938-939. doi:10.1016/j.jclinepi.2009.11.009

26. Breiman L. Random Forests. *Mach Learn*. 2001;45(1):5-32. doi:10.1023/A:1010933404324

27. Streck CJ, Vogel AM, Zhang J, et al. Identifying Children at Very Low Risk for Blunt Intra-Abdominal Injury in Whom CT of the Abdomen Can Be Avoided Safely. In: Vol 224. ; 2017:449-458.e3. doi:10.1016/j.jamcollsurg.2016.12.041

28. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Ann Intern Med*. 2006;144(3):201-209. doi:10.7326/0003-4819-144-3-200602070-00009

29. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med*. 2020;26(9):1320-1324. doi:10.1038/s41591-020-1041-y

30. Zorc JJ, Chamberlain JM, Bajaj L. Machine Learning at the Clinical Bedside-The Ghost in the Machine. *JAMA Pediatr*. 2019;162(1):W1-W73. doi:10.1001/jamapediatrics.2019.1075

31. Zihni E, Madai VI, Livne M, et al. Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome. *PLoS ONE*. 2020;15(4):1-15. doi:10.1371/journal.pone.0231166

32. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. :10.

33. Devlin S, Singh C, Murdoch WJ, Yu B. Disentangled Attribution Curves for Interpreting Random Forests and Boosted Trees. *ArXiv190507631 Cs Stat*. Published online May 18, 2019. Accessed February 10, 2022. http://arxiv.org/abs/1905.07631

34. Agarwal A, Tan YS, Ronen O, Singh C, Yu B. Hierarchical Shrinkage: improving the accuracy and interpretability of tree-based methods. *ArXiv220200858 Cs Stat*. Published online February 1, 2022. Accessed February 10, 2022. http://arxiv.org/abs/2202.00858

35. Singh C, Murdoch WJ, Yu B. Hierarchical Interpretations for Neural Network Predictions. *Int Conf Learn Represent*. Published online 2019:26.

36. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. *Proc Natl Acad Sci U S A*. 2019;116(44):22071-22080. doi:10.1073/pnas.1900654116

37. Pennell C, Polet C, Arthur LG, Grewal H, Aronoff S. Risk assessment for intraabdominal injury following blunt trauma in children. *J Trauma Acute Care Surg*. 2020;Publish Ah. doi:10.1097/ta.0000000000002717

38. Holmes JF, Sokolove PE, Brant WE, et al. Identification of children with intra-abdominal injuries after blunt trauma. *Ann Emerg Med*. 2002;39(5):500-509. doi:10.1067/mem.2002.122900

39. Capraro AJ, Mooney D, Waltzman ML. The use of routine laboratory studies as screening tools in pediatric abdominal trauma. *Pediatr Emerg Care*. 2006;22(7):480-484. doi:10.1097/01.pec.0000227381.61390.d7

40. Mahajan P, Kuppermann N, Tunik M, et al. Comparison of Clinician Suspicion Versus a Clinical Prediction Rule in Identifying Children at Risk for Intra-abdominal Injuries after Blunt Torso Trauma. *Acad Emerg Med*. 2015;22(9):1034-1041. doi:10.1111/acem.12739

41. Keller MS, Coln CE, Trimble JA, Green MC, Weber TR. The utility of routine trauma laboratories in pediatric trauma resuscitations. *Am J Surg*. 2004;188(6):671-678. doi:10.1016/j.amjsurg.2004.08.056

42. Toll DB, Janssen KJM, Vergouwe Y, Moons KGM. Validation, updating and impact of clinical prediction rules: A review. *J Clin Epidemiol*. 2008;61(11):1085-1094. doi:10.1016/j.jclinepi.2008.04.008

APPENDIX

Table A1. Predictor variables with 1:1 match between the two study datasets, PECARN and PedSRC.

	Predictor variable	PECARN variable	PedSRC variable
History	**Race**	*'Race'*	*'Race'*
	Vomiting*	*'VomitWretch'*	*'Emesis post injury'*
Vitals	**Heart rate**	*'InitHeartRate'*	*'Initial ED HR'*
	Blood pressure	*'InitSysBPRange'*	*'Initial ED systolic BP'*
	Hypotension	*'Hypotension'*	*'Hypotension'*
Exam	**Thoracic wall trauma***	*'ThoracicTrauma'*	*'Evidence of thoracic trauma (choice=None)'*
	Thoracic wall tenderness	*'ThoracicTender'*	*'Lower chest wall/costal margin tenderness to palpation (choice=None)'*
	Decreased breath sounds*	*'DecrBreathSound'*	*'Evidence of thoracic trauma (choice=Decreased breath sounds)'*
	Abdominal wall trauma*	*'AbdTrauma'*	*'Evidence of abdominal wall trauma'*
	Seat belt sign*	*'SeatBeltSign'*	*'Seatbelt sign'*
	Abdominal distention	*'AbdDistention'*	*'Abdominal distension'*
	Costal margin tenderness	*'CostalTender'*	*'Lower chest wall/costal margin tenderness to palpation'*

Predictor variables that match exactly. * Included in PECARN clinical decision instrument

Table A2. Predictor variables that were mapped between the two study datasets, PECARN and PedSRC.

	Feature	PECARN	Values (PECARN)		Values (PedSRC)	PedSRC	Notes
History	**Mechanism of injury**	*'RecodedMOI'*	Mechanism of injury (choice=Assault/struck) Mechanism of injury (choice=Other blunt mechanism)	↔	Object struck abdomen	*'Mechanism of injury'*	
			Mechanism of injury (choice=MVC)	↔	Motor vehicle collision		
			Mechanism of injury (choice=Motorcycle/dirt bike crash) Mechanism of injury (choice=ATV injury) Mechanism of injury (choice=Golf cart injury)	↔	Motorcycle/ATV/Scooter collision		
			Mechanism of injury (choice=Bike crash)	↔	Bike collision/fall		
			Mechanism of injury (choice=Bike struck by auto) Mechanism of injury (choice=Pedestrian struck by auto)	↔	Pedestrian/bicyclist struck by moving vehicle		
			Mechanism of injury (choice=Fall > 10 ft. height)	↔	Fall from an elevation		
			Unknown	↔	All other mechanisms of injury		
	Age	*'ageinyrs'*	0-2 (years) 2-17 (years)	↔ ↔	0-23 (months) 2-17 (years)	*'Age in years'* *'Age in months'*	Converted ages in months to years with decimal points
	Abdominal pain*	*'AbdomenPain'*	Yes No Other Unknown	↔ ↔ ↔ ↔	Yes No Other Unknown	*'Complain abdominal pain'*	
			Unable to assess	↔	Non-verbal Incoherent Intubated		
Vitals	**GCS***	*'GCSScore'*	3-15	↔	3-15	*'Initial GCS'*	Missing PECARN 'GCSScore' values replaced with 'AggregateGCS'
Exam	**Abdominal tenderness***	*'AbdTenderDegree'*	Mild Moderate	↔ ↔	Mild Moderate	*'Abdominal tenderness to palpation'*	
			Severe Unable to assess	↔ ↔	Severe Limited exam (intubation/sedation)		
			No	↔	None		

Table A3. Predictor variables that were adjudicated or left out between the two study datasets, PECARN and PedSRC.

	Feature	PECARN	Values (PECARN)		Values (PedSRC)	PedSRC	Notes
History	**Sex**	*'Sex'*	Sex	↔	?	Absent from data set	No adequate proxy found, so omitted from analysis
	Distracting injury	*'Distracting Pain'*	Yes	↔	Presence of any of following: skull fracture, facial fracture, clavicle fracture, rib fracture, pelvic fracture, femur fracture, dislocation, crush, or burn injury	Absent from dataset	We defined distracting pain in the PedSRC dataset as the presence of any of the injuries listed on the left
			No	↔	Absence of all of above		
			Unknown	↔	Absence of all of above, some missing values		
	Femur fracture	Absent from data set	?	↔	Yes No Unknown	*'Femur fracture'*	No adequate proxy found, so omitted from analysis

Table A4. Two children with intra-abdominal injury requiring acute intervention predicted very low risk by the original PECARN clinical decision instrument on the PedSRC external validation dataset.

(Age Range*), (Race/Ethnicity)	Mechanism	Additional Clinical Findings	Intra-abdominal Injury	Acute Intervention
Infant (0-5 years) (White, non-Hispanic)	Rollover MVC	Traumatic brain injury with skull fracture; femur fracture	Grade 1 liver laceration; Grade 3 splenic laceration	Exploratory laparotomy
Child (6-10 years) (Black, African American)	Fell > 10 feet	Extremity fracture	Grade 5 liver laceration	Angio-embolization

**Exact age concealed per policy; MVC: motor vehicle collision*

Figure A1. Five stages of the Predictability Computability Stability (PCS) framework as adapted for clinical decision instrument assessment. First, the PCS framework (1) defines the clinical problem, then reviews all aspects of (2) collecting and preprocessing data, and (3) models clinical decision instruments using interpretable and rule-based models. Next, the PCS framework (4) performs a validity (predictability), stability, and validation analysis. Last, the PCS framework (5) supports the interpretation of the results by identifying limitations in all the PCS steps, ensuring clinical decision instruments are developed to be supported by both data and domain expertise.

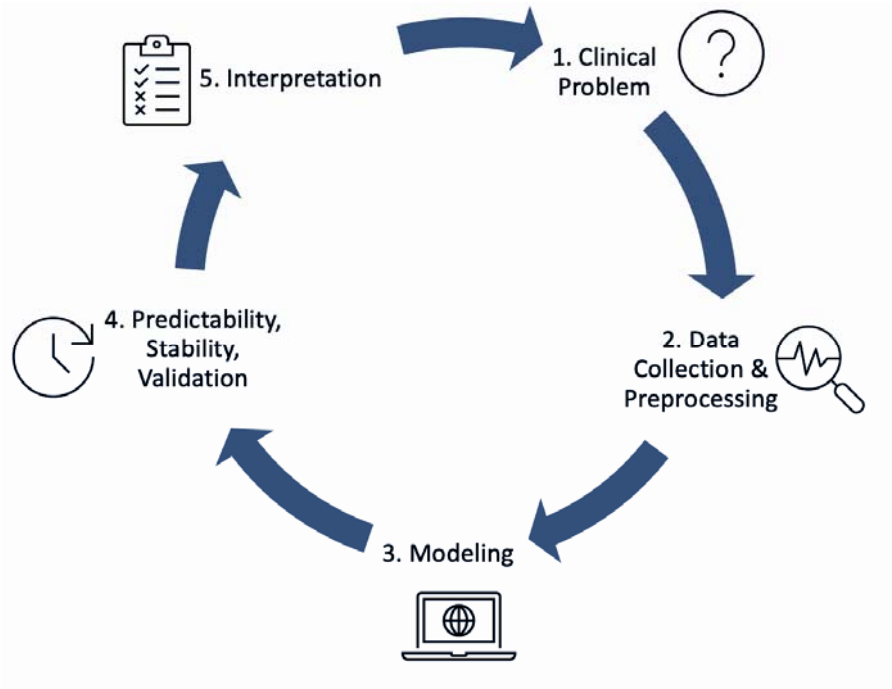


Figure A2. Redundant predictor variables are compared in this heatmap of (a) all predictors and (b) subset of key predictors. Darker blue signifies a direct correlation. Darker red signifies an inverse correlation. White signifies no correlation.

a.

b.

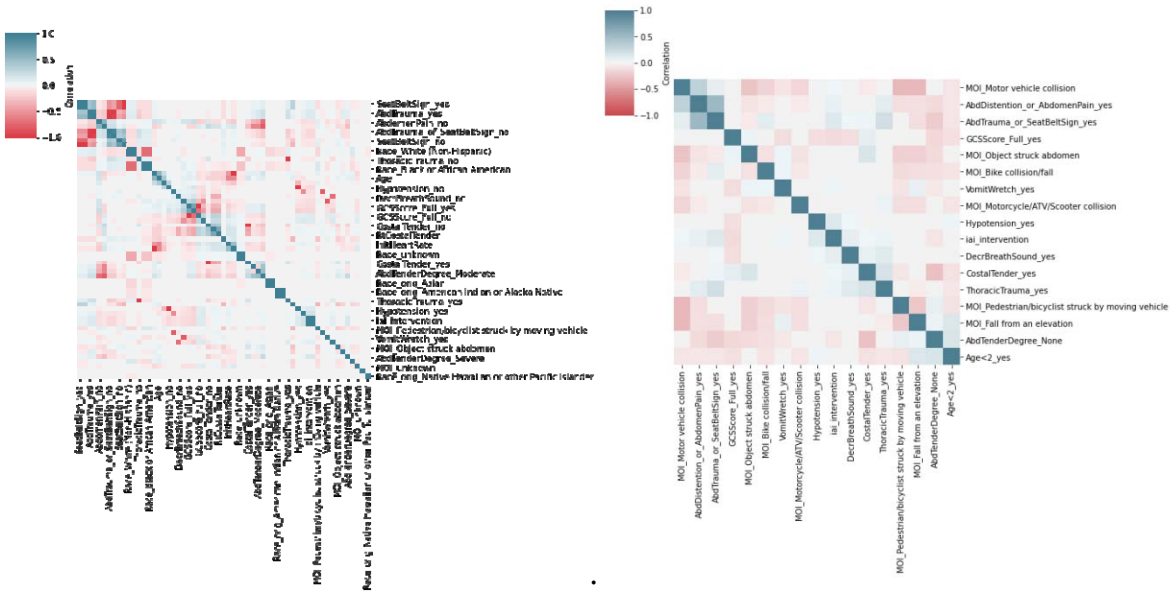


Figure A3. CART Decision tree.

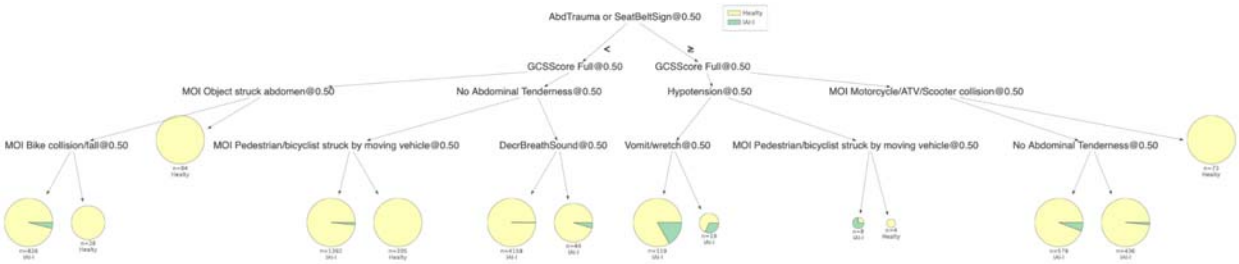
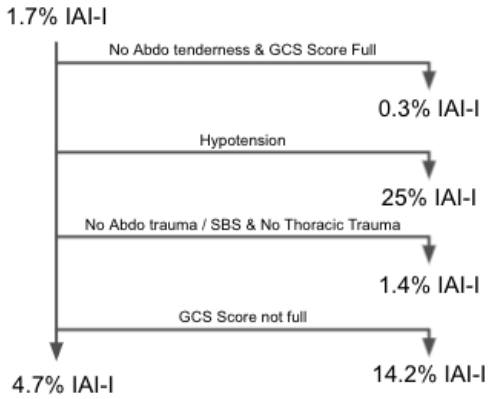


Figure A4. Bayesian rule list



IAI-I: intra-abdominal injury requiring acute intervention; Abdo:abdominal;
GCS: Glasgow Coma Scale score; SBS: seatbelt sign

Figure A5. RuleFit

Rule	Coefficient
No Abdominal Tenderness	-0.009
Abdominal Trauma or Seatbelt Sign	0.032
Glasgow Coma Scale score = 15	-0.034
Mechanism of Injury = Motor Vehicle Collision	0.001

Figure A6. CART Rule List

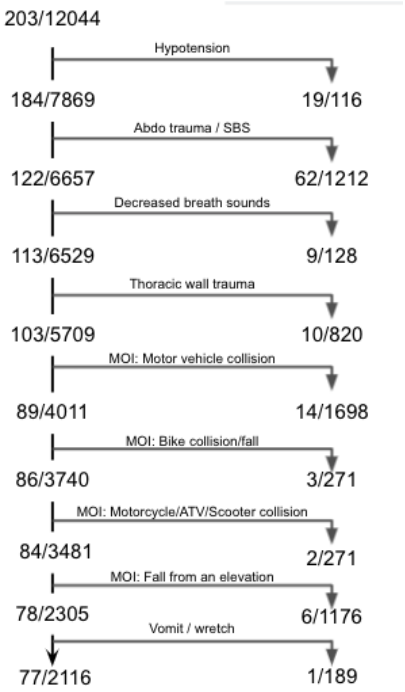
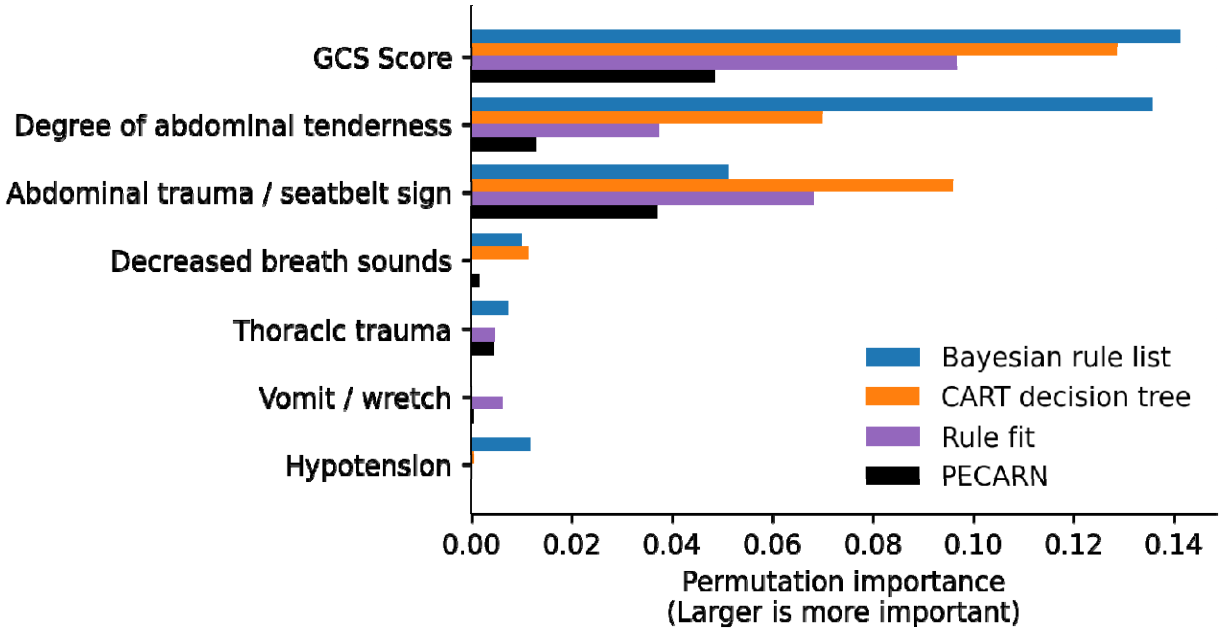


Figure A7. Iterative Random Forest permutation importance scores.

Rule	Permutation importance
Glasgow Coma Scale score = 15	0.094
Thoracic Trauma	0.09
Abdominal Trauma or Seatbelt Sign	0.09
Hypotension	0.089
Emesis/retching	0.083
Costal tenderness	0.073
No Abdominal Tenderness	0.073
Mechanism of Injury = Motor Vehicle Collision	0.071
Decreased breath sounds	0.064
Age < 2	0.062
Mechanism of Injury = Pedestrian/bicyclist struck by moving vehicle	0.047
Mechanism of Injury = Fall from an elevation	0.045
Abdominal Distention or Abdominal Pain	0.035
Mechanism of Injury = Object struck abdomen	0.034
Mechanism of Injury = Bike collision/fall	0.032
Mechanism of Injury = Motorcycle/ATV/Scooter collision	0.018

Figure A8. Non-zero permutation importance scores for predictor variables across high-performing clinical decision instruments.



**GCS Score: Glasgow Coma Scale score; PECARN: Pediatric Emergency Care Applied Research Network*