Group Probability-Weighted Tree Sums for Interpretable Modeling of Heterogeneous Data

Keyan Abou-Nasseri

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2022-167
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2022/EECS-2022-167.html

May 29, 2022

Group Probability-Weighted Tree Sums for Interpretable Modeling of Heterogeneous Data

by Keyan Nasseri

A Technical Report

Submitted to the Department of Electrical Engineering and Computer Sciences

University of California at Berkeley

in Partial Satisfaction of the Requirements

for the Degree of Master of Science, Plan II

Approval for the Report and Comprehensive Examination:

Committee:

Bin Yu

Bin Yu (May 29, 2022 17:34 PDT)

Professor Bin Yu
Advisor

May 29, 2022

Date

* * * * * * *

Aaron Kornblith (May 29, 2022 17:29 PDT)

Professor Aaron Kornblith
Second Reader

May 29, 2022

Date

Acknowledgments

This project is related to a broader effort on interpretable machine learning led by Professor Bin Yu as well as current and former members of the Yu Group including Chandan Singh, James Duncan, Yan Shuo Tan, and Abhineet Agarwal, all of whose work has been instrumental in providing the foundation for the ideas explored here. Further, Chandan Singh and James Duncan provided crucial guidance, mentorship, and assistance with this project in particular. Additionally, the experiments included were facilitated by the imodels Python package (`https://github.com/csinva/imodels`), which was co-authored with Singh, Tan, and Tiffany Tang.

I would also like to thank Dr. Aaron Kornblith for imbuing this work with his expertise in emergency medicine and pediatrics. More generally, this project builds on a 20-year history of research supervised by the Pediatric Emergency Care Applied Research Network (PECARN).

Lastly, I would like to acknowledge the work of Florica Constantine, Hyunsuk Kim, and Mark Oussoren in preprocessing the traumatic brain injury (TBI) dataset, and Chengzhong Ye and Albert Qu in doing so for the cervical spine injury (CSI) dataset, as a part of the final project for Statistics 215A in Fall 2021.

Abstract

Machine learning in high-stakes domains, such as healthcare, faces two critical challenges: (1) generalizing to diverse data distributions given limited training data while (2) maintaining interpretability. To address these challenges, we propose an instance-weighted tree-sum method that effectively pools data across diverse groups to output a concise, rule-based model. Given distinct groups of instances in a dataset (e.g., medical patients of different ages or from different treatment sites), our method first estimates group membership probabilities for each instance. Then, it uses these estimates as instance weights in FIGS [1], an existing greedy tree-sums method, to grow an set of decision trees whose values sum to the final prediction. We call this new method Group Probability-Weighted Tree Sums (G-FIGS). Extensive experiments on important clinical decision instruments datasets show that G-FIGS achieves state-of-the-art prediction performance; e.g., holding the level of sensitivity fixed at 92%, G-FIGS increases specificity for identifying cervical spine injury (CSI) by up to 10% over CART and up to 3% over FIGS alone, with larger gains at higher sensitivity levels. By keeping the total number of tree splits below 16 in FIGS, the final models remain interpretable, and we find that they match medical domain expertise. All code, data, and models are released on Github: Group Probability-Weighted Tree Sums is integrated into the Python package imodels [2] with an sklearn-compatible API, and experiments for reproducing the results here can be found at Yu-Group/imodels-experiments.

Table of Contents

1. Introduction

Recent advances in machine learning have led to impressive increases in predictive performance. However, machine learning has high stakes in the healthcare domain, with two critical challenges to effective adoption.

First, models must adapt to heterogenous data from diverse groups of patients [3]. Groups may differ dramatically and require distinct features for high predictive performance on the same outcome; e.g., infants may be nonverbal, disallowing features that require a verbal response, which in turn may be highly predictive in adults. A potential solution is to simply fit a unique model to each group (e.g., Kuppermann et al. 2009), but this discards valuable information that can be shared across groups.

Second, interpretability is essential for the development and implementation of models in healthcare and many other domains [5, 6]. Interpretability is required to ensure that models behave reasonably, identify when models will make errors, and make the models amenable to inspection by domain experts. Moreover, interpretable models tend to be much more computationally efficient than larger black-box models, often making them easier to use with humans in the loop, such as in medical diagnosis.

Here, we (1) address the challenge of sensibly sharing data across groups using group membership probability estimates and (2) address the challenges of interpretability by outputting a concise rule-based model. Specifically, we introduce Group Probability-Weighted Tree Sums (G-FIGS[1]), a two-step algorithm which takes in training data divided into known groups (e.g., patients in distinct age ranges), and outputs a rule-based model (Figure 1). G-FIGS first fits a classifier to predict group membership probabilities for each input instance (Figure 1A). Next, it uses these estimates as soft instance weights in the loss function of FIGS. The output is an ensemble of decision trees where the contribution from each tree is summed to yield a final prediction.

By sharing data sensibly across groups during training, G-FIGS results in a separate highly accurate rule-based model for each group. We test G-FIGS on three real-world clinical datasets (Chapter 4) and for two age groups commonly used in ER medicine; we find that G-FIGS outperforms state-of-the-art clinical decision instruments and competing machine learning methods in terms of specificity achieved at the high levels of sensitivity

[1] Our method is abbreviated as G-FIGS because we use an instance-weighted version of Fast Interpretable Greedy-tree sums (FIGS, Tan et al. 2022) to output a rule-based model.

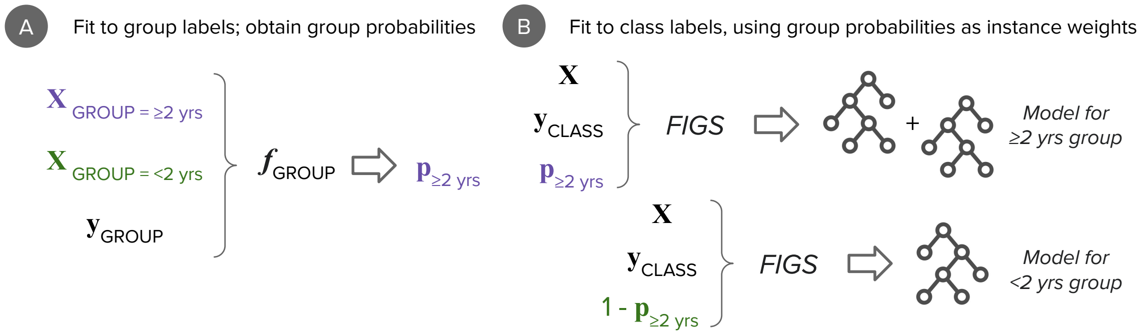

Figure 1. Overview of G-FIGS. **(A)** First, the covariates of each instance in a dataset are used to estimate an instance-specific probability of membership in each of the pre-specified groups in the data (e.g., patients of age <2 *yrs* and ≥ 2 *yrs*). **(B)** Next, these membership probabilities are used as instance weights when fitting an interpretable model for each group.

required in many clinical contexts. Moreover, G-FIGS maintains interpretability and ease-of-vetting with small (1-3 trees per group) and concise (≤ 6 splits per tree) clinical decision instruments by limiting the total number of rules across the trees for a given group.

1.1 Background and related work

We study the problem of sharing data across diverse groups in a supervised setting. Our methodology relies on estimates of group membership probabilities as instance weights in each group's outcome model, selected via cross-validation among multiple probability estimation methods. More weight is placed on instances that have higher estimated group-specific membership probability. In their role as group-balancing weights, we use these probabilities in a manner that is mathematically (though not conceptually) analogous to the use of propensity scores in causal inference for adjusting treatment-effect estimates [7]. More generally, this work is related to the literature on transfer learning [8], but we focus on transfer in the setting where outcomes are known for all training instances and interpretability is crucial.

Intrinsically interpretable methods, such as decision trees, have had success as highly predictive and interpretable models [9, 10]. Recent works have focused on improving the predictive performance of intrinsically interpretable methods [11, 12], particularly for rule-based models [13, 14, 1, 15], without degrading interpretability.

A key domain problem involving interpretable models is the development of clinical decision instruments, which can assist clinicians in improving the accuracy and efficiency of diagnostic strategies. Recent works have developed and validated clinical decision instruments using interpretable machine learning models, particularly in emergency medicine [16, 17, 18, 19].

2. Method: G-FIGS

Setup We assume a supervised learning setting (classification or regression) with features X (e.g., *blood pressure*, *signs of vomiting*), and an outcome Y (e.g., *cervical spine injury*). We are also given a group label G, which is specified using the context of the problem and domain knowledge; for example, G may correspond to different sites at which data is collected, different demographic groups which are known to require different predictive models, or data before/after a key temporal event. G should be discrete, as G-FIGS will produce a separate model for each unique value of G, but may be a discretized continuous or count feature.

Fitting group membership probabilities The first stage of G-FIGS fits a classifier to predict group membership probabilities $P(G|X)$ (Figure 1A). In estimating $P(G = g|X)$, we exclude features that trivially identify G (e.g., we exclude *age* when values of G are age ranges). Intuitively, these probabilities inform the degree to which a given instance is representative of a particular group; the larger the group membership probability, the more the instances should contribute to the model for that group. Any classifier can be used; we find that logistic regression and gradient-boosted decision trees perform best. The group membership probability classifier can be selected using cross-validation, either via group-label classification metrics or downstream performance of the weighted prediction model; we take the latter approach.

Fitting group probability-weighted FIGS In the second stage (Figure 1B), for each group $G = g$, G-FIGS uses the estimated group membership probabilities, $P(G = g|X)$, as instance weights in the loss function of a ML model for each group $P(Y|X, G = g)$. Intuitively, this allows the outcome model for each group to use information from out-of-group instances when their covariates are sufficiently similar. While the choice of outcome model is flexible, we find that the Fast Interpretable Greedy-Tree Sums (FIGS) model [1] performs best when both interpretability and high predictive performance are required. By greedily fitting a sum of trees, FIGS effectively allocates a small budget of rules to different types of structure in data. When interpretability is not critical, the same weighting procedure could also be applied to black-box models, such as Random Forest [20].

3. Datasets

3.1 Overview

Table 1 shows the main datasets under consideration here. Each publicly-available dataset constitutes a large-scale multi-site data aggregation generated by the Pediatric Emergency Care Applied Research Network (PECARN) with a relevant clinical outcome. For each of these datasets, we use their natural grouping of patients into <2 *yrs* and ≥ 2 *yrs* groups, where the young group includes only patients whose age is less than two years. This age-based threshold is commonly used for emergency-based diagnostic strategies (e.g., Kuppermann et al. [4]), because it follows a natural stage of development, including a child's ability to participate in their care. At the same time, the natural variability in early childhood development also creates opportunities to share information across this threshold. These datasets are non-standard for ML; as such, we spend considerable time cleaning and preprocessing these features along with medical expertise included in the authorship team.

We use 60% of the data for training, 20% for tuning hyperparameters (including estimation of $P(G|X)$), and 20% for evaluating test performance of the final models. More details on data splitting are provided in Chapter 4.

Unprocessed data is available at `https://pecarn.org/datasets/` and clean data is available on github at `https://github.com/csinva/imodels-data` (easily accessibly through the imodels [2] package). The final set of features used for fitting outcome models is shown in Table 2.

Name	Patients	Outcome	% Outcome	Features
TBI	42428	376	0.9	61
IAI	12044	203	1.7	21
CSI	3313	540	16.3	34

Table 1. Clinical decision-instrument datasets for traumatic brain injury (TBI) [4], intra-abdominal injury (IAI) [19], and cervical spine injury (CSI) [21].

3.2 Preprocessing

Traumatic brain injury (TBI) To screen patients, we follow the inclusion and exclusion criteria from Kuppermann et al. [4], which exclude patients with Glasgow Coma Scale (GCS) scores under 14 or no signs or symptoms of head trauma, among other disqualifying factors. No patients were dropped due to missing values: the majority of patients have about 1% of features missing, and are at maximum still under 20%. We utilize the same set of features as Kuppermann et al. [4].

Our strategy for imputing missing values differed between features according to clinical guidance. For features that are unlikely to be left unrecorded if present, such as paralysis or skull fracture, missing values were assumed to be negative. For other features that could be unnoticed by clinicians or guardians, such as loss of consciousness, missing values are assumed to be positive. For features that did not fit into either of these groups or were numeric, missing values are imputed with the median.

Cervical spine injury (CSI) Leonard et al. [21] engineered a set of 22 expert features from 609 raw features; we utilize this set but add back features that provide information on the following:

- Patient position after injury
- Clinical intervention received by patients prior to arrival (immobilization, intubation)
- Pain and tenderness of the head, face, torso/trunk, and extremities
- Age and gender
- Whether the patient arrived by emergency medical service (EMS)

We follow the same imputation strategy described in the preceding TBI subsection. Features that are assumed to be negative if missing include focal neurological findings, motor vehicle collision, and torticollis, while the only feature assumed to be positive if missing is loss of consciousness.

Intra-abdominal injury (IAI) We follow the data preprocessing steps described in Holmes et al. [22] and Kornblith et al. [18]. In particular, all features of which at least 5% of values are missing are removed, and variables that exhibit insufficient interrater agreement (lower bound of 95% CI under 0.4) are removed. The remaining missing values are imputed with the median. In addition to the 18 original variables, we engineered three additional features:

- *Full Glasgow Coma Scale (GCS) score*: True when GCS is equal to the maximum score of 15
- *Abd. Distention or abd. pain*: Either abdominal distention or abdominal pain
- *Abd. trauma or seatbelt sign*: Either abdominal trauma or seatbelt sign

Data for predicting group membership probabilities The data preprocessing steps for the group membership models in the first step of G-FIGS are identical to that above, except that missing values are not imputed at all for categorical features, such that “missing", or NaN, is allowed as one of the feature labels in the data. We find that this results in more accurate group membership probabilities, since for some features, such as those requiring a verbal response, missing values are predictive of age group.

Traumatic brain injury		
Feature Name	% Missing	% Nonzero
Altered Mental Status	0.74	12.95
Alt. Mental Status: Agitated	87.05	1.79
Alt. Mental Status: Other	87.05	1.82
Alt. Mental Status: Repetitive	87.05	1.04
Alt. Mental Status: Sleepy	87.05	6.67
Alt. Mental Status: Slow to respond	87.05	3.22
Acting normally per parents	7.09	85.38
Age (months)	0.00	N/A
Verbal amnesia	38.41	10.45
Trauma above clavicles	0.30	64.38
Trauma above clav.: Face	35.92	29.99
Trauma above clav.: Scalp-frontal	35.92	20.48
Trauma above clav.: Neck	35.92	1.38
Trauma above clav.: Scalp-occipital	35.92	9.62
Trauma above clav.: Scalp-parietal	35.92	7.79
Trauma above clav.: Scalp-temporal	35.92	3.39
Drugs suspected	4.19	0.87
Fontanelle bulging	0.37	0.06
Sex	0.01	N/A
Headache severity	2.38	N/A
Headache start time	3.09	N/A
Headache	32.76	29.94
Hematoma	0.69	39.42
Hematoma location	0.47	N/A
Hematoma size	1.67	N/A
Severity of injury mechanism	0.74	N/A
Injury mechanism	0.67	N/A
Intubated	0.73	0.01
Loss of consciousness	4.05	10.37
Length of loss of consciousness	5.39	N/A
Neurological deficit	0.85	1.3
Neurological deficit: Cranial	98.70	0.18
Neurological deficit: Motor	98.70	0.28
Neurological deficit: Other	98.70	0.71
Neurological deficit: Reflex	98.70	0.03
Neurological deficit: Sensory	98.70	0.26
Other substantial injury	0.43	10.07
Other sub. injury: Abdomen	89.93	1.25
Other sub. injury: Cervical spine	89.93	1.37
Other sub. injury: Cut	89.93	0.12
Other sub. injury: Extremity	89.93	5.49
Other sub. injury: Flank	89.93	1.56
Other sub. injury: Other	89.93	1.65
Other sub. injury: Pelvis	89.93	0.44
Paralyzed	0.75	0.01
Basilar skull fracture	0.99	0.68
Basilar skull frac.: Hemotympanum	99.32	0.35
Basilar skull frac.: CSF otorrhea	99.32	0.04
Basilar skull frac.: Periorbital	99.32	0.19
Basilar skull frac.: Retroauricular	99.32	0.08
Basilar skull frac.: CSF rhinorrhea	99.32	0.03
Skull fracture: Palpable	0.24	0.38
skull frac.: Palpable and depressed	99.69	0.18
Sedated	0.76	0.08
Seizure	1.70	1.17
Length of seizure	0.18	N/A
Time of seizure	0.12	N/A
Vomiting	0.71	13.1
Time of last vomit	89.04	N/A
Number of times vomited	0.60	N/A
Vomit start time	0.87	N/A

Intra-abdominal injury		
Abdominal distention	4.38	2.3
Abdominal distention or pain	0.00	4.93
Degree of abdominal tenderness	70.13	N/A
Abdominal trauma	0.56	15.48
Abd. trauma or seat belt sign	0.00	16.3
Abdomen pain	15.38	30.06
Age (years)	0.00	N/A
Costal margin tenderness	0.00	11.33
Decreased breath sound	1.93	2.13
Distracting pain	7.38	23.29
GCS (Glasgow Coma Scale)	0.00	N/A
Full GCS score	0.00	86.21
Hypotension	0.00	1.44
Left costal margin tenderness	0.00	N/A
Method of injury	3.95	N/A
Right costal margin tenderness	0.00	N/A
Seat belt sign	3.30	4.93
Sex	0.00	N/A
Thoracic tenderness	9.99	15.96
Thoracic trauma	0.63	16.95
Vomiting	3.92	9.57

Cervical spine injury		
Age (years)	0.00	N/A
Altered mental status	2.05	24.72
Axial load to head	0.00	24.0
Clotheslining	3.38	0.94
Focal neurological findings	9.84	14.67
Method of injury: Diving	0.03	1.3
Method of injury: Fall	2.44	3.83
Method of injury: Hanging	0.03	0.15
Method of injury: Hit by car	0.03	15.09
Method of injury: Auto collision	7.73	14.73
Method of injury: Other auto	0.03	3.11
Arrived by EMS	0.00	77.24
Loss of consciousness	8.03	42.68
Neck pain	5.25	38.42
Posterior midline tenderness	2.57	29.88
Patient position on arrival	61.52	N/A
Predisposed	0.00	0.66
Pain: Extremity	18.35	25.87
Pain: Face	18.35	7.58
Pain: Head	18.35	29.04
Pain: Torso/trunk	18.35	28.95
Tenderness: Extremity	20.37	15.15
Tenderness: Face	20.37	3.83
Tenderness: Head	20.37	7.79
Tenderness: Torso/trunk	20.37	25.87
Substantial injury: Extremity	1.03	10.87
Substantial injury: Face	1.06	5.67
Substantial injury: Head	1.00	15.88
Substantial injury: Torso/trunk	1.03	7.3
Neck tenderness	2.48	39.3
Torticollis	7.03	5.77
Ambulatory	5.77	21.46
Axial load to top of head	0.00	2.35
Sex	0.00	N/A

Table 2. Final features used for fitting the *outcome* models. Features include information about patient history (i.e. *mechanism of injury*), physical examination (i.e. *Abdominal trauma*), and mental condition (i.e. *Altered mental status*). Percentage of nonzero values is marked *N/A* for non-binary features.

4. Results

4.1 G-FIGS predicts well

Table 3 shows the prediction performance of G-FIGS and a subset of baseline methods. Sensitivity is extremely important for these settings, as a false negative (missing a diagnosis) has much more severe consequences than a false positive. For high levels of sensitivity, G-FIGS generally improves the model's specificity against the baselines. We compare to three baselines: CART [10], FIGS [1], and Tree-Alternating Optimization TAO [23]). For each baseline, we either (i) fit one model to all the training data or (ii) fit a separate model to each group (denoted with *-SEP*). Limits on the total number of rules for each model are varied over a range which yields interpretable models, from 8 to 16 maximum rules (full details of this and other hyperparameters are in Section 4.4).

	Traumatic brain injury				Cervical spine injury				Intra-abdominal injury			
Sensitivity:	92%	94%	96%	98%	92%	94%	96%	98%	92%	94%	96%	98%
TAO	6.2	6.2	0.4	0.4	41.5	21.2	0.2	0.2	0.2	0.2	0.0	0.0
TAO-SEP	26.7	13.9	10.4	2.4	32.5	7.0	5.4	2.5	12.1	8.5	2.0	0.0
CART	20.9	14.8	7.8.	2.1	38.6	13.7	1.5	1.1	11.8	2.7	1.6	1.4
CART-SEP	26.6	13.8	10.3	2.4	32.1	7.8	5.4	2.5	11.0	9.3	2.8	0.0
G-CART	15.5	13.5	6.4	3.0	38.5	15.2	4.9	3.9	11.7	10.1	3.8	0.7
FIGS	23.8	18.2	12.1	0.4	39.1	33.8	24.2	**16.7**	**32.1**	13.7	1.4	0.0
FIGS-SEP	39.9	19.7	**17.5**	2.6	38.7	33.1	20.1	3.9	18.8	9.2	2.6	0.9
G-FIGS	**42.0**	**23.0**	14.7	**6.4**	**42.2**	**36.2**	**28.4**	15.7	29.7	**18.8**	**11.7**	**3.0**

Table 3. Best test set specificity when sensitivity is constrained to be above a given threshold. G-FIGS provides the best performance overall in the high-sensitivity regime. *-SEP* models fit a separate model to each group, and generally outperform fitting a model to the entire dataset. G-CART follows the same approach as G-FIGS but uses weighted CART instead of FIGS for each final group model. Averaged over 10 random data splits into training, validation, and test sets, with hyperparameters chosen independently for each split. See Table 7 for more detail.

4.2 Interpreting the group membership model

In this clinical context, we begin by fitting several logistic regression and gradient-boosted decision tree group membership models to each of the training datasets to predict whether a patient is in the <2 *yrs* or ≥ 2 *yrs* group. Random forests and CART were tried as well, but were found to lead to worse performance (see Section 4.4 for more detail). For the instance-weighted methods, we treat the choice of group membership model as a

Traumatic brain injury		Cervical spine injury		Intra-abdominal injury	
Variable	Coef	Variable	Coef	Variable	Coef
No fontanelle bulging	3.62	Neck tenderness	2.44	Bike injury	2.01
Amnesia	2.07	Neck pain	2.18	Abdomen pain	1.66
Struck by vehicle	1.44	Motor vehicle injury: other	1.54	Thoracic tenderness	1.43
Headache	1.39	Hit by car	1.47	Hypotension	1.23
Bike injury	1.26	Substantial injury: extremity	1.35	No abdomen pain	0.98

Table 4. Logistic regression coefficients for features that contribute to high $P(\geq 2 \, yrs \mid X)$ reflect known medical knowledge. For example, features with large coefficients require verbal responses (e.g., *Amnesia*, *Headache*, *Pain*), relate to activities not typical for the <2 *yrs* group (*Bike injury*), or are specific to older children, e.g., older children should have *No fontanelle bulging*, as cranial soft spots typically close by 2 to 3 months after birth.

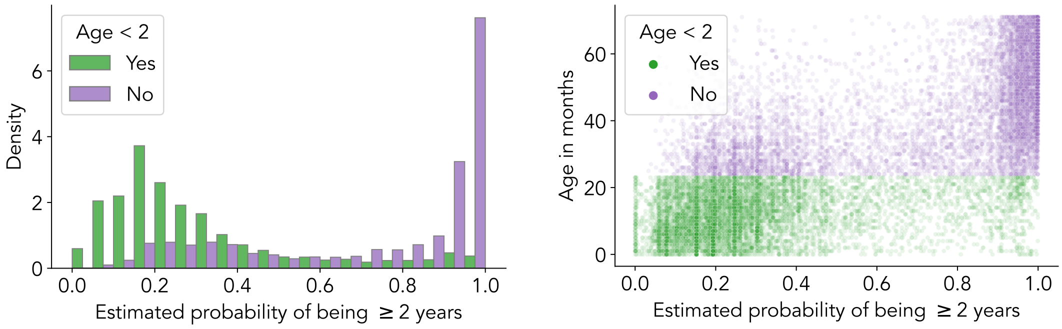

Figure 2. Visualizations of the group membership probabilities estimated in the first step of G-FIGS. Note that the bulk of the distribution for the <2 group is under 0.4, and that for the ≥ 2 group is over 0.8, indicating most points are estimated to be in the correct group. However, there is a noticeable cluster of ≥ 2 points which have scores between 0.1 and 0.4, and another small cluster of <2 points which have scores of about 0.95.

hyperparameter, and select the best model according to the downstream performance of the final decision rule on the validation set.

Table 4 shows the coefficients of the most important features for each logistic regression group membership model when predicting whether a patient is in the ≥ 2 *yrs* group. The coefficients reflect existing medical expertise. For example, the presence of verbal response features (e.g., *Amnesia*, *Headache*) increases the probability of being in the ≥ 2 *yrs* group, as does the presence of activities not typical for the <2 *yrs* group (e.g. *Bike injury*).

Figure 2 visualizes how information is shared between groups in the data. Most points are correctly classified within their group, but there is a small fraction of points which are found by the group membership model to be more similar to points in the other group. Additionally, there are some borderline points which the model finds equally likely to be in either group. These scores allow the final outcome model to navigate the age cutoff with more nuance than would be possible with two completely separate models.

4.3 Interpreting the outcome models

Figure 3, Figure 4, and Figure 5 shows the G-FIGS models fitted to the entire TBI, CSI, and IAI datasets respectively, selected via cross-validation. Outcome predictions for a group are made by summing the predicted risk contribution (Δ *Risk*) from the appropriate leaf of each tree in the group's fitted tree ensemble. Δ Risk is not simply equivalent to the fraction of patients with the condition since (i) G-FIGS uses patients from both groups, (ii) each tree in FIGS fits the residuals of the others, and (iii) positive examples are upweighted in the loss function of FIGS (see Section 4.4 for more detail).

The models are concise, highly predictive, and match existing medical knowledge. In general, we find that the tree ensemble of G-FIGS allows it to adapt a succinct model to independent risk factors in the data, whereas individual tree models (i.e., CART, TAO) are not flexible enough to model additive effects in the data.

4.3.1 TBI

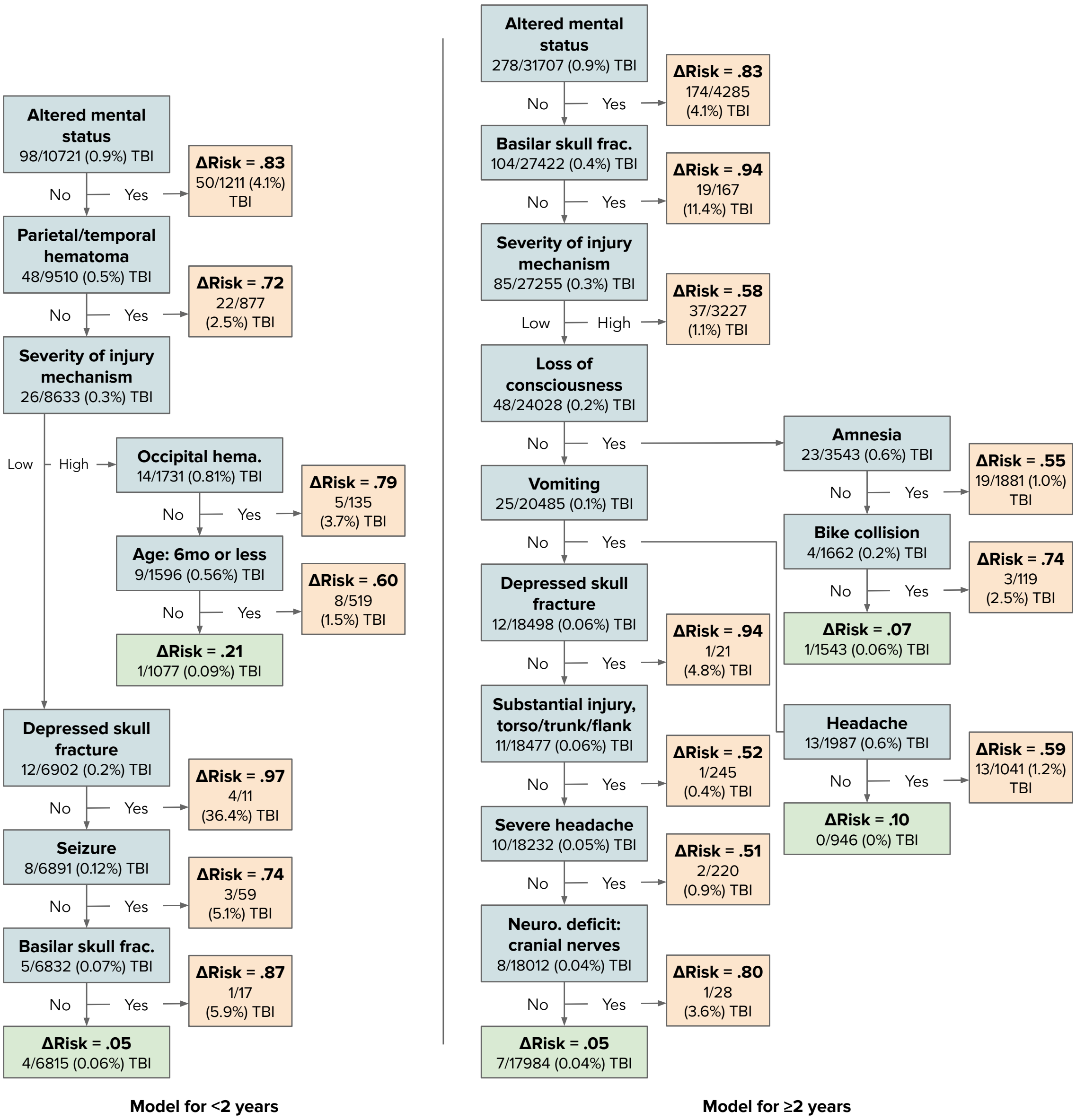

Figure 3. G-FIGS model fitted to the TBI dataset. Achieves 97.1% sensitivity and 58.9% specificity (training). The features used by each group are overlapping and reasonable, matching medical domain knowledge and partially matching previous work [4]; e.g., features such as *altered mental status*, *basilar skull fracture*, and *loss of consciousness* are all known to increase the risk of TBI. Features unique to each group largely relate to the age cutoff; the <2 *yrs* features only include those that clinicians can assess without asking the patient (e.g., *parietal hematoma*), while two of the ≥ 2 *yrs* features require verbal responses (*severe headache*, *headache*). Interestingly, in this case G-FIGS learns only a single tree for each group.

4.3.2 CSI

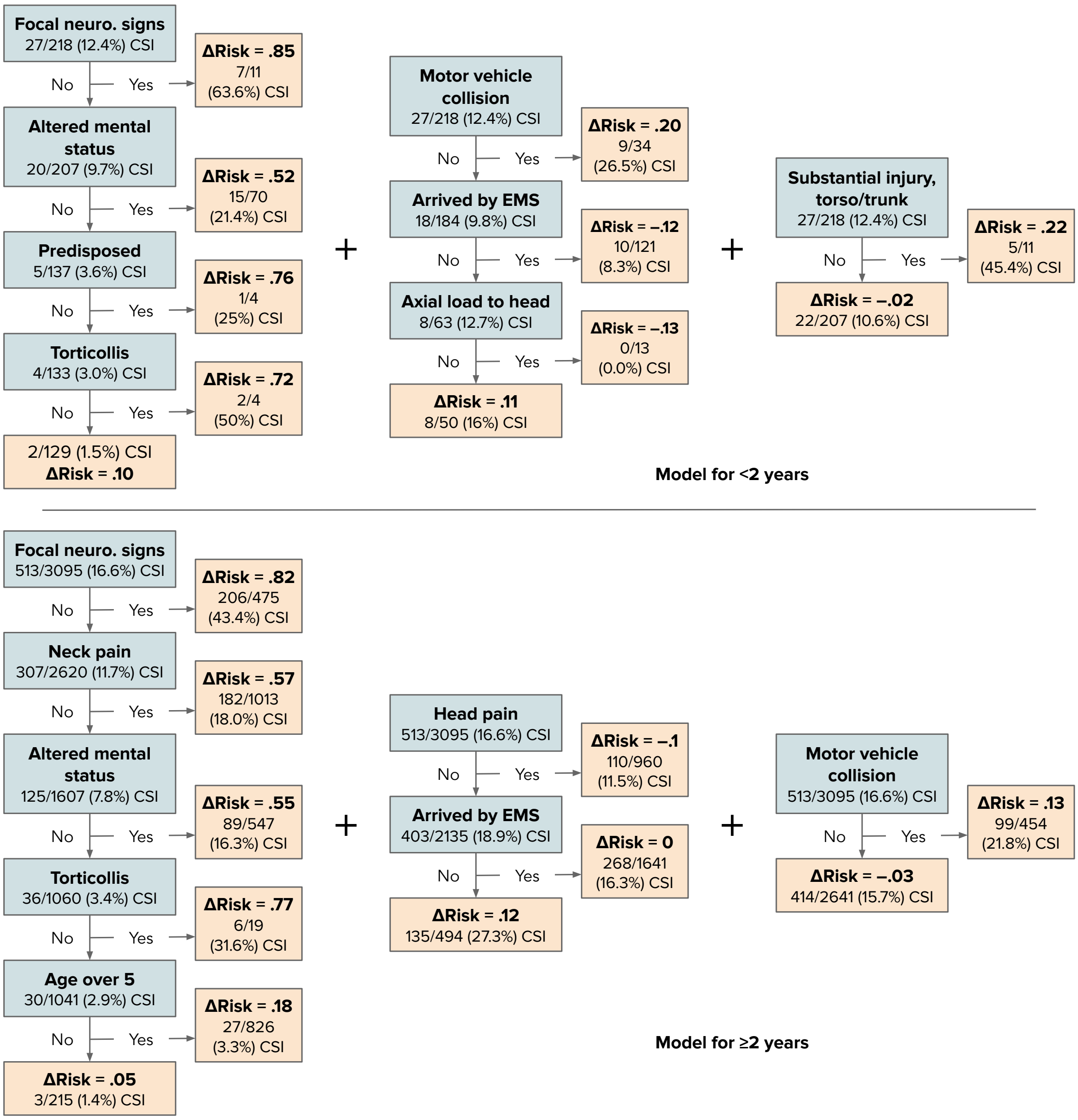

Figure 4. G-FIGS model fitted to the CSI dataset. Achieves 97.0% sensitivity and 33.9% specificity (training). The left tree for <2 *yrs* gives large Δ Risk to active features, and on its own provides sensitivity of 99%. Counterintuitively, the middle tree assigns Δ Risk < 0 for patients arriving by ambulance (*EMS*) or with head injuries that affect the spine (*axial load*). However, adding this second tree results in boosted specificity (increase of 8.7%) with a tiny reduction in sensitivity (decrease of 0.4%), indicating that G-FIGS adaptively tunes the sensitivity-specificity tradeoff.

Again, the features used by G-FIGS partially match previous work [21]; e.g., features such as *focal neurological signs*, *neck pain*, and *altered mental status* are all known to increase the risk of CSI. The ≥ 2 *yrs* model utilizes two features requiring verbal responses (*neck pain*, *head pain*), which are not in the <2 *yrs* model.

4.3.3 IAI

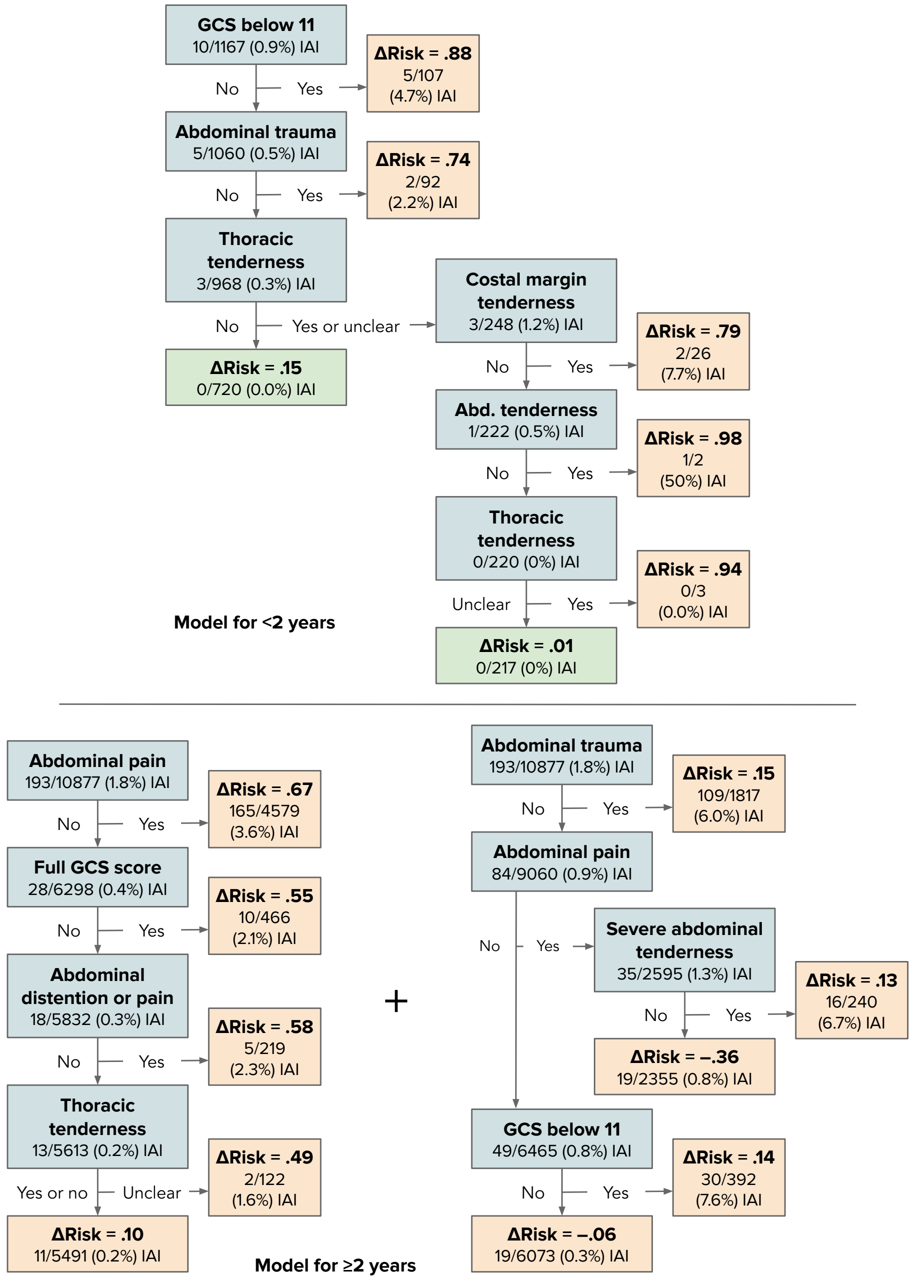

Figure 5. G-FIGS model fitted to the IAI dataset. Achieves 95.1% sensitivity and 50.8% specificity (training). Features predictive of intra-abdominal injury such as *abdominal trauma*, *GCS score*, and *abdominal pain* match previous work [22]. However, G-FIGS learns a different cutoff for *GCS score* (score of 11) compared to Holmes et al. (score of 13). Note that the younger group only uses *tenderness*, which can evaluated without verbal input from the patient, whereas the older group uses *pain*, which requires a verbal response.

	$<$2 *yrs* group			$\geq$2 *yrs* group		
Maximum tree splits:	8	12	16	8	12	16
TAO (1 iter)	**15.1** (6.7)	15.1 (6.7)	14.4 (6.1)	**14.1** (7.8)	14.1 (7.8)	8.9 (5.9)
TAO (5 iter)	**14.4** (6.1)	0.0 (0.0)	0.0 (0.0)	**8.9** (5.9)	3.1 (0.9)	1.5 (0.7)
CART-SEP	**15.1** (6.7)	14.4 (6.1)	0.0 (0.0)	**14.0** (7.8)	8.9 (5.9)	3.1 (0.9)
FIGS-SEP	**13.7** (5.9)	0.0 (0.0)	0.0 (0.0)	**23.1** (8.8)	13.0 (7.4)	7.8 (5.6)
G-CART w/ LR ($C = 2.8$)	**7.9** (6.7)	3.1 (2.1)	3.5 (1.7)	19.0 (8.8)	**21.8** (8.4)	2.1 (0.6)
G-CART w/ LR ($C = 0.1$)	**20.4** (8.6)	8.3 (6.6)	10.1 (6.7)	12.7 (7.6)	**14.9** (7.1)	3.6 (0.9)
G-CART w/ GB ($N = 100$)	**19.8** (8.3)	7.2 (6.3)	7.6 (6.1)	13.3 (8.0)	**21.4** (8.5)	9.0 (5.6)
G-CART w/ GB ($N = 50$)	**26.8** (9.7)	8.1 (6.3)	8.4 (6.1)	13.3 (8.0)	**21.4** (8.5)	9.7 (5.6)
G-FIGS w/ LR ($C = 2.8$)	**14.9** (8.5)	7.5 (5.4)	8.1 (6.9)	41.0 (8.7)	**48.1** (8.2)	35.6 (8.9)
G-FIGS w/ LR ($C = 0.1$)	**31.0** (9.4)	23.1 (9.1)	25.9 (9.7)	46.9 (8.4)	**48.2** (8.4)	33.7 (8.9)
G-FIGS w/ GB ($N = 100$)	**24.5** (8.6)	24.0 (9.3)	21.2 (8.7)	**47.5** (8.5)	47.5 (8.2)	27.9 (8.6)
G-FIGS w/ GB ($N = 50$)	**32.1** (9.6)	18.3 (8.2)	12.7 (6.9)	47.5 (8.5)	**53.2** (7.3)	28.4 (8.3)

(a)

Group membership model:	LR ($C = 2.8$)	LR ($C = 0.1$)	GB ($N = 100$)	GB ($N = 50$)
G-CART ($<$2, $\geq$2 combined)	**27.8** (6.0)	21.5 (5.9)	19.0 (5.7)	27.1 (6.5)
G-FIGS ($<$2, $\geq$2 combined)	51.3 (5.8)	54.5 (6.2)	**57.4** (5.6)	44.6 (7.4)

(b)

Table 5. Hyperparameter selection table for the TBI dataset; the metric shown is specificity at 94% sensitivity on the validation set. Standard errors are shown in parentheses. First, the best-performing maximum of tree splits is selected for each method or combination of method and membership model (a). This is done separately for each data group. Next, the best membership model is selected for G-CART and G-FIGS using the overall performance of the best models from (a) across both data groups (b). The two-stage validation process ensures that the $<$2 *yrs* and $\geq$2 *yrs* groups use the same group membership probabilities, which we have found leads to better performance than allowing them to use different membership models. Metrics shown are averages across the 10 validation sets, but hyperparameter selection was done independently for each of the 10 data splits.

4.4 Hyperparameter selection

Data splitting We use 10 random training/validation/test splits for each dataset, performing hyperparameter selection separately on each. There are two reasons we choose not to use a fixed test set. First, the small number of positive instances in our datasets makes our primary metrics (specificity at high sensitivity levels) noisy, so averaging across multiple splits makes the results more stable. Second, the works that introduced the TBI, IAI, and CSI datasets did not publish their test sets, as it is not as common to do so in the medical field as it is in machine learning, making the choice of test set unclear. For TBI and CSI, we simply use the random seeds 0 through 10. For IAI, some filtering of seeds is required due to the low number of positive examples; we reject seeds that do not allocate positive examples evenly enough between each split (a ratio of negative to positive outcomes over 200 in any split).

Class weights Due to the importance of achieving high sensitivity, we upweight positive instances in the loss by the inverse proportion of positive instances in the dataset. This results in class weights of about 7:1 for CSI, 112:1 for TBI, and 60:1 for IAI. These weights are fixed for all methods.

Hyperparameter settings Due to the relatively small number of positive examples in all datasets, we keep the hyperparameter search space small to avoid overfitting. We vary the maximum number of tree splits from 8 to 16 for all methods and the maximum number of update iterations from 1 to 5 for TAO. The options of group membership model are logistic regression with L2 regularization and gradient-boosted trees friedman2001greedy. For both models, we simply include two hyperparameter settings: a less-regularized version and a more-regularized version, by varying the inverse regularization strength (C) for logistic regression and the number of trees (N) for gradient-boosted trees. We initially experimented with random forests and CART, but found them to lead to poor downstream performance. Random forests tended to separate the groups too well in terms of estimated probabilities, leading to little information sharing between groups, while CART did not provide unique enough membership probabilities, since CART probability estimates are simply within-node class proportions.

Validation metrics We use the highest specificity achieved when sensitivity is at or above 94% as the metric for validation. If this metric is tied between different hyperparameter settings of the same model, specificity at 90% sensitivity is used as the tiebreaker. For the IAI dataset, only specificity at 90% sensitivity is used, since the relatively small number of positive examples makes high sensitivity metrics noisier than usual. If there is still a tie at 90% sensitivity, the smaller model in terms of number of tree splits is chosen.

Validation of group membership model Hyperparameter selection for G-FIGS and G-CART is done in two stages due to the need to select the best group membership model. First, the best-performing maximum of tree splits is selected for each combination of method and membership model. This is done separately for each data group. Next, the best membership model is selected using the overall performance of the best models across both data groups. The two-stage validation process ensures that the <2 *yrs* and ≥ 2 *yrs* groups use the same group membership probabilities, which we have found performs better than allowing different sub-models of G-FIGS to use different membership models.

5. Simulation

In addition our evaluations on clinical datasets, we evaluate G-FIGS under a simple simulation involving heterogeneous data. The data-generating process is multivariate Gaussian with four clusters and two meta-clusters which share the same relationship between X and Y, visualized in Figure 6. There are two variables of interest, X_1 and X_2, and 10 noise variables. Each cluster is centered at a different value of X_1; the first meta-cluster consists of the clusters centered at $X_1 = 0$ and $X_1 = 2$, which share the relationship $Y = X_2 > 0$, while the second consists of the clusters centered at $X_1 = 4$ and $X_1 = 6$, which share the relationship $Y = X_2 > 2$. X_1 and X_2 have variance 1 and all noise variables have variance 2; additionally, zero-mean noise with variance 2 is added to X_1 and X_2.

The four clusters are then treated as four groups, to which separate models are fitted. If the intuition behind G-FIGS is correct, G-FIGS should assign relatively higher probabilities to points that are within a given cluster's meta-cluster, and relatively lower probabilities to points in the other meta-cluster. In comparison to fitting completely separate models, this should increase the amount of data available for learning the two rules, thereby counteracting noise and resulting in better performance. On the other hand, if one model is fit to all of the data, we expect the lack of group-awareness to hurt performance (i.e. the crucial split at $X_1 = 3$ may be missed since it does not significantly reduce entropy). Our evaluation suggests that this is the case; as shown in Table 6, G-CART and G-FIGS significantly outperform baseline methods.

We do not perform any hyperparameter selection; we fix the maximum number of tree splits to be 1 for the probability-weighted models and -*SEP* models, and 4 for the models fit to all the data. The rationale for this is that 3 splits are sufficient to ideally model the entire data-generating process (splits at $X_1 = 3$, $X_2 = 0$, and $X_2 = 2$) and 1 split is sufficient for each cluster. Note that when only one split is used, G-CART and G-FIGS are the same algorithm. Logistic regression is used to fit the group membership model.

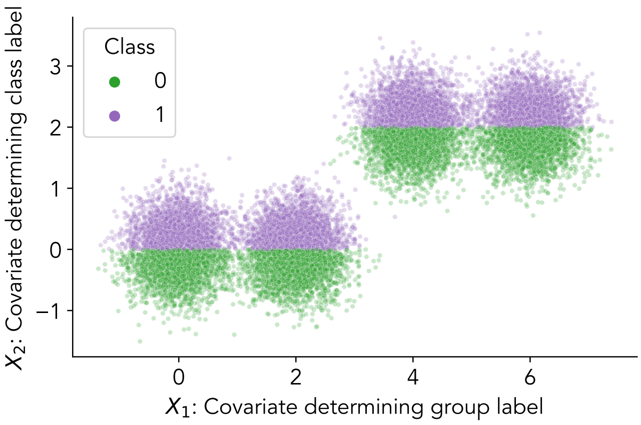

Figure 6. Visualization of the data-generating process for the simulation. Each cluster represents a group for G-FIGS. The two clusters on the left and two clusters on the right share a prediction rule, presenting a simple case where sharing data between groups can help performance. Noise variables are not pictured, and the variances of X_1 and X_2 are reduced for a clearer visualization.

	ROC AUC	APS	Accuracy	F1
TAO	.376 (.07)	.498 (.04)	59.0 (.02)	58.0 (.04)
TAO-SEP	.475 (.04)	.573 (.03)	58.3 (.02)	60.4 (.03)
CART	.370 (.07)	.495 (.04)	56.5 (.02)	54.7 (.03)
CART-SEP	.475 (.04)	.573 (.03)	58.3 (.02)	60.4 (.03)
FIGS	.470 (.04)	.539 (.04)	58.5 (.02)	55.5 (.03)
FIGS-SEP	.475 (.04)	.573 (.03)	58.3 (.02)	60.4 (.03)
G-CART / G-FIGS	**.550** (.03)	**.644** (.03)	**65.8** (.03)	**63.9** (.04)

Table 6. Unlike the clinical datasets, the simulation data is class-balanced and lacks a medical context, so we report area under the ROC curve, average precision score, accuracy, and F1 score instead of specificity metrics. Because only one split per cluster is computed for G-CART and G-FIGS they reduce to the exact same algorithm, so their results are shown together.

6. Discussion

G-FIGS makes an important step towards interpretable modeling of heterogeneous data in the context of high-stakes clinical decision-making, with interesting avenues for future work. The fitted models show promise, but require external clinical validation before potential use. Our scope was limited to age-based splits in the clinical domain, but the behavior of G-FIGS with temporal, geographical, or demographic splits could be studied as well, on these or other datasets.

Here we utilized datasets that used prospective data collection (TBI, IAI) or case-matched data (CSI) to avoid bias. However, future work will focus on prognostic models, (e.g. risk for future disease) requiring the evaluation of data that may be collected pre- and post-diagnosis or therapeutic measures. For instance, datasets that are collected both before and after the introduction of a vaccine, development of a new form of treatment, or arrival of a new disease variant present a problem in terms of temporal heterogeneity. When demographic heterogeneity is present within data, the ability of our new method to improve fairness metrics could be evaluated. Additionally, there are many methodological extensions to explore, such as data-driven identification of input data groups and schemes for feature weighting in addition to instance weighting.

References

[1] Yan Shuo Tan et al. "Fast interpretable greedy-tree sums (FIGS)". In: *arXiv preprint arXiv:2201.11931* (2022).

[2] Chandan Singh et al. "imodels: a python package for fitting interpretable models". In: *Journal of Open Source Software* 6.61 (2021), p. 3192. DOI: `10.21105/joss.03192`. URL: `https://doi.org/10.21105/joss.03192`.

[3] Geoffrey S Ginsburg and Kathryn A Phillips. "Precision medicine: from science to value". In: *Health Affairs* 37.5 (2018), pp. 694–701.

[4] Nathan Kuppermann et al. "Identification of children at very low risk of clinically-important brain injuries after head trauma: a prospective cohort study". In: *The Lancet* 374.9696 (2009), pp. 1160–1170.

[5] W James Murdoch et al. "Definitions, methods, and applications in interpretable machine learning". In: *Proceedings of the National Academy of Sciences* 116.44 (2019), pp. 22071–22080. DOI: `10.1073/pnas.1900654116`.

[6] Cynthia Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature Machine Intelligence* 1.5 (2019), pp. 206–215. DOI: `10.1038/s42256-019-0048-x`.

[7] Shenyang Guo and Mark W Fraser. *Propensity score analysis: Statistical methods and applications*. Vol. 11. SAGE publications, 2014.

[8] Fuzhen Zhuang et al. "A comprehensive survey on transfer learning". In: *Proceedings of the IEEE* 109.1 (2020), pp. 43–76.

[9] J. Ross Quinlan. "Induction of decision trees". In: *Machine learning* 1.1 (1986), pp. 81–106.

[10] Leo Breiman et al. *Classification and regression trees*. Chapman and Hall/CRC, 1984.

[11] Berk Ustun and Cynthia Rudin. "Supersparse linear integer models for optimized medical scoring systems". In: *Machine Learning* 102.3 (2016), pp. 349–391. DOI: `10.1007/s10994-015-5528-6`.

[12] Wooseok Ha et al. "Adaptive wavelet distillation from neural networks through interpretations". In: *Advances in Neural Information Processing Systems* 34 (2021).

[13] Jerome H Friedman, Bogdan E Popescu, et al. "Predictive learning via rule ensembles". In: *The Annals of Applied Statistics* 2.3 (2008), pp. 916–954.

[14] Abhineet Agarwal et al. “Hierarchical Shrinkage: improving the accuracy and interpretability of tree-based methods”. In: *arXiv preprint arXiv:2202.00858* (2022).

[15] Jimmy Lin et al. “Generalized and scalable optimal sparse decision trees”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 6150–6160.

[16] Dimitris Bertsimas et al. “Prediction of cervical spine injury in young pediatric patients: an optimal trees artificial intelligence approach”. In: *Journal of Pediatric Surgery* 54.11 (2019), pp. 2353–2357.

[17] Ian G Stiell et al. “The Canadian CT Head Rule for patients with minor head injury”. In: *The Lancet* 357.9266 (2001), pp. 1391–1396.

[18] Aaron E Kornblith et al. “Predictability and Stability Testing to Assess Clinical Decision Instrument Performance for Children After Blunt Torso Trauma”. In: *medRxiv* (2022).

[19] James F Holmes et al. “Identification of children with intra-abdominal injuries after blunt trauma”. In: *Annals of emergency medicine* 39.5 (2002), pp. 500–509.

[20] Leo Breiman. “Random forests”. In: *Machine learning* 45.1 (2001), pp. 5–32.

[21] Julie C Leonard et al. “Cervical spine injury risk factors in children with blunt trauma”. In: *Pediatrics* 144.1 (2019).

[22] James F Holmes et al. “Identification of children with intra-abdominal injuries after blunt trauma”. In: *Annals of emergency medicine* 62.2 (2013), pp. 107–116.

[23] Miguel A Carreira-Perpinán and Pooya Tavallali. “Alternating optimization of decision trees, with application to learning sparse oblique trees”. In: *Advances in neural information processing systems* 31 (2018).

Extended results

	Traumatic brain injury				Cervical spine injury				Intra-abdominal injury			
	92%	94%	96%	98%	92%	94%	96%	98%	92%	94%	96%	98%
TAO	6.2 (5.9)	6.2 (5.9)	0.4 (0.4)	0.4 (0.4)	41.5 (0.9)	21.2 (6.6)	0.2 (0.2)	0.2 (0.2)	0.2 (0.2)	0.2 (0.2)	0.0 (0.0)	0.0 (0.0)
TAO-SEP	26.7 (6.4)	13.9 (5.4)	10.4 (5.5)	2.4 (1.5)	32.5 (4.9)	7.0 (1.6)	5.4 (0.7)	2.5 (1.0)	12.1 (1.7)	8.5 (2.0)	2.0 (1.3)	0.0 (0.0)
CART	20.9 (8.8)	14.8 (7.6)	7.8 (5.8)	2.1 (0.6)	38.6 (3.6)	13.7 (5.7)	1.5 (0.6)	1.1 (0.4)	11.8 (5.0)	2.7 (1.0)	1.6 (0.5)	1.4 (0.5)
CART-SEP	26.6 (6.4)	13.8 (5.4)	10.3 (5.5)	2.4 (1.5)	32.1 (5.1)	7.8 (1.5)	5.4 (0.7)	2.5 (1.0)	11.0 (1.6)	9.3 (1.8)	2.8 (1.4)	0.0 (0.0)
G-CART	15.5 (5.5)	13.5 (5.7)	6.4 (2.2)	3.0 (1.5)	38.5 (3.4)	15.2 (4.8)	4.9 (1.0)	3.9 (1.1)	11.7 (1.3)	10.1 (1.6)	3.8 (1.3)	0.7 (0.4)
FIGS	23.8 (9.0)	18.2 (8.5)	12.1 (7.3)	0.4 (0.3)	39.1 (3.0)	33.8 (2.4)	24.2 (3.2)	**16.7** (3.9)	**32.1** (5.5)	13.7 (6.0)	1.4 (0.8)	0.0 (0.0)
FIGS-SEP	39.9 (7.9)	19.7 (6.8)	**17.5** (7.0)	2.6 (1.6)	38.7 (1.6)	33.1 (2.0)	20.1 (2.6)	3.9 (2.2)	18.8 (4.4)	9.2 (2.2)	2.6 (1.7)	0.9 (0.8)
G-FIGS	**42.0** (6.6)	**23.0** (7.8)	14.7 (6.5)	**6.4** (2.8)	**42.2** (1.3)	**36.2** (2.3)	**28.4** (3.8)	15.7 (3.9)	29.7 (6.9)	**18.8** (6.6)	**11.7** (5.1)	**3.0** (1.3)

Table 7. Test set prediction results averaged over 10 random data splits, with corresponding standard error in parentheses. Values in columns labeled with a sensitivity percentage (e.g. 92%) are best specificity achieved at the given level of sensitivity or greater. G-FIGS provides the best performance overall in the high-sensitivity regime. G-CART attains the best ROC curves, while TAO is strongest in terms of F1 score.