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

Keyan Nasseri [1] **Chandan Singh** [1] **James Duncan** [2] **Aaron Kornblith** [3 4] **Bin Yu** [1 2 5]

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 grouped by age or treatment site), our method first estimates group membership probabilities for each instance. Then, it uses these estimates as instance weights in FIGS (Tan et al., 2022), to grow a set of decision trees whose values sum to the final prediction. We call this new method Group Probability-Weighted Tree Sums (G-FIGS). G-FIGS achieves state-of-the-art prediction performance on important clinical datasets; 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 rules below 16 in FIGS, the final models remain interpretable, and we find that their rules match medical domain expertise. All code, data, and models are released on Github. [1]

1. Introduction

Recent advances in machine learning (ML) have led to impressive increases in predictive performance. However, ML 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 (Ginsburg & Phillips, 2018). Groups may differ dramatically and require distinct features for high predictive performance on the same outcome; e.g., infants may be nonverbal, excluding 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, a lack of interpretability is unacceptable in healthcare and many other domains (Murdoch et al., 2019; Rudin, 2019). 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[2]), 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 (Fig 1). G-FIGS first fits a classifier to predict group membership probabilities for each input instance (Fig 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-

[1]EECS Department, UC Berkeley [2]Group in Biostatistics, UC Berkeley [3]Emergency Medicine, UC San Francisco [4]Pediatrics Department, UC San Francisco [5]Statistics Department, UC Berkeley. Correspondence to: Bin Yu <binyu@berkeley.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

[1]Group Probability-Weighted Tree Sums is integrated into the imodels package csinva/imodels (Singh et al., 2021) with an sklearn-compatible API. Experiments for reproducing the results here can be found at Yu-Group/imodels-experiments.

[2]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.

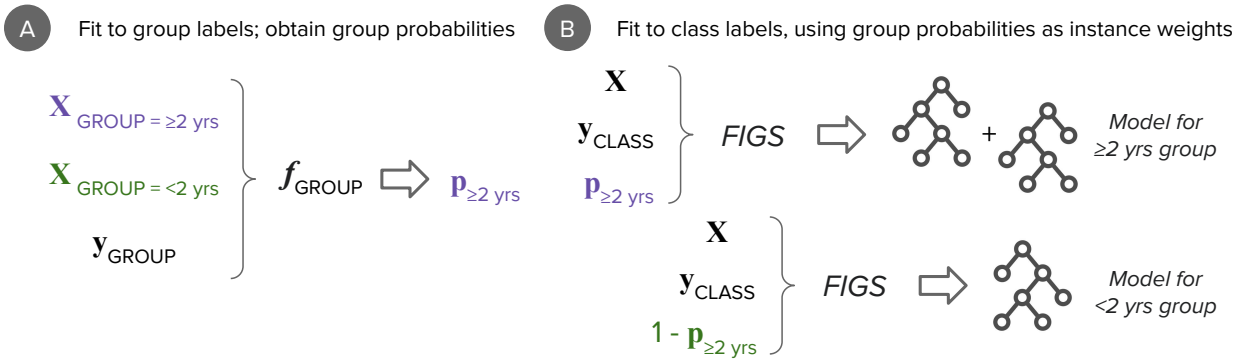

Figure 1. Overview of G-FIGS. **(A)** First, the covariates of each instance in a dataset are used to estimate an instance-specific probability of membership in each of the pre-specified groups in the data (e.g., patients of age <2 *yrs* and ≥ 2 *yrs*). **(B)** Next, these membership probabilities are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

FIGS results in a separate highly accurate rule-based model for each group. We test G-FIGS on three real-world clinical datasets (Sec 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 ML methods in terms of specificity achieved at the high levels of sensitivity 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.

2. 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 (Guo & Fraser, 2014). More generally, this work is related to the literature on transfer learning (Zhuang et al., 2020), 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 (Quinlan, 1986; Breiman et al., 1984). Recent works have focused on improving the predictive performance of intrinsically interpretable methods (Ustun & Rudin, 2016; Ha et al., 2021), particularly for rule-based models (Friedman et al., 2008; Agarwal et al., 2022; Tan et al., 2022; Lin et al., 2020), 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 ML models, particularly in emergency medicine (Bertsimas et al., 2019; Stiell et al., 2001; Kornblith et al., 2022; Holmes et al., 2002).

3. 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 $P(G|X)$ (Fig 1A).[3] 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.

[3] 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).

Fitting group probability-weighted FIGS. In the second stage (Fig 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 (Tan et al., 2022) performs best when both interpretability and high predictive performance are required.[4] By greedily fitting a sum of trees, FIGS effectively allocates a small budget of rules to different types of structure in data.

4. Results and discussion

Datasets and data cleaning. Table 1 shows the main datasets under consideration here. They each constitute a large-scale multi-site data aggregation by the Pediatric Emergency Care Applied Research Network, 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. 2009), 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. [5] 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.

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) (Kuppermann et al., 2009), intra-abdominal injury (IAI) (Holmes et al., 2002), and cervical spine injury (CSI) (Leonard et al., 2019).

G-FIGS predicts well. Table 2 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 (Breiman et al., 1984), FIGS (Tan et al., 2022), and Tree-Alternating Optimization TAO (Carreira-Perpinán & Tavallali, 2018)). 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 2 to 16 maximum rules (full details of this and other hyperparameters are in Supp C).

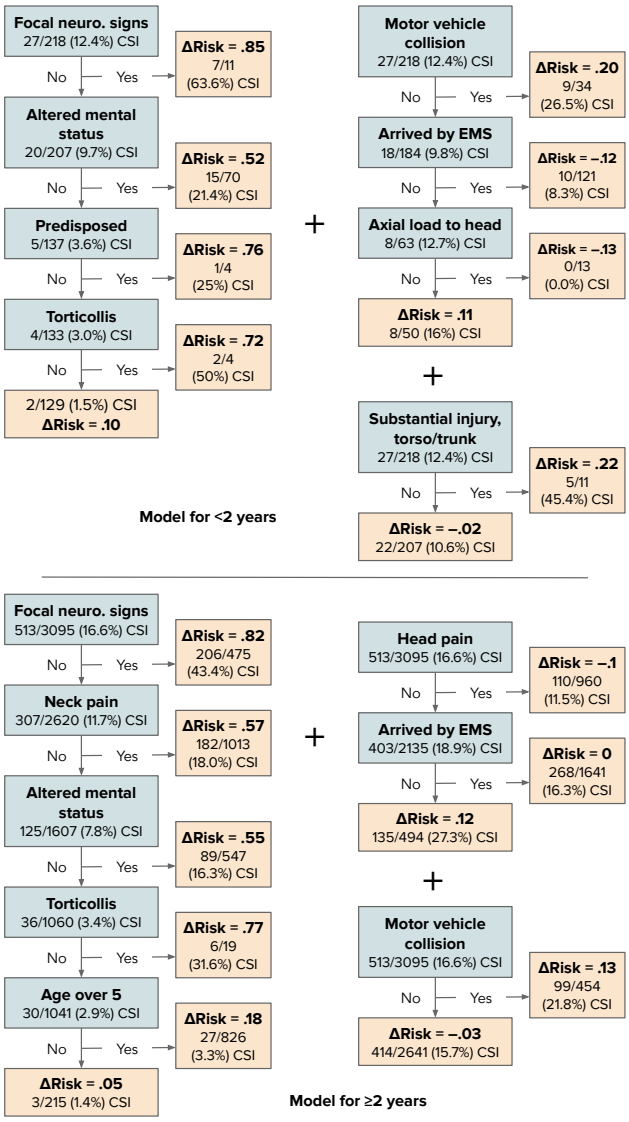

Figure 2. G-FIGS models fitted to the CSI dataset are concise, highly predictive, and match known medical knowledge. The left tree for <2 *yrs* has high sensitivity (99%); adding the upper right tree boosts specificity by 8.7% and decreases sensitivity by 0.4%.

[4] When interpretability is not critical, the same weighting procedure could also be applied to black-box models, such as Random Forest (Breiman, 2001).

[5] Details, along with the openly released clean data can be found in Supp D. Additionally, simulation results showing the effectiveness of G-FIGS are shown in Supp B.

	Traumatic brain injury				Cervical spine injury				Intra-abdominal injury			
Sensitivity level:	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 2. 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.

Traumatic brain injury		Cervical spine injury		Intra-abdominal injury	
Variable	Coefficient	Variable	Coefficient	Variable	Coefficient
No fontanelle bulging	3.62	Neck tenderness	2.44	Bike injury	2.01
Amnesia	2.07	Neck pain	2.18	Abdomen pain	1.66
Pedestrian 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 3. 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.

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. For the instance-weighted methods, we treat the choice of group membership model as a hyperparameter, and select the best model according to the downstream performance of the final decision rule on the validation set.

Table 3 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*).

Interpreting the outcome model. Fig 2 shows the G-FIGS model on the CSI dataset, 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.[6] The features used by each group are overlapping and reasonable, matching medical domain knowledge and partially matching previous work (Leonard et al., 2019); e.g., features such as *focal neuro signs*, *neck pain*, and *altered mental status* are all known to increase the risk of CSI. 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., *substantial torso injury*), while two of the ≥ 2 *yrs* features require verbal responses (*neck pain*, *head pain*). Supp A shows fitted models for other datasets/methods; 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.

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 is 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. Additionally, there are many methodologi-

[6] Δ Risk is not simply equivalent to the fraction of patients with CSI since (i) G-FIGS uses patients from both groups and (ii) each tree in FIGS fits the residuals of the others.

cal extensions to explore, such as data-driven identification of input data groups and schemes for feature weighting in addition to instance weighting.

Acknowledgements

We gratefully acknowledge partial support from NSF Grants DMS-1613002, 1953191, 2015341, IIS 1741340, the Center for Science of Information (CSoI), an NSF Sci- ence and Technology Center, under grant agreement CCF- 0939370, NSF grant 2023505 on Collaborative Research: Foundations of Data Science Institute (FODSI), the NSF and the Simons Foundation for the Collaboration on the Theoretical Foundations of Deep Learning through awards DMS-2031883 and 814639, and a grant from the Weill Neurohub.

References

Agarwal, A., Tan, Y. S., Ronen, O., Singh, C., and Yu, B. Hierarchical shrinkage: improving the accuracy and interpretability of tree-based methods. *arXiv preprint arXiv:2202.00858*, 2022.

Bertsimas, D., Masiakos, P. T., Mylonas, K. S., and Wiberg, H. Prediction of cervical spine injury in young pediatric patients: an optimal trees artificial intelligence approach. *Journal of Pediatric Surgery*, 54(11):2353–2357, 2019.

Breiman, L. Random forests. *Machine learning*, 45(1):5–32, 2001.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. J. *Classification and regression trees*. Chapman and Hall/CRC, 1984.

Carreira-Perpinán, M. A. and Tavallali, P. Alternating optimization of decision trees, with application to learning sparse oblique trees. *Advances in neural information processing systems*, 31, 2018.

Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189–1232, 2001.

Friedman, J. H., Popescu, B. E., et al. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3):916–954, 2008.

Ginsburg, G. S. and Phillips, K. A. Precision medicine: from science to value. *Health Affairs*, 37(5):694–701, 2018.

Guo, S. and Fraser, M. W. *Propensity score analysis: Statistical methods and applications*, volume 11. SAGE publications, 2014.

Ha, W., Singh, C., Lanusse, F., Upadhyayula, S., and Yu, B. Adaptive wavelet distillation from neural networks through interpretations. *Advances in Neural Information Processing Systems*, 34, 2021.

Holmes, J. F., Sokolove, P. E., Brant, W. E., Palchak, M. J., Vance, C. W., Owings, J. T., and Kuppermann, N. Identification of children with intra-abdominal injuries after blunt trauma. *Annals of emergency medicine*, 39(5):500–509, 2002.

Holmes, J. F., Lillis, K., Monroe, David Borgialli, D., Kerrey, B. T., et al. Identifying children at very low risk of clinically important blunt abdominal injuries. *Annals of emergency medicine*, 62(2): 107–116, 2013.

Kornblith, A. E., Singh, C., Devlin, G., Addo, N., Streck, C. J., Holmes, J. F., Kuppermann, N., Grupp-Phelan, J., Fineman, J., Butte, A. J., et al. Predictability and stability testing to assess clinical decision instrument performance for children after blunt torso trauma. *medRxiv*, 2022.

Kuppermann, N., Holmes, J. F., Dayan, P. S., Hoyle, J. D., Atabaki, S. M., Holubkov, R., Nadel, F. M., Monroe, D., Stanley, R. M., Borgialli, D. A., et al. Identification of children at very low risk of clinically-important brain injuries after head trauma: a prospective cohort study. *The Lancet*, 374(9696):1160–1170, 2009.

Leonard, J. C., Browne, L. R., Ahmad, F. A., Schwartz, H., Wallendorf, M., Leonard, J. R., Lerner, E. B., and Kuppermann, N. Cervical spine injury risk factors in children with blunt trauma. *Pediatrics*, 144(1), 2019.

Lin, J., Zhong, C., Hu, D., Rudin, C., and Seltzer, M. Generalized and scalable optimal sparse decision trees. In *International Conference on Machine Learning*, pp. 6150–6160. PMLR, 2020.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116 (44):22071–22080, 2019. doi: 10.1073/pnas.1900654116.

Quinlan, J. R. Induction of decision trees. *Machine learning*, 1(1): 81–106, 1986.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. doi: 10.1038/s42256-019-0048-x.

Singh, C., Nasseri, K., Tan, Y. S., Tang, T., and Yu, B. imodels: a python package for fitting interpretable models. *Journal of Open Source Software*, 6(61):3192, 2021. doi: 10.21105/joss.03192. URL https://doi.org/10.21105/joss.03192.

Stiell, I. G., Wells, G. A., Vandemheen, K., Clement, C., Lesiuk, H., Laupacis, A., McKnight, R. D., Verbeek, R., Brison, R., Cass, D., et al. The canadian ct head rule for patients with minor head injury. *The Lancet*, 357(9266):1391–1396, 2001.

Tan, Y. S., Singh, C., Nasseri, K., Agarwal, A., and Yu, B. Fast interpretable greedy-tree sums (figs). *arXiv preprint arXiv:2201.11931*, 2022.

Ustun, B. and Rudin, C. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3): 349–391, 2016. doi: 10.1007/s10994-015-5528-6.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

A. Fitted models

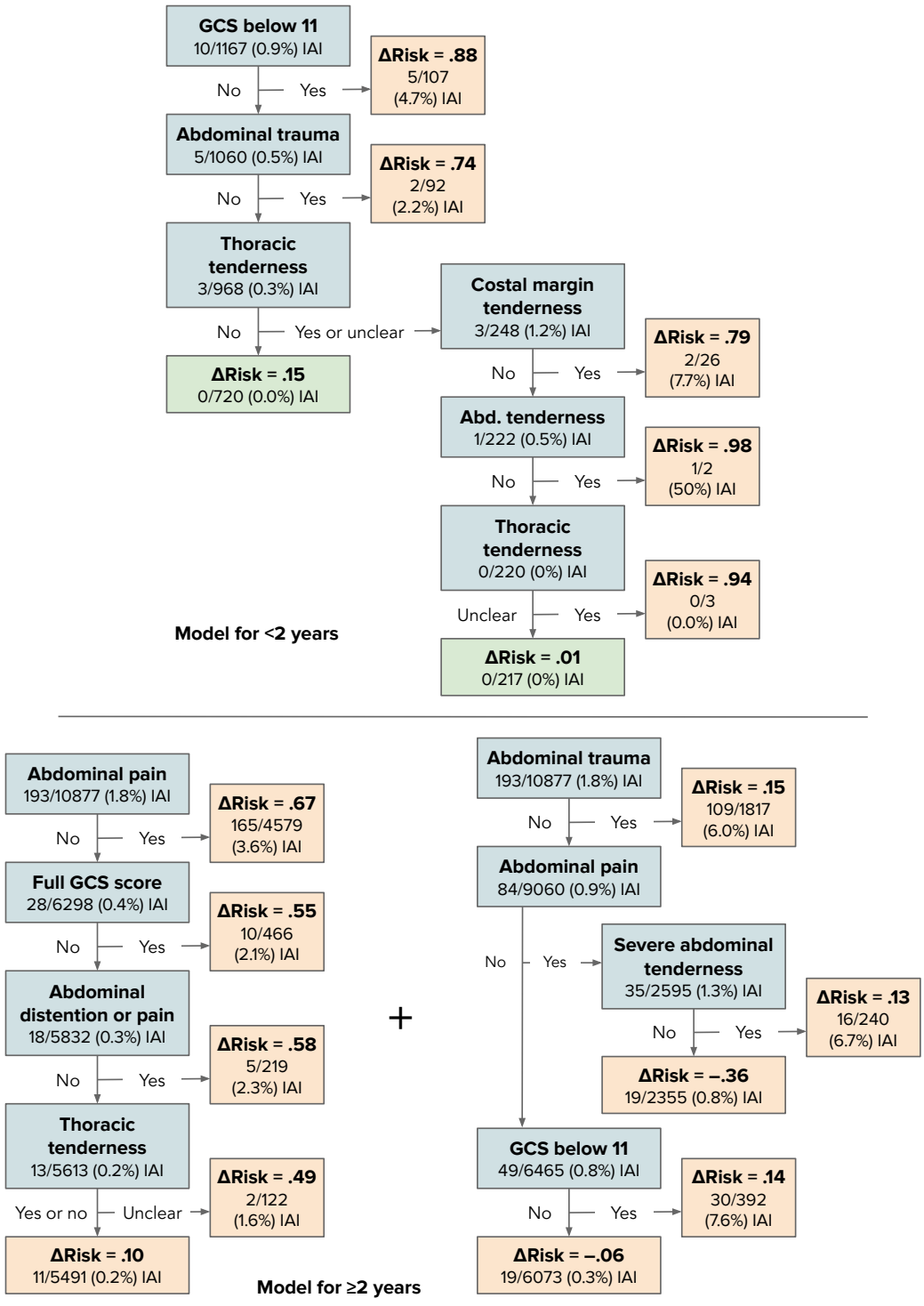

Figure 3. G-FIGS model fitted to the IAI dataset. Note that the younger group only uses *tenderness*, which can evaluated without verbal input from the patient, whereas the older group uses *pain*, which requires a verbal response. Achieves 95.1% sensitivity and 50.8% specificity (training).

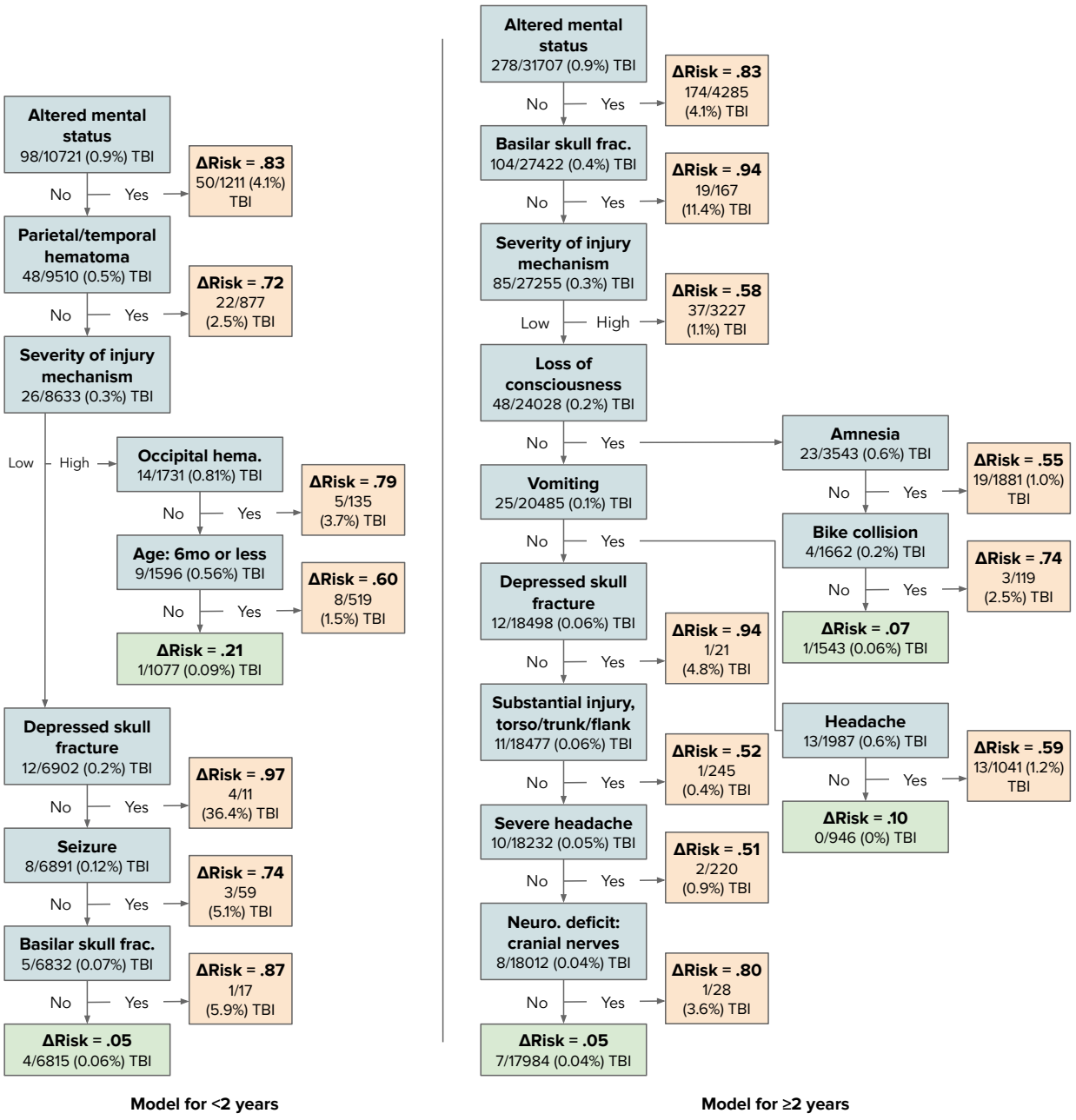

Figure 4. G-FIGS model fitted to the TBI dataset. Interestingly, in this case G-FIGS learns only a single tree for each group. Note that the model for the older group utilizes the *Headache* and *Severe Headache* features, which require a verbal response. Achieves 97.1% sensitivity and 58.9% specificity (training).

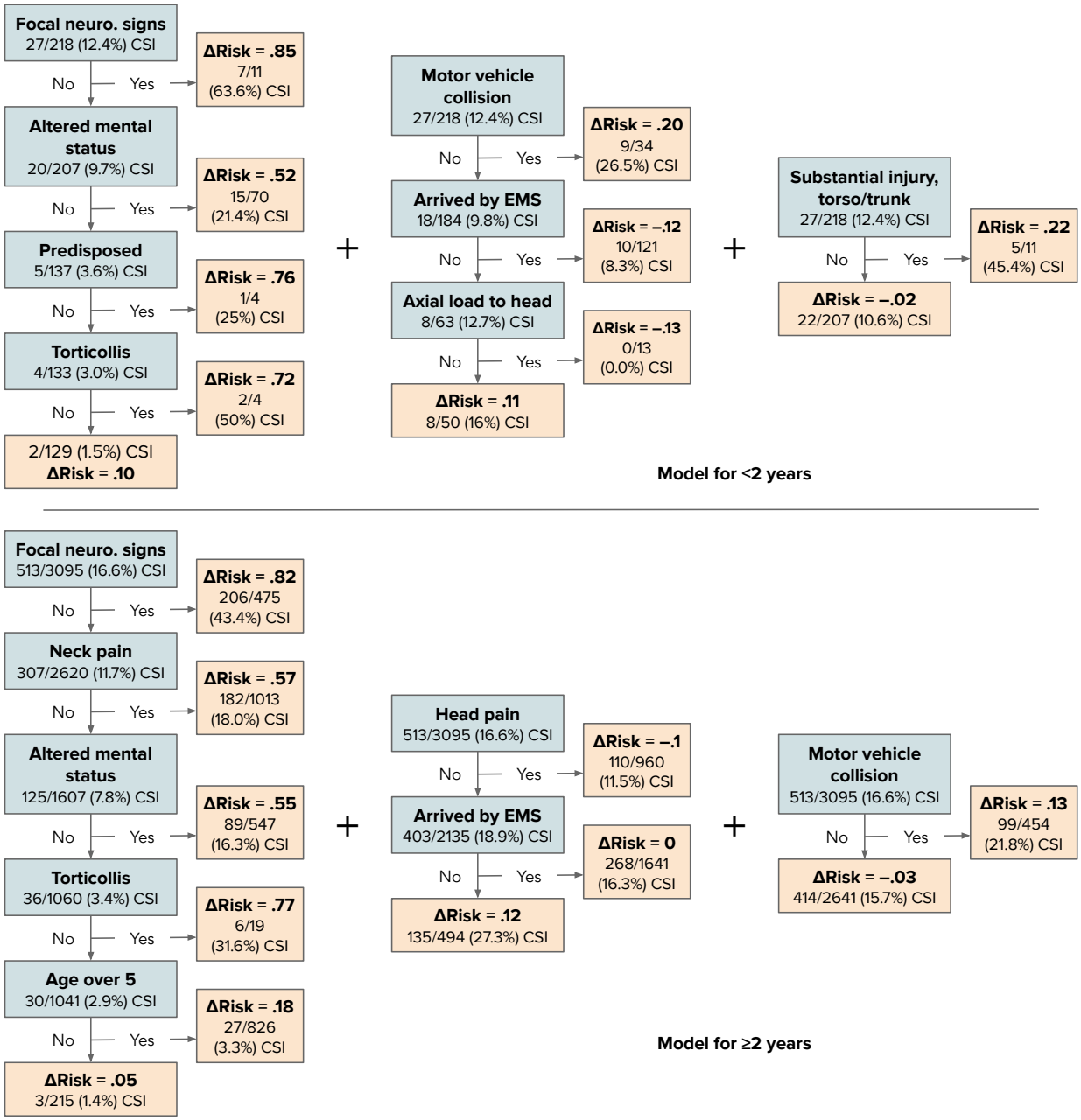

Figure 5. G-FIGS model fitted to the CSI dataset (duplicated from the main text Fig 2). Achieves 97.0% sensitivity and 33.9% specificity (training). The left tree for <2 *yrs* gives large Δ Risk to active features, and on its own provides sensitivity of 99%. Counterintuitively, the middle tree assigns 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.

B. 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 Fig 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 4, 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.

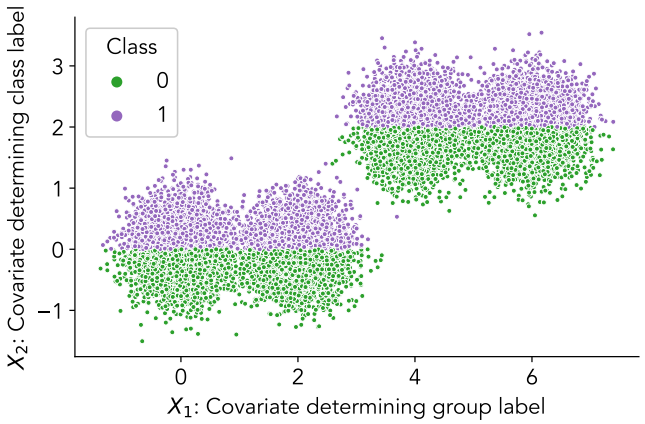

Figure 6. Visualization of the data-generating process for the simulation. Each cluster represents a group for G-FIGS. The two clusters on the left and two clusters on the right share a prediction rule, presenting a simple case where sharing data between groups can help performance. Noise variables are 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 4. 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.

C. Hyperparameter selection

	<2 *yrs* group			≥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 *yrs*, ≥2 *yrs* models combined)	**27.8** (6.0)	21.5 (5.9)	19.0 (5.7)	27.1 (6.5)
G-FIGS (<2 *yrs*, ≥2 *yrs* models 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, with corresponding standard error 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 ≥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.

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 (Friedman, 2001). 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.

D. Data preprocessing details

Traumatic brain injury (TBI) To screen patients, we follow the inclusion and exclusion criteria from Kuppermann et al. (2009), 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. (2009).

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, 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. (2019) 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 TBI subsection above. 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. (2013) and (Kornblith et al., 2022). 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 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.

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 package (Singh et al., 2021)).

Traumatic brain injury		
Feature Name	% Missing	% Nonzero
Altered Mental Status	0.74	12.95
Altered Mental Status: Agitated	87.05	1.79
Altered Mental Status: Other	87.05	1.82
Altered Mental Status: Repetitive	87.05	1.04
Altered Mental Status: Sleepy	87.05	6.67
Altered 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 clavicles: Face	35.92	29.99
Trauma above clavicles: Scalp-frontal	35.92	20.48
Trauma above clavicles: Neck	35.92	1.38
Trauma above clavicles: Scalp-occipital	35.92	9.62
Trauma above clavicles: Scalp-parietal	35.92	7.79
Trauma above clavicles: 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 substantial injury: Abdomen	89.93	1.25
Other substantial injury: Cervical spine	89.93	1.37
Other substantial injury: Cut	89.93	0.12
Other substantial injury: Extremity	89.93	5.49
Other substantial injury: Flank	89.93	1.56
Other substantial injury: Other	89.93	1.65
Other substantial injury: Pelvis	89.93	0.44
Paralyzed	0.75	0.01
Basilar skull fracture	0.99	0.68
Basilar skull fracture: Hemotympanum	99.32	0.35
Basilar skull fracture: CSF otorrhea	99.32	0.04
Basilar skull fracture: Periorbital	99.32	0.19
Basilar skull fracture: Retroauricular	99.32	0.08
Basilar skull fracture: CSF rhinorrhea	99.32	0.03
Skull fracture: Palpable	0.24	0.38
Skull fracture: 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
Abdominal 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
Glasgow Coma Scale (GCS) score	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 neck 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 6. 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.

E. Extended results

	Traumatic brain injury						Cervical spine injury		
	92%	94%	96%	98%	ROC AUC	F1	92%	94%	96%
TAO	6.2 (5.9)	6.2 (5.9)	0.4 (0.4)	0.4 (0.4)	.294 (.05)	5.2 (.00)	41.5 (0.9)	21.2 (6.6)	0.2 (0.2)
TAO-SEP	26.7 (6.4)	13.9 (5.4)	10.4 (5.5)	2.4 (1.5)	.748 (.02)	**5.8** (.00)	32.5 (4.9)	7.0 (1.6)	5.4 (0.7)
CART	20.9 (8.8)	14.8 (7.6)	7.8 (5.8)	2.1 (0.6)	.702 (.06)	5.7 (.00)	38.6 (3.6)	13.7 (5.7)	1.5 (0.6)
CART-SEP	26.6 (6.4)	13.8 (5.4)	10.3 (5.5)	2.4 (1.5)	.753 (.02)	5.6 (.00)	32.1 (5.1)	7.8 (1.5)	5.4 (0.7)
G-CART	15.5 (5.5)	13.5 (5.7)	6.4 (2.2)	3.0 (1.5)	**.758** (.01)	5.5 (.00)	38.5 (3.4)	15.2 (4.8)	4.9 (1.0)
FIGS	23.8 (9.0)	18.2 (8.5)	12.1 (7.3)	0.4 (0.3)	.380 (.07)	4.8 (.00)	39.1 (3.0)	33.8 (2.4)	24.2 (3.2)
FIGS-SEP	39.9 (7.9)	19.7 (6.8)	**17.5** (7.0)	2.6 (1.6)	.619 (.05)	5.1 (.00)	38.7 (1.6)	33.1 (2.0)	20.1 (2.6)
G-FIGS	**42.0** (6.6)	**23.0** (7.8)	14.7 (6.5)	**6.4** (2.8)	.696 (.04)	4.7 (.00)	**42.2** (1.3)	**36.2** (2.3)	**28.4** (3.8)

	Cervical spine injury (cont.)			Intra-abdominal injury					
	98%	ROC AUC	F1	92%	94%	96%	98%	ROC AUC	F1
TAO	0.2 (0.2)	.422 (.04)	44.5 (.01)	0.2 (0.2)	0.2 (0.2)	0.0 (0.0)	0.0 (0.0)	.372 (.04)	**13.9** (.01)
TAO-SEP	2.5 (1.0)	.702 (.01)	44.4 (.01)	12.1 (1.7)	8.5 (2.0)	2.0 (1.3)	0.0 (0.0)	.675 (.01)	12.9 (.00)
CART	1.1 (0.4)	.617 (.06)	**45.8** (.01)	11.8 (5.0)	2.7 (1.0)	1.6 (0.5)	1.4 (0.5)	.688 (.06)	13.4 (.00)
CART-SEP	2.5 (1.0)	.707 (.00)	44.2 (.01)	11.0 (1.6)	9.3 (1.8)	2.8 (1.4)	0.0 (0.0)	.688 (.01)	13.0 (.01)
G-CART	3.9 (1.1)	**.751** (.01)	45.2 (.01)	11.7 (1.3)	10.1 (1.6)	3.8 (1.3)	0.7 (0.4)	**.732** (.02)	12.5 (.01)
FIGS	**16.7** (3.9)	.664 (.03)	43.0 (.01)	**32.1** (5.5)	13.7 (6.0)	1.4 (0.8)	0.0 (0.0)	.541 (.04)	9.4 (.01)
FIGS-SEP	3.9 (2.2)	.643 (.02)	41.4 (.01)	18.8 (4.4)	9.2 (2.2)	2.6 (1.7)	0.9 (0.8)	.653 (.02)	8.0 (.00)
G-FIGS	15.7 (3.9)	.700 (.01)	42.6 (.01)	29.7 (6.9)	**18.8** (6.6)	**11.7** (5.1)	**3.0** (1.3)	.671 (.03)	9.1 (.01)

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.

We include the results from above with their standard errors, as well as additional metrics (Area under the ROC Curve and F1 score) for each dataset.