EXPLAINABLE LONGITUDINAL PREDICTION WITH COMPLETE FUZZY MEASURES AND HIGHER-ORDER FEATURE INTERACTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding longitudinal predictions requires identifying how several variables' histories jointly shape future outcomes. Explanations limited to individual variables or pairs can miss these higher-order interactions. We introduce recursive Choquet, a longitudinal model that summarizes each feature's history separately and combines the summaries through a complete fuzzy measure. Our central construction recursively learns each feature set's value from those of its immediate subsets. We prove that this construction guarantees normalization and monotonicity without restricting interaction order. We then use the Choquet integral to produce forecasts with an exact decomposition into signed feature interactions. We compare our model with boosted trees, GRU-D, and a Transformer using common prediction-explanation methods. In controlled experiments, we recover positive and negative interactions; on a higher-order synthetic data, we obtain the lowest attribution error among these predictors at every order through six under both Faith-Shap and Shapley-Taylor. Across a harmonized set of clinical cohorts containing 166,836 participants, we predict 12 biomarkers and two depression scores with standardized test mean squared error 0.5215. For depression predictions, we obtain the most stable pair and triple explanations among these predictors, with triple cosine similarity 0.677 across training seeds. Our results support higher-order interaction recovery in controlled data and repeatable clinical explanations, providing a basis for further clinical evaluation.

1 INTRODUCTION

Explainable longitudinal prediction requires understanding both how variables are summarized over time and how they act together. Individual feature importance can miss joint effects, while pairwise explanations miss interactions involving three or more variables. For any feature set, we therefore ask what value the model assigns to it, what changes when a feature is added, and what interaction remains after lower-order effects are removed.

Fuzzy measures provide such an answer. A normalized monotone fuzzy measure, also called a capacity, assigns a value to each subset of inputs. The Choquet integral uses these values to aggregate the inputs, allowing their joint influence to depart from an additive combination (Choquet, 1954; Grabisch, 1996a). For a given observation, differences between sorted input values weight the capacity values along a nested sequence of feature sets. Thus the fuzzy measure is part of the prediction calculation itself. A complete measure retains interactions of every order, at the cost of exponentially many set values and monotonicity constraints.

We introduce a longitudinal Choquet model that preserves feature identity from past measurements to future predictions. Each variable is transformed and aggregated over its own history before entering a complete fuzzy measure. Our recursive parameterization guarantees monotonicity while retaining interactions of every order. The resulting prediction can be decomposed into terms for specific feature sets, making their contributions directly accessible.

Our contributions are: (1) a longitudinal prediction model that preserves feature identity and represents interactions of every order through a complete, monotone fuzzy measure; (2) an explanation analysis connecting the classical signed interaction decomposition to forecasts and establishing recovery

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

conditions under controlled inputs; and (3) a comparison of higher-order prediction explanations with boosted trees, GRU-D, and a Transformer, alongside clinical prediction and stability analyses. Synthetic experiments test recovery against known functions through six-feature interactions. Clinical experiments test whether explanations remain consistent across fitted models. Attention and lower-order capacity ablations assess the value and cost of the complete representation.

2 Related work

Capacities and Choquet integrals provide a classical way to model interacting criteria (Choquet, 1954; Grabisch, 1996a; 2003). The Möbius representation and fuzzy-measure interaction indices long predate this work: prior papers use fuzzy measures to represent feature importance and interaction, develop k-additive and interaction representations, and study transforms and axioms for higher-order interactions (Grabisch, 1996b; 1997; Denneberg & Grabisch, 1999; Grabisch & Roubens, 1999). We use the classical Möbius coefficient $d(S)$ to quantify signed interactions in the learned capacity.

Learning a complete capacity is difficult because it contains 2^n values and must respect monotonicity. Earlier work learns fuzzy measures from data, develops alternative coordinates that simplify constraints, embeds Choquet fusion in neural networks, and studies explanation and data-support problems for learned measures (Beliakov & Divakov, 2020; Islam et al., 2020; Murray et al., 2021; Kakula et al., 2021). Our recursive construction is used here as a practical parameterization that keeps direct set values and all orders.

Attention has been used for permutation-invariant set processing (Lee et al., 2019), while Deep Sets gives a general invariant construction (Zaheer et al., 2017). Our attention ablation parameterizes the capacity using representations organized by subset size. Explanations remain defined by the resulting capacity over the original features.

GRU-D incorporates missingness and elapsed time into a recurrent model (Che et al., 2018); Transformers use attention over sequences (Vaswani et al., 2017); and boosted trees remain strong tabular predictors (Chen & Guestrin, 2016). These methods are useful predictive comparisons, but they do not natively enumerate a normalized monotone value for every feature set. Interaction attribution methods such as Shapley-Taylor and Faith-Shap provide different allocations with their own axioms (Sundararajan et al., 2020; Tsai et al., 2023). We therefore define each reported set quantity explicitly rather than treating Möbius coefficients, fuzzy-measure interaction indices, and prediction-attribution indices as interchangeable.

3 Explainable prediction with a complete fuzzy measure

The model preserves variables through two stages: a separate temporal summary for each feature, followed by aggregation with a complete fuzzy measure. This separates how a feature's history is summarized from how feature sets enter the prediction.

3.1 Feature-specific longitudinal aggregation

Let $x_{ij\ell}$ be the ℓth available value of feature j for participant i, measured no later than landmark t_i. Training-participant statistics standardize each feature, and a scalar monotone map ϕ_j sends it to $[0, 1]$. For weights $w_{ij\ell} \geq 0$ that sum to one over available past measurements,

$$z_{ij} = \sum_{\ell : t_{ij\ell} \leq t_i} w_{ij\ell} \phi_j(x_{ij\ell}). \tag{1}$$

The main model uses softmax weights. Each weight score uses only feature j's value, elapsed time, and $\log(1 + \text{elapsed time})$. Thus changing feature j cannot change the time summary of feature $k \neq j$. Missing and future measurements receive zero weight. Each interaction is therefore among scalar history summaries, rather than complete trajectories; the capacity coefficients do not vary with forecast time.

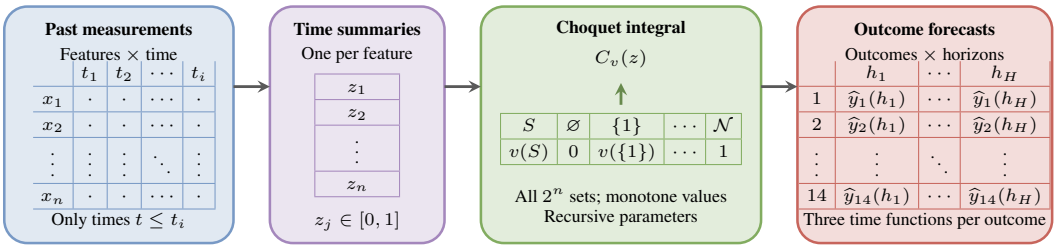


Figure 1: Recursive Choquet. Each feature retains its identity through temporal aggregation. We use a complete, recursively parameterized capacity to combine the summaries, and outcome-specific time functions produce longitudinal forecasts.

3.2 How the fuzzy measure enters prediction

Let $\mathcal{N} = \{1, \dots, n\}$ and let $v : 2^{\mathcal{N}} \to [0,1]$ satisfy $v(\varnothing) = 0$, $v(\mathcal{N}) = 1$, and $v(A) \le v(B)$ whenever $A \subseteq B$. For $z \in [0,1]^n$, sort its entries as $z_{[1]} \ge \cdots \ge z_{[n]}$ and let S_k contain the indices of the k largest entries. The discrete Choquet integral is

$$C_v(z) = \sum_{k=1}^{n} (z_{[k]} - z_{[k+1]}) v(S_k), \qquad z_{[n+1]} = 0. \tag{2}$$

For a fixed observation, each term connects a set S_k to the prediction through its capacity value and the input gap $z_{[k]} - z_{[k+1]}$. A large capacity value alone therefore does not imply a large contribution to that observation's aggregate. The calculation follows one sorted chain, while the complete fuzzy measure defines the model's behavior across all 2^n sets.

3.3 Learning a valid complete fuzzy measure

For every proper nonempty S, define

$$L(S) = \max_{j \in S} v(S \setminus \{j\}), \qquad v(S) = L(S) + [1 - L(S)] \operatorname{sigmoid}(\theta_S), \tag{3}$$

with sets evaluated in increasing size and fixed boundary values. Every immediate predecessor is included.

Proposition 1 (Validity and representation)**.** *Equation* (3) *produces a normalized monotone capacity for any finite parameters. Conversely, every normalized capacity whose proper nonempty sets have values strictly above those of their immediate subsets and strictly below one has a finite recursive parameterization.*

The proof is in Appendix A.1. Equality between a set and an immediate subset is obtained as a parameter limit. We call this model recursive Choquet. Its validity does not depend on a penalty coefficient. An attention parameterization and restrictions to lower interaction orders are evaluated as ablations (Appendix D).

3.4 Direct longitudinal prediction

For outcome o and time function c, let

$$f_{oc} = b_{oc} + a_{oc}\{2C_{v_{oc}}(z) - 1\}, \qquad r_{oc}(h) = 1 - \exp(-\kappa_{oc} h), \tag{4}$$

where h is the forecast horizon and $\kappa_{oc} > 0$. The selected relaxation form is

$$\widehat{y}_o(h) = y_o(0) \left[1 - \sum_c \pi_{oc} r_{oc}(h) \right] + \sum_c \pi_{oc} r_{oc}(h) f_{oc}, \tag{5}$$

with $\pi_{oc} \ge 0$ and $\sum_c \pi_{oc} = 1$. Here $y_o(0)$ is the available outcome anchor; when unavailable, it is estimated by a learned affine function of the first Choquet aggregate. We use three time

functions per outcome. Biomarkers use a multivariate Gaussian likelihood after log transformation and standardization using training participants. The 0-18 depression score uses a bounded censored-continuous logistic distribution; the 0-60 score uses a bounded discrete logistic distribution. No multilayer perceptron follows Equation (2).

The impact of a fuzzy-measure term on the predicted trajectory depends on the outcome coefficient a_{oc} and horizon weight $\pi_{oc} r_{oc}(h)$. Set values and gains are reported on the normalized capacity scale; they are not changes in clinical outcome units.

4 Explaining higher-order feature interactions

4.1 Set values, addition gains, and signed interactions

The fuzzy-measure value $v(S)$ describes a set's role in aggregation, including effects inherited from smaller subsets. It cannot be compared across sizes as a pure interaction strength, and $v(\mathcal{N}) = 1$ by definition. To explain its structure, we use three related quantities. For $j \notin A$, the directed addition is

$$\Delta_j v(A) = v(A \cup \{j\}) - v(A). \tag{6}$$

This measures the change in capacity when feature j joins the particular predecessor set A. For a nonempty target set S, its mean immediate gain is

$$G(S) = \frac{1}{|S|} \sum_{j \in S} \left[v(S) - v(S \setminus \{j\}) \right]. \tag{7}$$

It averages the gain of completing S over its immediate predecessors, so it describes the set's incremental value without isolating its pure interaction. Directed marginal contributions are standard objects for set functions; $G(S)$ is the simple average used here for set-level reporting. Finally, the classical Möbius coefficient, written as the signed finite difference

$$d(S) = \sum_{T \subseteq S} (-1)^{|S|-|T|} v(T) \tag{8}$$

isolates what remains after all proper-subset effects are removed. For $|S| \geq 2$, a positive or negative value indicates a positive or negative residual interaction in the fuzzy measure. Monotonicity requires nonnegative addition gains but permits negative signed interactions, since lower-order terms can offset them. Equation (8) is used to explain the learned direct set values.

These interactions connect exactly to the Choquet aggregate through $C_v(z) = \sum_{\varnothing \neq S \subseteq \mathcal{N}} d(S) \min_{j \in S} z_j$ (Appendix B, Equation (18)). Each signed interaction is weighted by the smallest transformed input in its set. This is an algebraic decomposition of the aggregate; it does not describe the effect of intervening on a clinical variable.

For example, the generating pairs $\{1, 3\}$ and $\{5, 6\}$ both have $v(S) = 0.2833$, but their signed interactions are $+0.20$ and -0.20. Their mean gains are 0.2417 and 0.0417. Larger singleton values offset the negative interaction. Appendix A.2 gives the full calculation.

Theorem 1 (Exact set-gain identities)**.** *For every nonempty S,*

$$\Delta_j v(A) = \sum_{\substack{B \subseteq A \cup \{j\} \\ j \in B}} d(B), \qquad G(S) = \frac{1}{|S|} \sum_{B \subseteq S} |B| d(B). \tag{9}$$

Let $A_r(S)$ be the mean of $v(T)$ over the r-subsets of S, with $A_0(S) = 0$, and let $P_r(S)$ be the mean of $G(T)$ over the same r-subsets. Then

$$P_r(S) = A_r(S) - A_{r-1}(S), \qquad v(S) = \sum_{r=1}^{|S|} P_r(S). \tag{10}$$

If v is monotone, every $P_r(S)$ is nonnegative.

These identities show that gains can contain interactions of several orders. The nonnegative increments $P_r(S)$ sum to $v(S)$; they do not isolate order-r interactions. Appendix A.3 gives the counting proof.

4.2 IDENTIFICATION AND ACCURACY OF THE EXPLANATIONS

Theorem 2 (Complete-probe identification). *For $0 < t \leq 1$ and $z = t\mathbf{1}_S$, $C_v(z) = tv(S)$. One probe for every proper nonempty S therefore identifies the complete capacity after fixing $v(\varnothing) = 0$ and $v(\mathcal{N}) = 1$. If $\|\widehat{v} - v\|_\infty \leq \delta$, then*

$$|\widehat{G}(S) - G(S)| \leq 2\delta, \qquad |\widehat{d}(S) - d(S)| \leq 2^{|S|}\delta. \tag{11}$$

At a fixed order, every planted gain ranks above every null gain whenever their true separation exceeds 4δ.

The identity provides a controlled identification test. It does not guarantee recovery from correlated observational data or convergence of the optimizer. The exponential error bound makes higher-order interactions particularly sensitive to capacity error. Appendix A.4 gives the proof and a design-matrix extension.

5 EXPERIMENTS

We evaluate two questions separately: whether the explanations reflect known feature interactions in controlled data, and how the complete fuzzy-measure model performs in a clinical prediction task. The clinical analysis describes the fitted measure where the underlying interactions are unknown.

The experiments were run in an workstation with Intel Xeon Silver 4216 CPU (2.10 GHz), 512GB RAM and 3 Nvidia RTX A4000 48GB GPUs. The models did not need to go beyond a single GPU but we ran them concurrently optimize time requirements. Appendix provides a CPU runtime information for fare estimations under certain comparisons.

5.1 CONTROLLED VALIDATION OF FEATURE-SET EXPLANATIONS

The controlled study contains six correlated histories. Feature-specific power means use exponents from 1.3 to 2.3. The generating finite differences are

$$d(\{1,3\}) = d(\{2,4\}) = 0.20, \quad d(\{1,2,3\}) = 0.15, \quad d(\{5,6\}) = -0.20, \tag{12}$$

with singleton values $1/24$ for features 1-4 and $29/120$ for features 5-6; all remaining higher-order terms, including the full set, are zero. The large fifth and sixth singleton values keep the capacity monotone despite the negative pair. Here, we test signed cancellation rather than simply rewarding every planted set with a large positive value.

We use moderate correlation/noise ($\rho = 0.55, \sigma = 0.01$) and strong correlation/noise ($\rho = 0.85, \sigma = 0.03$). Each fit uses 20,000 training, 5,000 validation, and 20,000 independent test observations. Training and validation combine ordinary histories with 32 and 8 repeats, respectively, of one scaled Boolean probe per set. Model selection uses noisy validation MSE and never uses the generating coefficients. We fit the recursive model with three seeds per condition and evaluate all 64 sets and 192 directed additions. We also match the proposed recursive model against an attention based progressive ablation(Appendix B.3).

5.2 COMPARING EXPLANATIONS OF FINAL PREDICTIONS

We perform a separate synthetic longitudinal evaluation of final-prediction explanations using known pair-only and higher-order data-generating processes. Each synthetic dataset generated with six-features across 100,000 training, 20,000 validation, and 20,000 test samples. We explain both outcomes at seven years for 128 test histories per dataset, across three seeds. For each predictor, absent features are replaced by their training means throughout the history while outcome anchors and observation times remain fixed. Evaluating all 64 feature subsets gives exact Faith-Shap and Shapley-Taylor explanations through order six. Reference explanations use the same replacements in the generating function. These errors are measured in standardized outcome units; the preceding recovery study measures internal capacity coefficients. The design and results are provided in Appendix D.6.

5.3 Clinical prediction and fitted fuzzy measures

We use a harmonized longitudinal dataset combining the Coronary Artery Risk Development in Young Adults (CARDIA) study and the Women's Health Initiative (WHI), two longitudinal U.S. cohorts (Friedman et al., 1988; The Women's Health Initiative Study Group, 1998; Gurugubelli et al., 2022). To address missingness in the longitudinal data, the source data-processing framework builds on multiple-imputation-based fuzzy clustering and validation methods developed for incomplete longitudinal data (Fang, 2017; Zhang & Fang, 2016; Zhang et al., 2016), with subsequent work extending these methods to harmonized longitudinal health datasets (Ngo et al., 2022; 2024). Because the imputation procedure is not a contribution of this work, we treat the imputation methods as part of the harmonized data and focus instead on evaluating whether the proposed model remains robust across imputed datasets. The resulting dataset contains 166,836 unique participants, represented in each of ten multiply imputed datasets. We use a fixed participant-level split of 116,613 participants for training, 16,699 for validation, 16,780 for calibration, and 16,744 for testing.

The 21 Choquet inputs are ten directly measured biomarkers and the eleven AHEI-2010 components (Chiuve et al., 2012). BMI and HOMA-IR are retained as targets and histories but excluded from the Choquet input set because they are derived from other inputs. CRP is excluded after its recorded visit timing leaves no supported primary measurement. Outcomes are 12 biomarkers plus the original 0-18 WHI and 0-60 CARDIA depression scores, which are modeled separately.

Data preparation converts CARDIA weight from pounds to kilograms, uses cohort-specific nominal examination times, restricts all history to the landmark, and projects AHEI-2010 component values to their defined $[0, 10]$ range. Participant identity never crosses a split. The original imputation program and observed-value indicators were unavailable, so we cannot determine whether imputation used information across participant splits or evaluate error exclusively on observed outcomes. GRU-D therefore uses availability after preprocessing rather than original missingness indicators.

Time aggregation, activation, trajectory form, capacity construction, learning rate, and comparison-model settings were selected using training and validation participants. The selected recursive model uses softmax time aggregation, scalar sigmoid transforms, the equilibrium trajectory in Equation (5), AdamW, batch size 4096, at most 300 epochs, and patience 40. The external comparisons are boosted trees, GRU-D, and a matched causal Transformer; progressive Choquet is an attention ablation. Recursive Choquet, boosted trees, GRU-D, and the Transformer have 30 fits across ten imputations and three random seeds; the larger progressive model has three predeclared seeds on imputation 1. After selection was fixed, calibration participants set split-conformal quantiles from each participant's maximum standardized trajectory error (Lei et al., 2018). Subsequent analyses and lower-order model comparisons are secondary because they were designed after the original test results were known.

6 Results

6.1 Explanations distinguish positive and negative interactions

Figure 2 shows how the explanatory quantities differ under strong correlation and noise. The Recursive Choquet model recovers every planted interaction sign. The positive pairs have the two highest mean pair gains, and the planted triple has the highest triple gain. For the negative pair, a small positive immediate gain coexists with a negative signed interaction: its large singleton values offset the negative pair term. The full set provides a second contrast, with capacity fixed at one and a generating signed interaction of zero. Thus capacity values, gains, and signed interactions answer different questions about the same fuzzy measure.

Across all 64 sets, recursive Choquet has interaction MAE 0.00680 in the moderate condition and 0.01401 in the strong condition. Strong-condition noise-free prediction RMSE is 0.00134. These results assess two distinct properties: prediction error measures agreement with the generating response, while interaction errors and individual-set comparisons assess the explanations. Appendix B.3 extends the individual-set analysis to every order. Signed averages within an order can obscure individual interactions through cancellation.

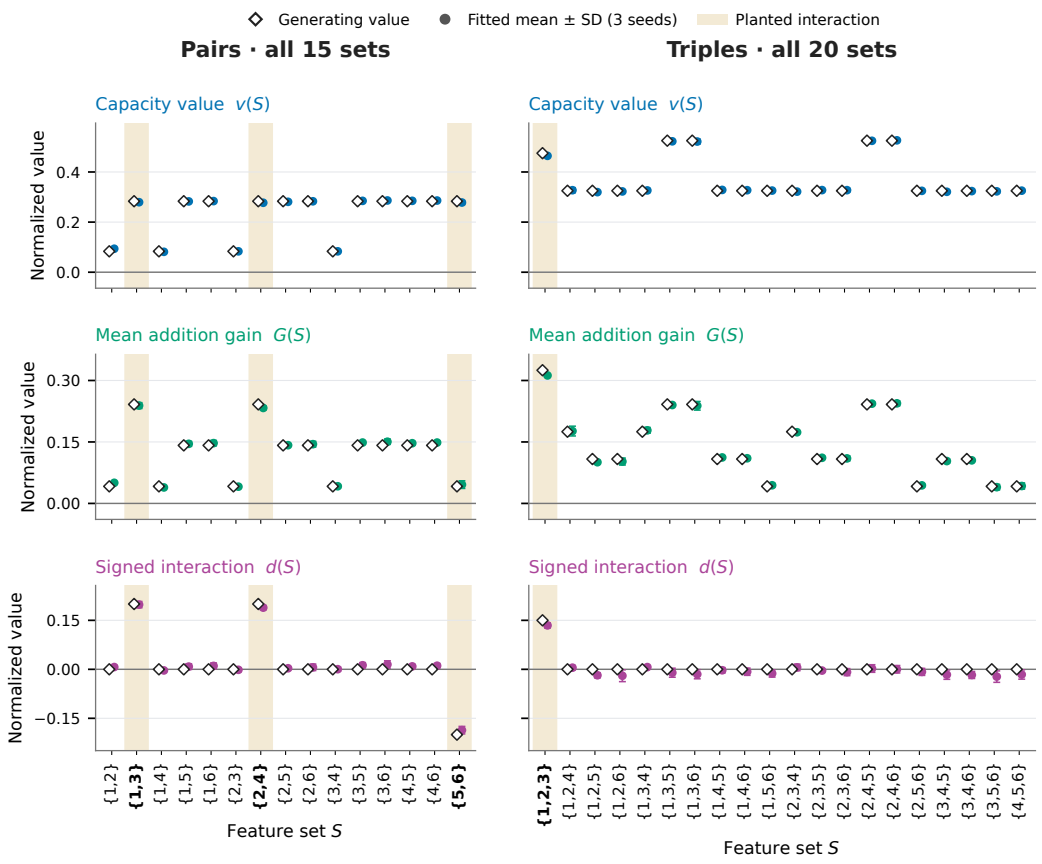


Figure 2: Feature-set explanations for all 15 pairs and 20 triples in the recursive model under strong correlation and noise. Rows distinguish capacity values $v(S)$, mean addition gains $G(S)$, and signed interactions $d(S)$, with shared scales across columns. Colored points show fitted means and error bars show sample SD over three seeds; open diamonds show generating values. Shading and bold set labels mark planted nonzero interactions. A positive gain can coexist with a negative interaction, as for $\{5, 6\}$. Other set orders and the progressive model are shown in Appendix B.3.

6.2 Higher-order prediction explanations

On the higher-order interaction synthetic data, recursive Choquet has the lowest mean attribution error at every order from one through six under both explanation methods. Table 1 summarizes triple and sixth-order Faith-Shap errors. Sign agreement for nonzero reference terms is 98.7% for triples and 91.7% for sixth-order terms. Reference signals become smaller and less frequent with order: sixth-order sign agreement uses 56 coefficients above 10^{-5} out of 768. Relative MAE is 0.007 for triples and 0.090 for sixth-order terms. On the pair-only synthetic data, recursive Choquet also has the lowest errors at orders two through six, indicating fewer spurious higher-order explanations.

The model also fits the higher-order generating response more accurately: prediction RMSE is 0.00057, compared with 0.021-0.137 for the external predictors. Thus the explanation results assess recovery of the generating function and its interactions together. They establish controlled recovery within the Choquet family. Appendix D.6 reports prediction errors, both explanation methods at every order, and the progressive ablation.

6.3 Clinical prediction

Table 2 reports test MSE averaged equally across standardized outcomes. Boosted trees have the lowest point estimate at 0.5043, followed by GRU-D at 0.5076. Recursive Choquet obtains 0.5215.

Table 1: Higher-order prediction explanations. Synthetic columns give Faith-Shap MAE ($\times 10^3$, mean $\pm$ SD over three seeds) on the higher-order generator; lower is better. Clinical columns give mean cosine similarity across training seeds for 64 participants; higher is more stable. Tree triple comparisons exclude jointly zero explanations. Full results and the progressive ablation are in Appendix D.6.

Model	Synthetic triples	Synthetic order six	Clinical pairs	Clinical triples
Recursive	0.04 ± 0.0095	0.0142 ± 0.011	0.778	0.677
Boosted trees	11.7 ± 1	0.439 ± 0.25	0.591	0.450
GRU-D	7.28 ± 0.36	0.937 ± 0.29	0.646	0.248
Transformer	5.63 ± 0.3	1.84 ± 0.29	0.449	0.104

Table 2: Clinical test prediction. SD is across imputations and random seeds. Intervals resample test participants and are not pairwise difference tests.

Model	Fits	Macro standardized MSE	Participant 95% interval
Boosted trees	30	0.50429 ± 0.00049	$[0.47355, 0.53504]$
GRU-D	30	0.50761 ± 0.00038	$[0.47692, 0.53900]$
Matched Transformer	30	0.51559 ± 0.00286	$[0.48565, 0.54575]$
Recursive Choquet	30	0.52149 ± 0.00021	$[0.49131, 0.55268]$

A secondary paired participant bootstrap found lower error for trees and GRU-D after adjustment for four comparisons; the Transformer interval included zero (Appendix E.4). These intervals condition on the fitted models.

Calibration targets simultaneous coverage of all available future measurements for a participant and outcome. Mean 90% trajectory coverage ranges from 0.890 to 0.895; mean 95% coverage ranges from 0.948 to 0.950. Beyond these aggregate measures, Figure 9 shows how error increases with forecast horizon, and Table 8 reports sensitivity to history deletion, measurement noise, and input shift. We show both analyses in Appendix C.

6.4 Clinical interaction stability

For final prediction explanations, all models use the same sample of 64 clinical participants, two cohort-specific depression outcomes at five years, one imputation, and three training seeds. Faith-Shap interactions are estimated through order three with two independent samples of 16,384 feature subsets. Recursive Choquet has the highest pair and triple cosine similarity among the five evaluated predictors (Table 1); its top-ten selection overlaps are 0.492 and 0.413. Trees have the most stable individual-feature explanations. Their triple result uses 21 nonempty participant/seed-pair comparisons, versus 192 for the other predictors. Full results, including the progressive ablation, numerical uncertainty, and prediction sensitivity, further expanded in Appendix D.6.

This relative advantage does not establish clinical correctness. Triple agreement between different predictors is weak, with cosine similarities from -0.036 to 0.066. Separately, exact internal capacity coefficients across all 30 recursive fits become less stable at higher orders: median selection frequency declines from 0.690 for pairs to 0.310 for triples and 0.103 for five-feature interactions. These coefficients describe the fitted measure, whereas Faith-Shap describes each participant's final prediction. Appendix E.3 provides the complete refitting analysis, and Appendix D.7 shows named clinical examples.

6.5 Ablation Study

We also examine two simpler variants of the main recursive model. First, we add attention to the recursive capacity parameterization, giving the Progressive Choquet model. On clinical imputation 1, it reaches an MSE of 0.51423, although its pair and triple prediction-explanation stability is lower than that of the main recursive model. We also restrict the capacity to individual and pair terms. This 2-additive model improves clinical prediction to an MSE of 0.51492, compared with 0.52148 for matched full recursive fits, while reducing mean training time from 3,676 to 517 seconds. We provide

the full ablation results in Appendices D and E.4. The lower-order variants were not included in the common Faith-Shap comparison as the focus is higher-order in the long term.

We further test how much the learned interactions depend on direct subset supervision. Under strong correlation and noise, the recursive model attains an interaction MAE of 0.0217 without probes and 0.0559 with randomly sampled partial probes, while prediction RMSE remains low at 0.0014 and 0.0031, respectively. These results show that good predictive performance does not by itself imply accurate recovery of the underlying interactions. We report the full set of probe designs, alternative generators, and capacity restrictions in Appendix E.1.

7 Discussion and conclusion

Our Recursive Choquet connects longitudinal predictions to explicit interactions among feature histories. Its complete fuzzy measure represents every interaction order while preserving feature identity and monotonicity. On the higher-order longitudinal generator, it recovers prediction explanations more accurately than the evaluated predictors at every order through six. Clinical results show the strongest pair and triple explanation stability among those predictors, alongside modestly higher prediction error than trees and GRU-D.

Retaining every order has a cost: the clinical model has 88.1 million parameters, and lower-order ablations improve prediction and reduce computation. This is expected due to the $O(2^N)$ scaling factor with N features. The benefit of the complete representation is access to interactions that a restricted capacity cannot express; their usefulness must be assessed for the task. Sparse capacities and stability-guided restrictions are promising directions.

Our synthetic experiments support interaction recovery against known synthetic interactions. However, clinical interpretation is limited by supplied imputations with unknown original missingness, correlated histories, and secondary analyses designed after the original test results were known. Full-scale clinical explanation requires extension across participants and outcomes, external validation, and human evaluation by clinicians and domain experts. Future work should establish which higher-order interactions are clinically meaningful and sufficiently stable at clinical level to inform scientific understanding.

AI use statement

Generative AI tools were used for drafting and editing this manuscript. The authors reviewed the generated material, checked numerical presentation, performed the implementations and take responsibility for the final paper.

Ethics statement

This study uses previously harmonized longitudinal cohort data. The paper reports aggregate results and does not release participant-level records. Clinical feature-set values are fitted associations and are not intended for diagnosis, treatment selection, or causal interpretation until expert clinical evaluation is performed.

Reproducibility statement

We have included appendices with proofs, data definitions, model settings, stopping rules, and secondary results. Supplementary Code is added with the paper as well. Participant records are not distributed; reproducing the clinical fits requires authorized access to the cohort data.

REFERENCES

Gleb Beliakov and Dmitriy Divakov. On representation of fuzzy measures for learning choquet and sugeno integrals. *Knowledge-Based Systems*, 189:105134, 2020. doi: 10.1016/j.knosys.2019.105134.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8:6085, 2018. doi: 10.1038/s41598-018-24271-9.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. doi: 10.1145/2939672.2939785.

Stephanie E. Chiuve, Teresa T. Fung, Eric B. Rimm, Frank B. Hu, Marjorie L. McCullough, Molin Wang, Meir J. Stampfer, and Walter C. Willett. Alternative dietary indices both strongly predict risk of chronic disease. *The Journal of Nutrition*, 142(6):1009–1018, 2012. doi: 10.3945/jn.111.157222.

Gustave Choquet. Theory of capacities. *Annales de l'Institut Fourier*, 5:131–295, 1954. doi: 10.5802/aif.53.

Dieter Denneberg and Michel Grabisch. Interaction transform of set functions over a finite set. *Information Sciences*, 121(1–2):149–170, 1999. doi: 10.1016/S0020-0255(99)00099-7.

Hua Fang. MIFuzzy clustering for incomplete longitudinal data in smart health. *Smart Health*, 1-2: 50–65, 2017. doi: 10.1016/j.smhl.2017.04.002.

Gary D. Friedman, Gary R. Cutter, Richard P. Donahue, George H. Hughes, Stephen B. Hulley, David R. Jacobs Jr., Kiang Liu, and Peter J. Savage. CARDIA: Study design, recruitment, and some characteristics of the examined subjects. *Journal of Clinical Epidemiology*, 41(11):1105–1116, 1988. doi: 10.1016/0895-4356(88)90080-7.

Michel Grabisch. The application of fuzzy integrals in multicriteria decision making. *European Journal of Operational Research*, 89(3):445–456, 1996a. doi: 10.1016/0377-2217(95)00176-X.

Michel Grabisch. The representation of importance and interaction of features by fuzzy measures. *Pattern Recognition Letters*, 17(6):567–575, 1996b. doi: 10.1016/0167-8655(96)00020-7.

Michel Grabisch. k-order additive discrete fuzzy measures and their representation. *Fuzzy Sets and Systems*, 92(2):167–189, 1997. doi: 10.1016/S0165-0114(97)00168-1.

Michel Grabisch. Modelling data by the choquet integral. In Vicenç Torra (ed.), *Information Fusion in Data Mining*, pp. 135–148. Springer, 2003. doi: 10.1007/978-3-540-36519-8_8.

Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 28(4):547–565, 1999. doi: 10.1007/s001820050125.

Venkata Sukumar Gurugubelli, Hua Fang, James M. Shikany, Salvador V. Balkus, Joshua Rumbut, Hieu Ngo, Honggang Wang, Jeroan J. Allison, and Lyn M. Steffen. A review of harmonization methods for studying dietary patterns. *Smart Health*, 23:100263, 2022. doi: 10.1016/j.smhl.2021.100263.

Muhammad Aminul Islam, Derek T. Anderson, Anthony J. Pinar, Timothy C. Havens, Grant J. Scott, and James M. Keller. Enabling explainable fusion in deep learning with fuzzy integral neural networks. *IEEE Transactions on Fuzzy Systems*, 28(7):1291–1300, 2020. doi: 10.1109/TFUZZ.2019.2917124.

Siva Krishna Kakula, Anthony J. Pinar, Muhammad Aminul Islam, Derek T. Anderson, and Timothy C. Havens. Novel regularization for learning the fuzzy choquet integral with limited training data. *IEEE Transactions on Fuzzy Systems*, 29(10):2890–2901, 2021. doi: 10.1109/TFUZZ.2020.3009722.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6696–6707, 2020.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 3744–3753. PMLR, 2019.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018. doi: 10.1080/01621459.2017.1307116.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Bryce J. Murray, Muhammad Aminul Islam, Anthony J. Pinar, Derek T. Anderson, Grant J. Scott, Timothy C. Havens, and James M. Keller. Explainable AI for the choquet integral. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(4):520–529, 2021. doi: 10.1109/TETCI.2020.3005682.

Hieu Ngo, Hua Fang, and Honggang Wang. Poster: Intelligent fuzzifier-based cluster validation for incomplete longitudinal digital trial data. In *2022 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pp. 146–147, 2022.

Hieu Ngo, Hua Fang, Joshua Rumbut, and Honggang Wang. Federated fuzzy clustering for decentralized incomplete longitudinal behavioral data. *IEEE Internet of Things Journal*, 11(8):14657–14670, 2024. doi: 10.1109/JIOT.2023.3343719.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The shapley taylor interaction index. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 9259–9268. PMLR, 2020.

The Women's Health Initiative Study Group. Design of the women's health initiative clinical trial and observational study. *Controlled Clinical Trials*, 19(1):61–109, 1998. doi: 10.1016/S0197-2456(97)00078-0.

Che-Ping Tsai, Chih-Kuan Yeh, and Pradeep Ravikumar. Faith-shap: The faithful shapley interaction index. *Journal of Machine Learning Research*, 24(94):1–42, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Zhaoyang Zhang and Hua Fang. Multiple- vs non- or single-imputation based fuzzy clustering for incomplete longitudinal behavioral intervention data. In *2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pp. 219–228. IEEE, 2016. doi: 10.1109/CHASE.2016.19.

Zhaoyang Zhang, Hua Fang, and Honggang Wang. Multiple imputation based clustering validation (MIV) for big longitudinal trial data with missing values in eHealth. *Journal of Medical Systems*, 40(6):146, 2016. doi: 10.1007/s10916-016-0499-0.

A MATHEMATICAL ANALYSIS

A.1 RECURSIVE CAPACITY VALIDITY AND EXPRESSIVENESS

Proof of Proposition 1. The boundary values are fixed as $v(\varnothing) = 0$ and $v(\mathcal{N}) = 1$. Consider a proper nonempty set S. Since $0 < \mathrm{sigmoid}(\theta_S) < 1$ for finite θ_S,

$$L(S) < v(S) < 1.$$

For every $j \in S$, $v(S \setminus \{j\}) \leq L(S) < v(S)$. Thus every addition of a feature to an immediate subset of S is monotone. If $A \subset B$, there is a chain $A = S_0 \subset S_1 \subset \cdots \subset S_m = B$ in which one feature is added at each step. Applying the immediate-subset inequalities along this chain gives $v(A) \leq v(B)$. The full-set boundary is at least every proper-set value, so normalization and monotonicity hold everywhere.

For the reverse direction, let v^* be a normalized capacity with strict increases from immediate subsets and $v^*(S) < 1$ for every proper nonempty set. Define

$$L^*(S) = \max_{j \in S} v^*(S \setminus \{j\}), \qquad q_S = \frac{v^*(S) - L^*(S)}{1 - L^*(S)}.$$

The assumptions give $q_S \in (0, 1)$. Setting $\theta_S = \log(q_S/(1 - q_S))$ and evaluating sets in increasing size reproduces $v^*(S)$ by induction. Equality with an immediate subset or a proper-set value of one corresponds to $q_S \in \{0, 1\}$ and is approached as $\theta_S \to -\infty$ or $+\infty$, respectively. □

The proof uses no loss penalty. In contrast, a finite constraint penalty can trade a small monotonicity violation against improved prediction loss. This is why the recursive construction is used for the primary model and loss-based constraints are treated as an ablation.

A.2 WORKED EXAMPLES OF SET VALUES AND INTERACTIONS

The generating measure in our controlled study makes the distinction concrete. For the positive pair $\{1, 3\}$, each singleton contributes $1/24$ and the pair interaction contributes 0.20. Hence $v(\{1, 3\}) = 2/24 + 0.20 \approx 0.2833$, while either addition to complete the pair has gain $1/24 + 0.20 \approx 0.2417$. For the negative pair $\{5, 6\}$, the larger singleton values $29/120$ offset the interaction -0.20: its capacity is also 0.2833, but its mean addition gain is only 0.0417. Identical set values can therefore conceal opposite interactions.

For the triple $S = \{1, 2, 3\}$, the measure contains three singleton terms, the pair interaction $d(\{1, 3\}) = 0.20$, and the additional triple interaction $d(S) = 0.15$. Thus $v(S) = 3/24 + 0.20 + 0.15 = 0.475$. Completing the triple from $\{1, 3\}$ adds $1/24 + 0.15 \approx 0.1917$, whereas completing it from either $\{1, 2\}$ or $\{2, 3\}$ also introduces the pair interaction and adds approximately 0.3917. Their mean is $G(S) = 0.325$, which includes both inherited pair structure and the triple interaction.

The contribution to a prediction also depends on the inputs. For illustrative transformed values $(z_1, z_2, z_3) = (0.8, 0.4, 0.6)$, the pair term contributes $0.20 \min(z_1, z_3) = 0.12$ to the Choquet aggregate, and the triple term contributes $0.15 \min(z_1, z_2, z_3) = 0.06$. These are separate terms in the exact signed decomposition, not total predictions or intervention effects. Equation (5) then scales their contribution to each outcome and forecast horizon.

A.3 SET-SPECIFIC GAIN IDENTITIES

Boolean-lattice inversion of Equation (8) gives

$$v(A) = \sum_{B \subseteq A} d(B). \tag{13}$$

Subtracting the expansions for $A \cup \{j\}$ and A cancels every term not containing j, hence

$$\Delta_j v(A) = \sum_{\substack{B \subseteq A \cup \{j\} \\ j \in B}} d(B).$$

For a fixed nonempty $B \subseteq S$, the term $d(B)$ appears in $v(S) - v(S \setminus \{j\})$ exactly when $j \in B$. It therefore appears $|B|$ times in the sum over $j \in S$, proving

$$G(S) = \frac{1}{|S|} \sum_{B \subseteq S} |B| d(B).$$

This also yields the signed order decomposition

$$Q_r(S) = \frac{r}{|S|} \sum_{\substack{B \subseteq S \\ |B|=r}} d(B), \qquad G(S) = \sum_{r=1}^{|S|} Q_r(S). \tag{14}$$

The $Q_r(S)$ values can be negative. They separate the singleton, pair, triple, and later finite-difference terms inherited by one target set.

We next prove the direct nonnegative decomposition. Expand the definition of $P_r(S)$:

$$\begin{aligned} P_r(S) &= \binom{|S|}{r}^{-1} \sum_{\substack{T \subseteq S \\ |T|=r}} \left[v(T) - \frac{1}{r} \sum_{j \in T} v(T \setminus \{j\}) \right] \\ &= A_r(S) - \frac{1}{r\binom{|S|}{r}} \sum_{\substack{T \subseteq S \\ |T|=r}} \sum_{j \in T} v(T \setminus \{j\}). \end{aligned}$$

Each $(r-1)$-subset U of S is obtained as $T \setminus \{j\}$ in exactly $|S| - r + 1$ ways. The combinatorial identity

$$r\binom{|S|}{r} = (|S| - r + 1)\binom{|S|}{r-1}$$

therefore turns the second term into $A_{r-1}(S)$. Thus $P_r(S) = A_r(S) - A_{r-1}(S)$. Summing over r telescopes to $A_{|S|}(S) - A_0(S) = v(S)$. Monotonicity makes each immediate difference inside every $G(T)$ nonnegative, so $P_r(S) \geq 0$.

For the full feature set, the order term can also be divided among features. Define

$$P_{r,j}(\mathcal{N}) = \frac{1}{r\binom{n}{r}} \sum_{\substack{T \subseteq \mathcal{N},\ |T|=r \\ j \in T}} \left[v(T) - v(T \setminus \{j\}) \right]. \tag{15}$$

Summing over j visits every directed addition into an order-r set once, so

$$\sum_{j=1}^{n} P_{r,j}(\mathcal{N}) = P_r(\mathcal{N}), \qquad \sum_{r=1}^{n} \sum_{j=1}^{n} P_{r,j}(\mathcal{N}) = v(\mathcal{N}) = 1. \tag{16}$$

This assigns every successive order increase to named features without creating a hidden representation.

A.4 Probe identification and finite-error ranking

Proof of Theorem 2. For $S = \varnothing$, both sides are zero. For nonempty S, let $z = t\mathbf{1}_S$ with $0 < t \leq 1$. After sorting, the first $|S|$ entries are t and the rest are zero. All adjacent differences in Equation (2) vanish except the difference at $k = |S|$, which equals t. Its sorted set is S. Therefore $C_v(t\mathbf{1}_S) = tv(S)$.

After fixing the empty and full boundaries, collect the $2^n - 2$ proper capacity values into a vector. One proper-set probe per coordinate gives a diagonal design matrix with positive diagonal entries t_S. It has full rank, so the noiseless solution is unique. For each set, repeated probes at a fixed positive scale with independent mean-zero finite-variance noise yield a consistent estimate after division by that scale.

For the error bounds,

$$|\widehat{\Delta}_j v(A) - \Delta_j v(A)| \leq |\widehat{v}(A \cup \{j\}) - v(A \cup \{j\})| + |\widehat{v}(A) - v(A)| \leq 2\delta.$$

Averaging cannot enlarge this uniform bound, so $|\widehat{G}(S) - G(S)| \le 2\delta$. Equation (8) contains $2^{|S|}$ terms with coefficients of absolute value one, giving $|\widehat{d}(S) - d(S)| \le 2^{|S|}\delta$ by the triangle inequality.

Let P be a planted set and U a null set of the same order. Then

$$\widehat{G}(P) - \widehat{G}(U) \ge G(P) - G(U) - 4\delta.$$

If the true separation is greater than 4δ, the fitted ordering is strict. Applying this to every planted-null pair proves the ranking statement. □

For ordinary observations, let X be the sorted-chain gap design matrix for the proper capacity values. If X has full column rank, unconstrained least squares has a unique solution. With response error ε,

$$\|\widehat{v} - v\|_2 \le \frac{\|\varepsilon\|_2}{\sigma_{\min}(X)}. \tag{17}$$

Strong feature correlation can make $\sigma_{\min}(X)$ small even when population support is full. The complete probes remove this ambiguity in the controlled experiment; they do not remove it from observational clinical data.

A.5 Progressive attention properties

The attention ablation is defined in Appendix D.1. At order one, there are n singleton representations. Suppose all representations through order $k-1$ have been constructed. For each order-k set S, the method forms exactly one item for every $j \in S$, namely $(h_{S\setminus\{j\}}, h_{\{j\}})$. Thus it visits every directed edge from an order-$(k-1)$ predecessor into S. Induction through $k = n$ visits all 2^n sets and

$$\sum_{k=1}^{n} k\binom{n}{k} = n2^{n-1}$$

directed additions, including the singleton edges from the empty set.

Cross-attention applies the same key and value maps to all elements of $\mathcal{D}_S$. Permuting their enumeration permutes scores and values together. Softmax normalization and their weighted sum are therefore unchanged. Induction gives enumeration-independent h_S. This is not invariance to changing the identity of a clinical variable: relabeling variables must also relabel their learned descriptors and residual coordinates.

The progressive score a_S enters the same recursive parameterization as b_S. Consequently, Proposition 1 applies for arbitrary finite attention parameters. Setting the final attention score to zero reduces the model to free residual coordinates, so the representation result for capacities with strict inequalities is retained. Large attention weights alone do not imply a large $G(S)$ or $d(S)$ because the residual coordinate and recursive normalization also affect the final capacity.

B Synthetic experiments

B.1 Generating capacity

For $\mathcal{N} = \{1, \dots, 6\}$, the nonzero finite differences are

$$\begin{aligned} d(\{j\}) &= 1/24 && j = 1,2,3,4, \\ d(\{j\}) &= 29/120 && j = 5,6, \\ d(\{1,3\}) &= d(\{2,4\}) = 1/5, \\ d(\{1,2,3\}) &= 3/20, \qquad d(\{5,6\}) = -1/5. \end{aligned}$$

All other $d(S)$ are zero. The coefficients sum to one. Every addition involving the negative pair retains a nonnegative total margin because adding feature 5 or 6 contributes its singleton value $29/120$ before the $-1/5$ pair term, leaving at least $1/24$. Hence the generating set function is normalized and monotone.

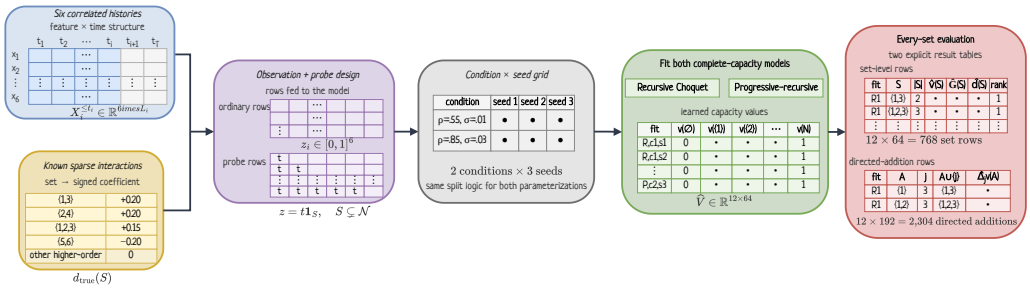


Figure 3: Controlled recovery design.

Table 3: All-condition synthetic summary, mean over three seeds. Noise-free RMSE is measured against the generating response; capacity and interaction errors use every one of the 64 sets.

Model	Condition	RMSE	Capacity MAE	Interaction MAE	Max interaction error
Progressive	Moderate	.000743	.002094	.005858	.036488
Progressive	Strong	.001341	.004662	.014242	.070034
Recursive	Moderate	.000583	.001764	.006798	.043686
Recursive	Strong	.001339	.004460	.014005	.081297

For inputs $z \in [0,1]^6$, Boolean inversion also gives the exact function

$$C_v(z) = \sum_{\varnothing \neq B \subseteq \mathcal{N}} d(B) \min_{j \in B} z_j. \tag{18}$$

To verify Equation (18), insert Equation (13) into the sorted-chain Choquet formula and collect the coefficient of each $d(B)$. A term $d(B)$ appears exactly until the smallest value among $\{z_j : j \in B\}$ leaves the sorted chain, so its telescoping coefficient is $\min_{j \in B} z_j$. This decomposition is exact and signed. It is used for controlled interpretation, while the forward model uses direct set values.

The six latent Gaussian trajectories have signed Toeplitz dependence before transformation. Each named history is transformed and aggregated with a feature-specific positive power between 1.3 and 2.3. Moderate and strong settings use the correlation and noise values stated in the main paper. Because nonlinear aggregation changes correlations, the empirical correlation of the aggregated features is reported separately and is not expected to reproduce the Gaussian matrix numerically.

B.2 Accuracy of fitted measures and signed interactions

Across the four model-condition groups, both positive pairs rank first and second by mean immediate gain among pairs, and the positive triple ranks first among triples. All four planted signs are correct in every group. Evaluation includes all 64 sets and 192 directed additions in each of the 12 fits. Table 4 provides the complete selected-set summary for both conditions.

B.3 Recovery by interaction order

Figures 4-8 compare both parameterizations at each proper nonempty set size under the strong condition. Each figure uses the same three quantities as the main explanation figure: capacity values, mean addition gains, and signed interactions. Columns share scales within each row, and explicit set labels align the two models. Points show means and error bars show sample SD across three seeds; open diamonds show generating values. Shading marks planted nonzero interactions, not statistical significance.

This arrangement separates inherited structure from new interactions. For example, sets of size four and five can have substantial capacity and addition gain even though their generating signed interactions are all zero. Their fitted finite differences show the remaining discrepancy directly. Table 5 completes the comparison with the empty and full sets, retaining all 64 sets per model.

Table 4: Planted interaction recovery in all model-condition groups. Values are mean ± SD over three seeds.

Model	Condition	Set	True d	Fitted d	Absolute error
Progressive	Moderate	$\{1,3\}$	.20	$.2036 \pm .0012$	.0036
Progressive	Moderate	$\{2,4\}$	.20	$.1960 \pm .0038$	.0040
Progressive	Moderate	$\{1,2,3\}$	.15	$.1371 \pm .0043$	.0129
Progressive	Moderate	$\{5,6\}$	$-.20$	$-.2036 \pm .0046$	.0047
Progressive	Strong	$\{1,3\}$	.20	$.2044 \pm .0147$	.0126
Progressive	Strong	$\{2,4\}$	.20	$.1939 \pm .0086$	.0066
Progressive	Strong	$\{1,2,3\}$	.15	$.1223 \pm .0204$	.0277
Progressive	Strong	$\{5,6\}$	$-.20$	$-.2099 \pm .0191$	.0163
Recursive	Moderate	$\{1,3\}$	.20	$.2005 \pm .0030$	.0024
Recursive	Moderate	$\{2,4\}$	.20	$.1962 \pm .0025$	.0038
Recursive	Moderate	$\{1,2,3\}$	.15	$.1445 \pm .0043$	.0055
Recursive	Moderate	$\{5,6\}$	$-.20$	$-.1961 \pm .0037$	.0039
Recursive	Strong	$\{1,3\}$	.20	$.1981 \pm .0097$	.0068
Recursive	Strong	$\{2,4\}$	.20	$.1888 \pm .0066$	.0112
Recursive	Strong	$\{1,2,3\}$	.15	$.1345 \pm .0080$	.0155
Recursive	Strong	$\{5,6\}$	$-.20$	$-.1863 \pm .0116$	.0137

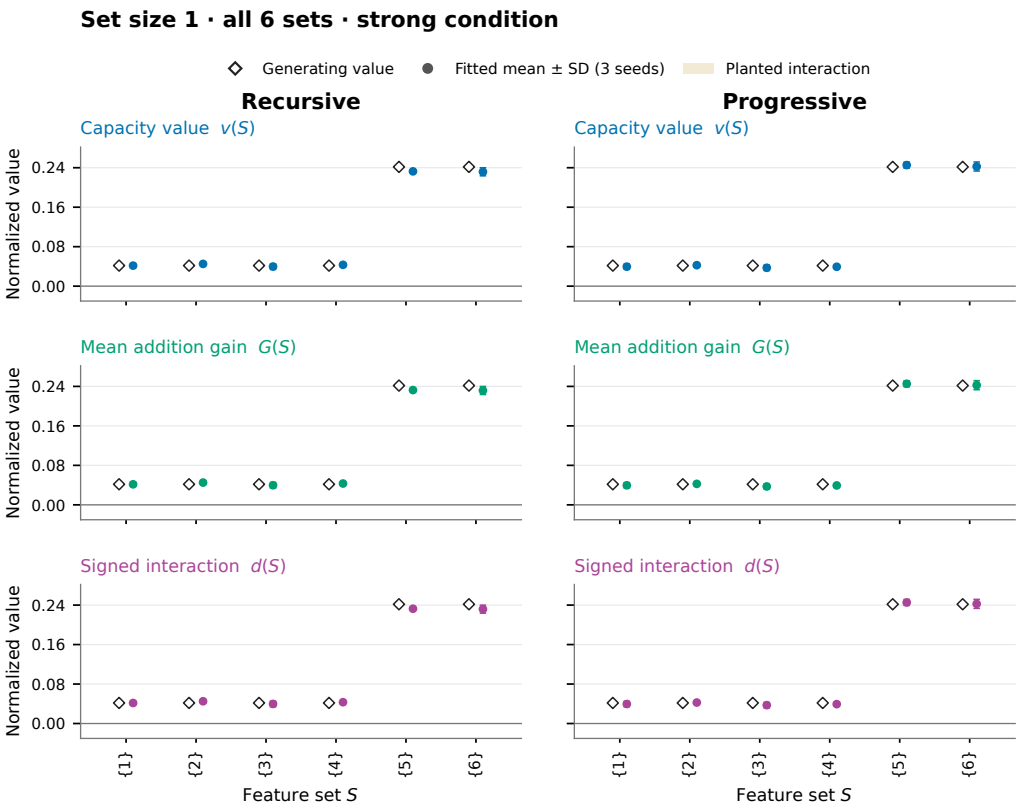


Figure 4: All six singletons under the strong condition. At this order, capacity, addition gain, and the singleton coefficient coincide. The larger values for features 5 and 6 allow the generating measure to remain monotone despite its negative pair interaction.

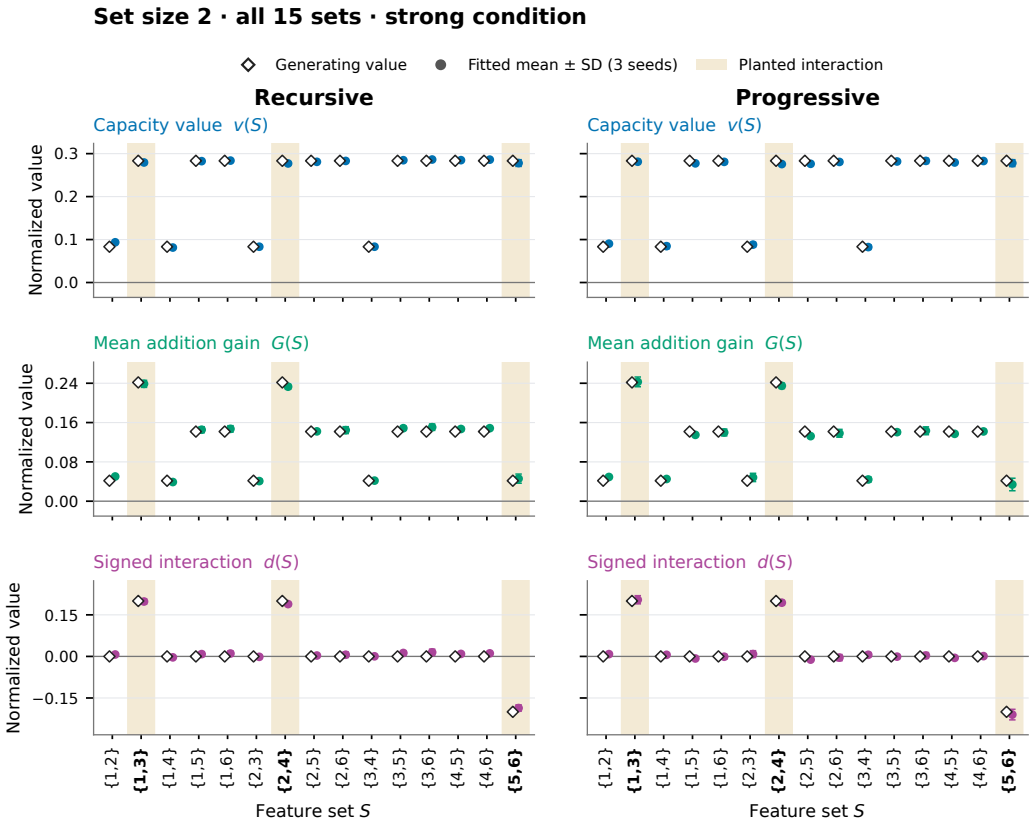


Figure 5: All 15 pairs under the strong condition. The positive pairs $\{1, 3\}$ and $\{2, 4\}$ and negative pair $\{5, 6\}$ are highlighted. The three rows distinguish similar capacity values from different addition gains and opposite signed interactions.

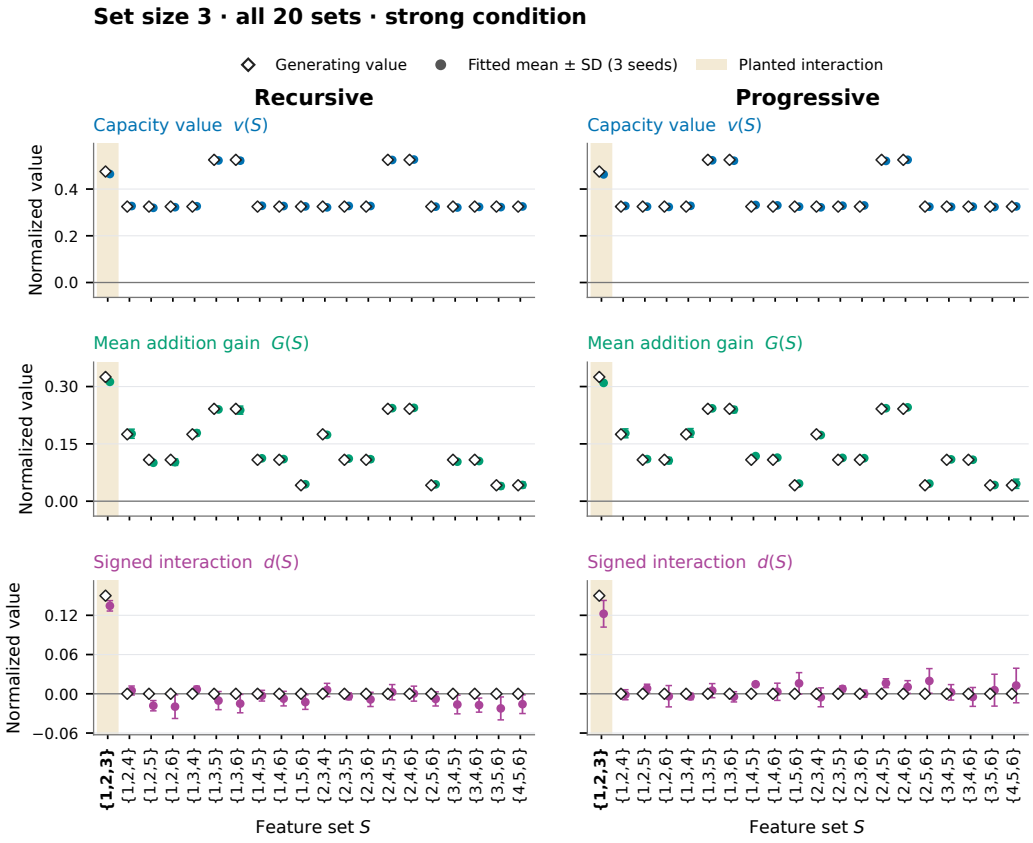


Figure 6: All 20 triples under the strong condition. Only $\{1, 2, 3\}$ has a nonzero generating triple interaction. Other triples can still have large capacity values or addition gains because they inherit singleton and pair effects.

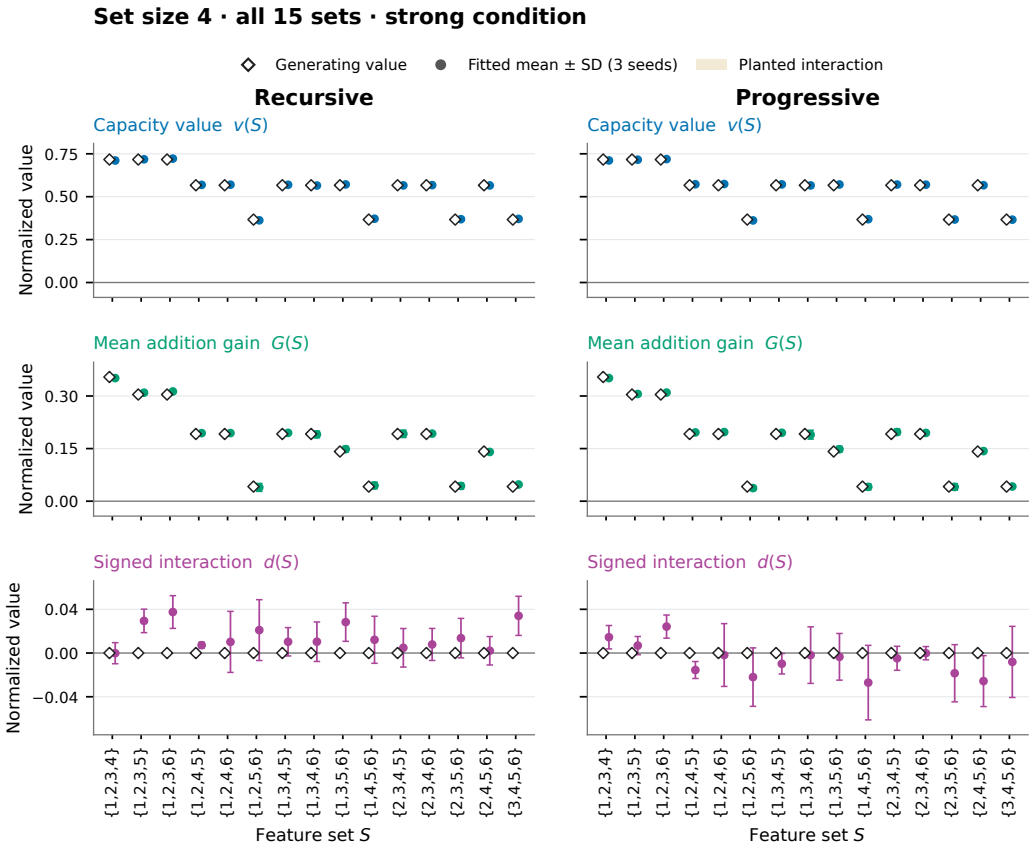


Figure 7: All 15 sets of size four under the strong condition. Generating signed interactions are zero throughout this order. Nonzero capacity values and addition gains arise from lower-order terms; fitted signed interactions display estimation discrepancies.

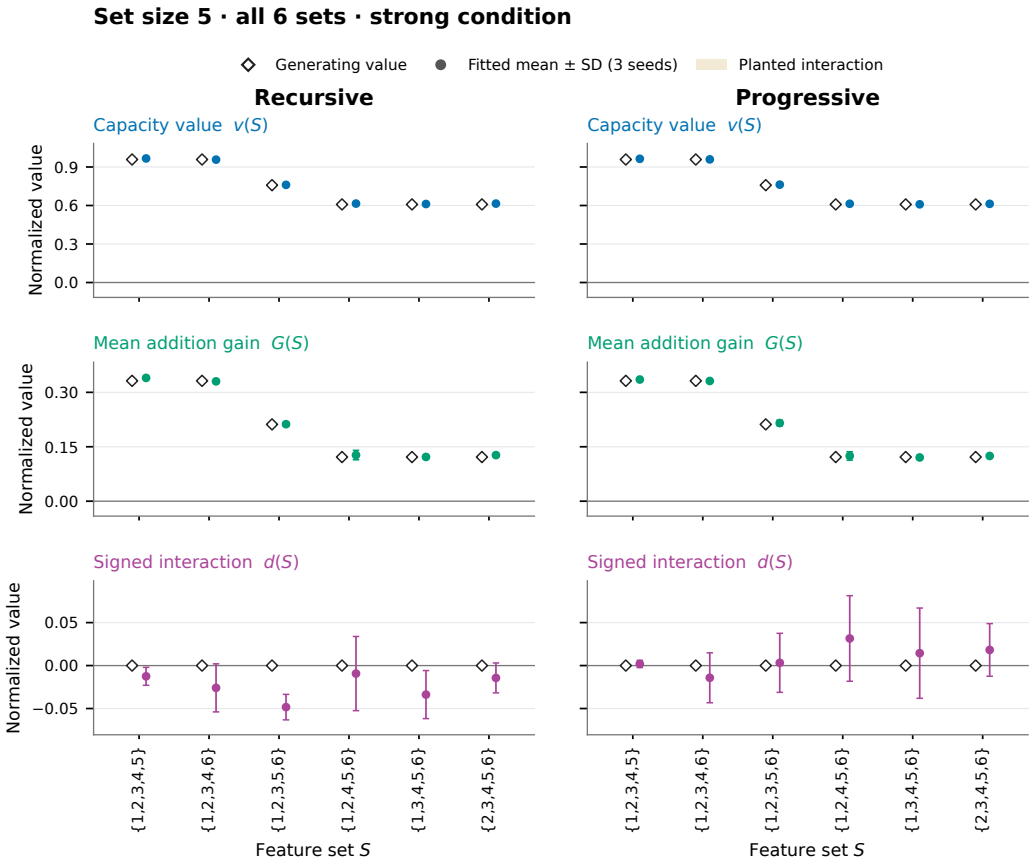


Figure 8: All six sets of size five under the strong condition. Generating signed interactions remain zero, while capacity and addition gains retain lower-order effects. Scales are shared between models within each row and include all error bars.

Table 5: Boundary sets under the strong condition. Normalization fixes the empty- and full-set capacity values. The generating full-set interaction is zero, although its immediate gain need not be zero. Fitted values are mean ± sample SD over three seeds. The empty-set interaction is zero; its mean addition gain is undefined.

Measure	$v(\varnothing)$	$v(\mathcal{N})$	$G(\mathcal{N})$	$d(\mathcal{N})$
Generating	0	1	0.2500	0
Recursive	0	1	0.2457 ± 0.0017	0.0269 ± 0.0479
Progressive	0	1	0.2465 ± 0.0021	-0.0180 ± 0.0750

Table 6: Static synthetic comparison averaged across dimensions and seeds.

Condition	Generator	Progressive	Recursive	Boosted trees
Smaller sample	Higher order	.01027	.00074	.00486
Smaller sample	Pairs	.00252	.00048	.00526
Strong	Higher order	.00386	.00099	.00252
Strong	Pairs	.00526	.00032	.00266
Moderate	Higher order	.00632	.00069	.00318
Moderate	Pairs	.00101	.00044	.00325

B.4 Additional synthetic comparisons

With a separate static study, we compare boosted trees, recursive capacity, and progressive capacity on pair-only and higher-order generators at $n \in \{6, 8, 10, 12\}$, three conditions varying input dependence or sample size, and three seeds. Table 6 averages the dimension-specific noise-free RMSE values only as a compact overview. Compared with the progressive model, recursive capacity has lower set-interaction error in 69 of 72 matched cases and lower prediction error in 53. Tree set diagnostics are evaluated at Boolean inputs because trees have no native fuzzy measure; they should not be confused with learned capacity parameters.

GRU-D was also evaluated through prediction-based Boolean probes on six-feature longitudinal data. Across three seeds, standardized MSE was 0.000309 ± 0.000045 for the pair generator and 0.000623 ± 0.000036 for the higher-order generator. Rankings of generating pair interactions were less consistent than rankings by addition gain. This is a function-level explanation of the comparator, not a native fuzzy measure.

C Clinical experiments

C.1 Variables and harmonization

The ten directly measured Choquet biomarkers are height, weight, waist circumference, systolic and diastolic blood pressure, glucose, insulin, HDL cholesterol, LDL cholesterol, and triglycerides. The eleven AHEI-2010 inputs are vegetables, fruit, whole grains, sugar-sweetened beverages and fruit juice, nuts and legumes, red and processed meat, trans fat, long-chain omega-3 fats, polyunsaturated fatty acids, sodium, and alcohol. The AHEI-2010 definition follows Chiuve et al. (2012); long-chain omega-3 and polyunsaturated fat remain distinct components.

The 14 outcomes are the ten input biomarkers, BMI, HOMA-IR, the WHI 0-18 depression score, and the CARDIA 0-60 depression score. The instruments are not pooled. Screening thresholds of 5 and 16, respectively, are secondary probability outputs rather than diagnostic labels.

Weight units differed by cohort and were corrected before modeling. We used nominal cohort-specific examination years, and values recorded before their documented examination were excluded. AHEI components outside $[0, 10]$ were clipped to that interval. This affected 3,360,612 of 19,737,124 available AHEI component entries in imputation 1. Clipping enforces the score range but does not recover the original dietary scores.

Table 7: Complete clinical prediction comparison. SD is across ten imputations and three seeds, except progressive Choquet, which uses three seeds on imputation 1. Intervals resample test participants and are not pairwise difference tests.

Model	Fits	Macro standardized MSE	Participant 95% interval
Boosted trees	30	0.50429 ± 0.00049	$[0.47355, 0.53504]$
GRU-D	30	0.50761 ± 0.00038	$[0.47692, 0.53900]$
Progressive Choquet	3	0.51423 ± 0.00032	$[0.48491, 0.54465]$
Matched Transformer	30	0.51559 ± 0.00286	$[0.48565, 0.54575]$
Recursive Choquet	30	0.52149 ± 0.00021	$[0.49131, 0.55268]$

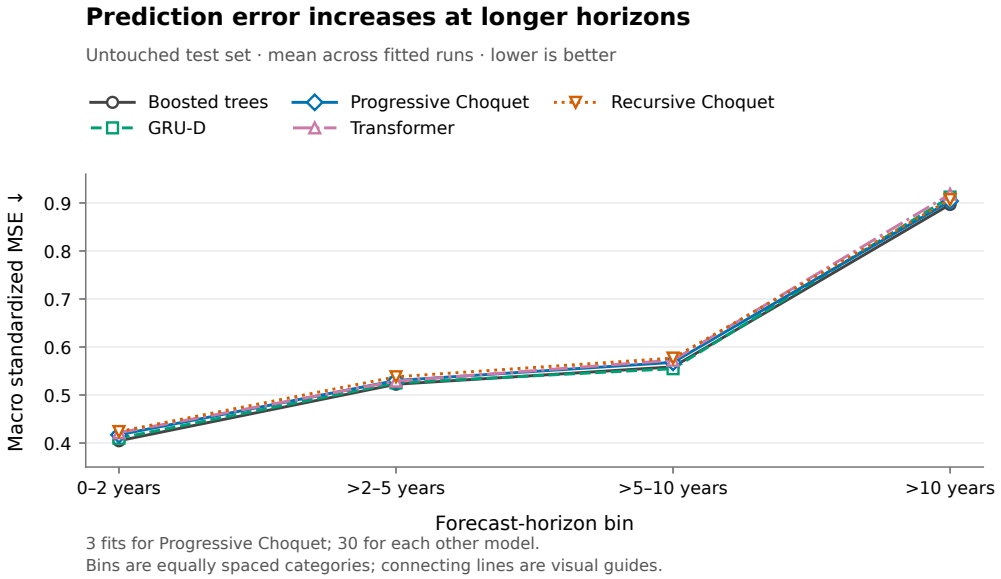


Figure 9: Test prediction error across forecast-horizon bins. Points show mean macro standardized MSE across three fits for Progressive Choquet and 30 fits for each other model. The bins are $(0, 2]$, $(2, 5]$, $(5, 10]$, and > 10 years; their spacing is categorical and connecting lines are visual guides. No uncertainty intervals are shown.

The ten supplied imputations share 166,836 participant identities and the same split. The original imputation method and observed-value indicators were unavailable. Available values may therefore include imputations. All models, including GRU-D, use availability after preprocessing.

C.2 Outcome and horizon results

The complete comparison in Table 7 averages across outcomes and forecast horizons. We examine these dimensions separately to show where that summary hides differences. For the two depression scores, boosted trees obtain MSE 0.7918 (WHI) and 0.9095 (CARDIA). GRU-D obtains 0.8088 and 0.9069; progressive Choquet 0.8115 and 0.9100; the Transformer 0.8120 and 0.9108; and recursive Choquet 0.8119 and 0.9078. These errors are measured after standardizing each outcome. Across forecast horizons, every model's mean error increases, with the highest values beyond ten years (Figure 9).

C.3 Robustness

To assess sensitivity to changes in the available inputs, we evaluated the fixed imputation-1, three-seed groups without retuning. The perturbations were 20% random deletion of available history, 0.1 training-SD measurement noise, and a $+0.2$ training-SD shift of named inputs. Table 8 reports the mean relative change in test MSE from each fit's unperturbed evaluation.

Table 8: Relative change in test macro standardized MSE under controlled perturbations, averaged over three seeds on imputation 1. Positive values indicate increased error. SD in the column headings refers to the training-data scale of the perturbation, not uncertainty.

Model	History deletion 20%	Measurement noise 0.1 SD	Input shift +0.2 SD
Boosted trees	+6.95%	+0.70%	+1.64%
GRU-D	+2.70%	+0.27%	+0.25%
Progressive Choquet	+2.14%	−0.03%	+1.13%
Transformer	+3.10%	+0.20%	+0.54%
Recursive Choquet	+0.87%	+0.06%	+1.48%

History deletion produces the largest change for boosted trees ($+6.95\%$), compared with $+0.87\%$ for recursive Choquet and $+2.14\%$ for progressive Choquet. Measurement noise changes mean error by less than 1% for every model. Input shift has a larger effect than deletion for recursive Choquet ($+1.48\%$ versus $+0.87\%$), illustrating that sensitivity depends on the perturbation. The small error reduction for progressive Choquet (-0.03%) does not establish a benefit from added noise. Random deletion measures sensitivity to missing history; it does not reconstruct the unavailable source missingness process. Original calibration quantiles are retained, so the accompanying coverage results measure degradation without recalibration.

C.4 Clinical fuzzy-measure explanations

The preceding comparisons assess prediction. To inspect how the fitted model combines features, we next examine its fuzzy measure. We evaluated 88,080,384 set values and 924,844,032 directed additions across 14 outcomes and three time functions for the representative recursive fit. Figure 10 provides an overview: each row follows one named feature, and each column averages its addition gain at a particular target-set size. The shared color scale makes these averages comparable across clinical measurements and diet component scores.

The largest broad feature/size averages are height at set size five (0.1118), weight at size six (0.1112), and weight at size seven (0.1108). These values locate prominent patterns in the fitted measure, but averaging cannot identify the particular sets responsible. Moreover, an addition to a larger set can inherit lower-order interactions, as Equation (9) shows. A question about a named combination therefore requires its individual addition gains and signed interaction. Relating those quantities to a prediction further requires the inputs, outcome coefficients, and horizon weights.

We calculated pair and triple interactions for the same recursive fit (imputation 1, seed 1729). Each of its 42 capacities contains 210 pairs and 1,330 triples, giving 64,680 coefficients in total. These quantities depend only on the fitted capacities.

Figures 11-16 cover every one of the 14 outcomes under each of the three time functions. For each outcome, time function, and order, the displayed positive term is the largest positive $d(S)$ and the displayed negative term is the most negative $d(S)$ among all 210 pairs or 1,330 triples. A missing negative pair is shown explicitly rather than replaced by a near-zero or positive term.

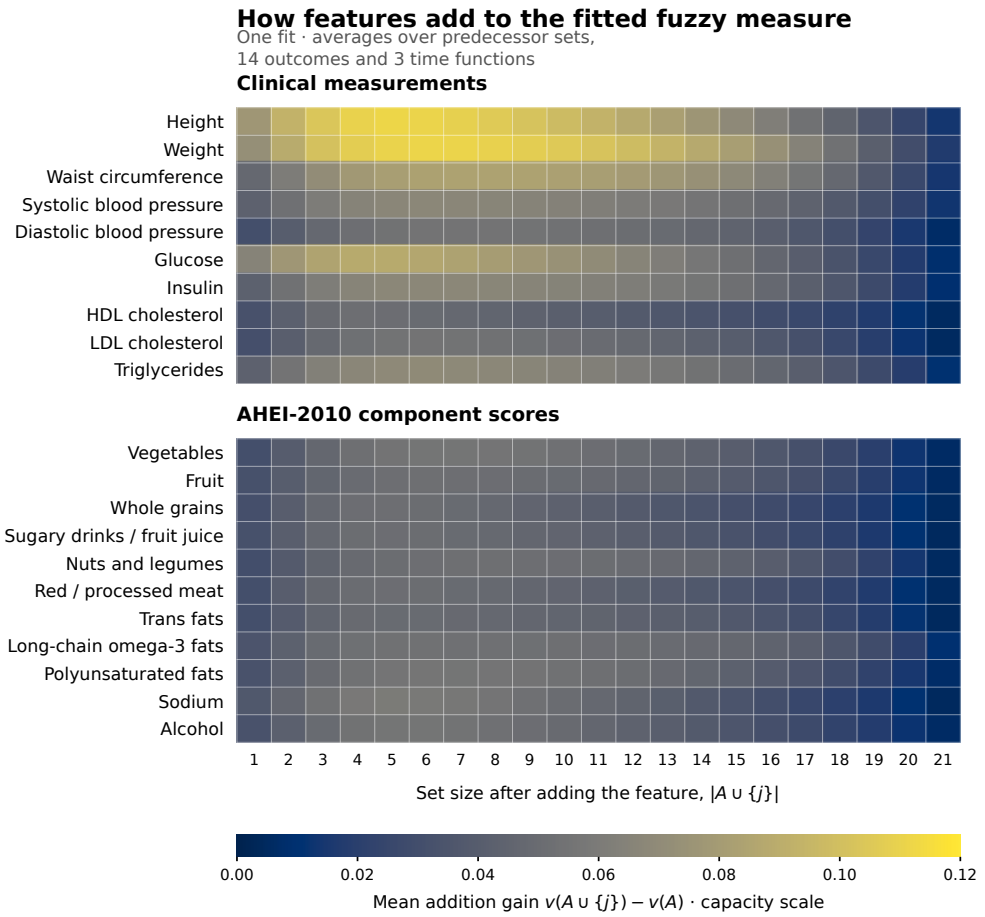
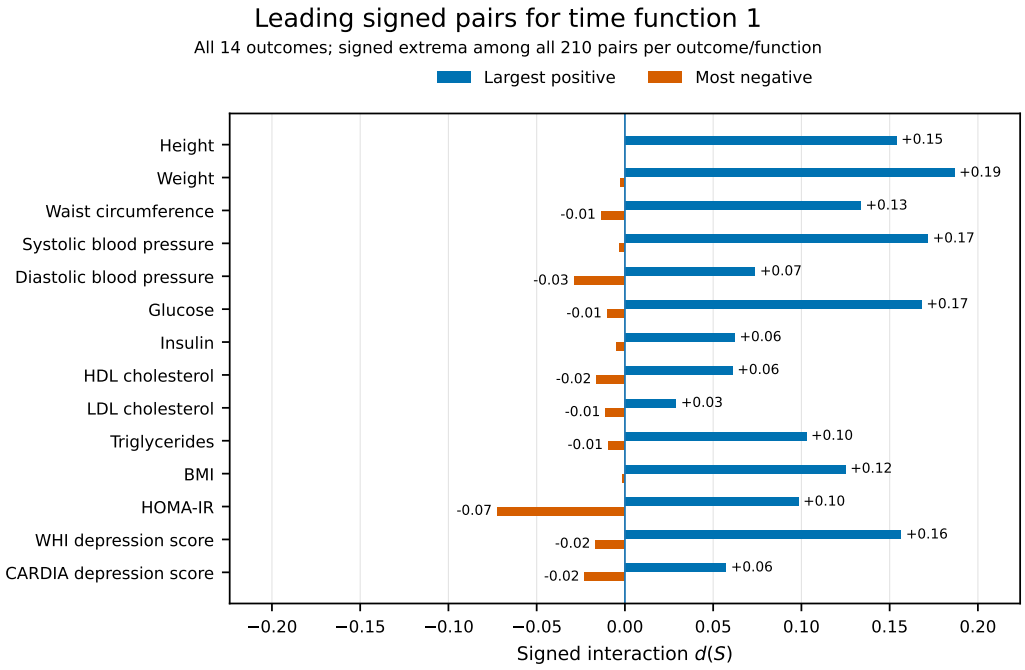
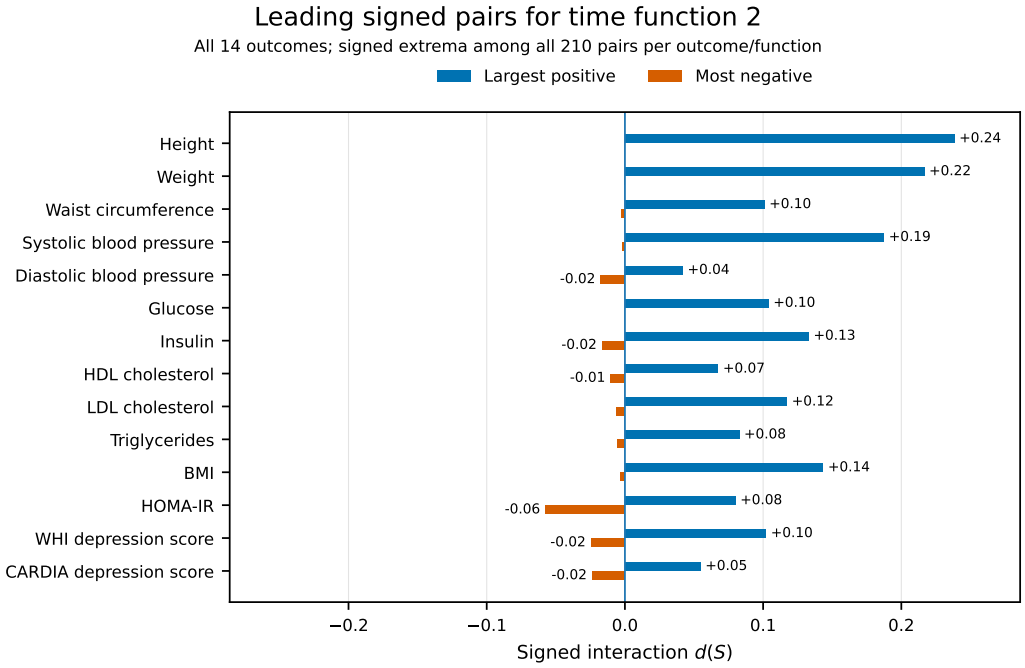


Figure 10: Mean feature-addition gains in the representative clinical fuzzy measure. Each cell averages $v(A \cup \{j\}) - v(A)$ over sets A for which $|A| + 1$ equals the displayed size, then equally across 14 outcomes and three time functions from one fitted model. Clinical measurements and AHEI-2010 component scores share a linear color scale starting at zero. The x-axis gives the size after addition; it does not indicate a pure interaction order. Gains describe the fitted capacity, not changes in outcome units or causal effects.



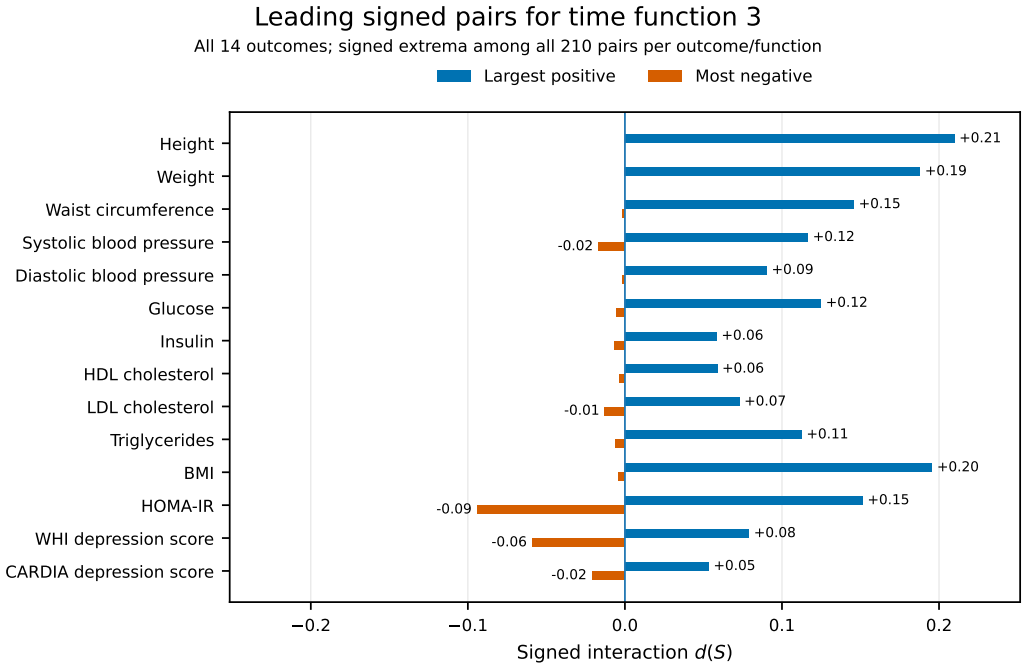
Outcome	Selected negative set	Selected positive set
Height	none < 0	Height+TG
Weight	LDL+Vegetables	Weight+Waist
Waist circumference	Height+Vegetables	Weight+Waist
Systolic blood pressure	Systolic BP+Diastolic BP	Systolic BP+Sodium
Diastolic blood pressure	Vegetables+Sodium	Waist+Glucose
Glucose	Glucose+Insulin	Weight+Glucose
Insulin	Weight+PUFA	Weight+Waist
HDL cholesterol	TG+Alcohol	Systolic BP+Insulin
LDL cholesterol	Height+Alcohol	Insulin+PUFA
Triglycerides	Systolic BP+Vegetables	Diastolic BP+TG
BMI	Height+Vegetables	Weight+Fruit
HOMA-IR	Weight+Glucose	Height+Weight
WHI depression score	Insulin+Vegetables	Diastolic BP+Alcohol
CARDIA depression score	Glucose+Sodium	Insulin+TG

Figure 11: Selected pair interactions for clinical time function 1. Each outcome shows its largest positive and most negative pair Möbius coefficient $d(S)$ among all 210 pairs; a missing sign is labeled explicitly. The table below the bars names the selected negative and positive sets.



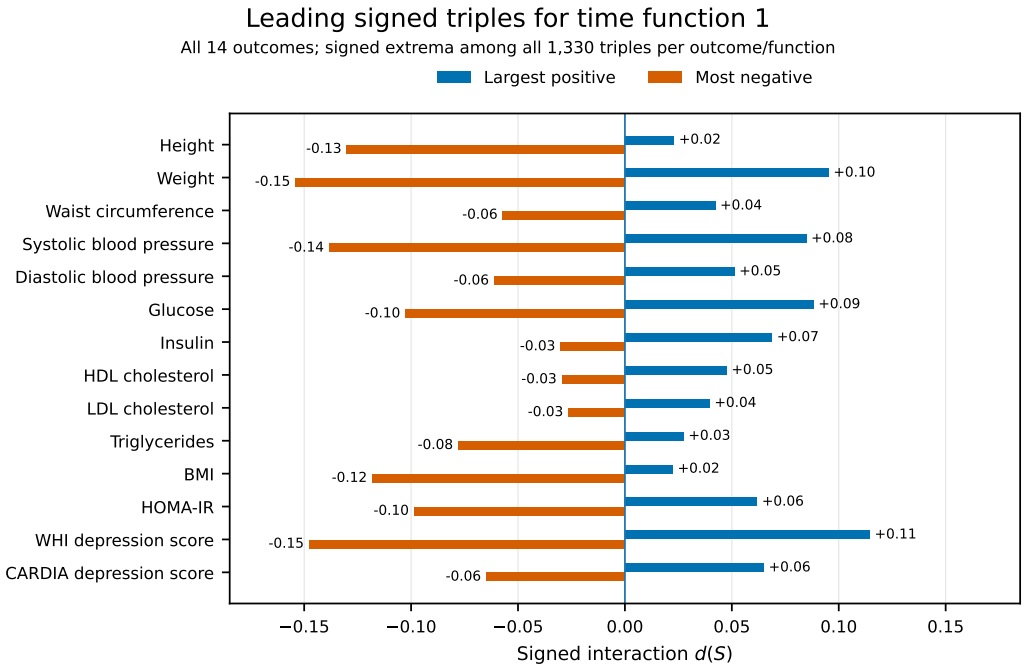
Outcome	Selected negative set	Selected positive set
Height	none < 0	Height+Vegetables
Weight	LDL+Vegetables	Height+Weight
Waist circumference	Height+Vegetables	Weight+Diastolic BP
Systolic blood pressure	Height+Alcohol	Systolic BP+Sodium
Diastolic blood pressure	Glucose+Omega-3	Systolic BP+Fruit
Glucose	none < 0	Glucose+Insulin
Insulin	Weight+Systolic BP	Weight+Insulin
HDL cholesterol	Vegetables+Alcohol	Insulin+Nuts/legumes
LDL cholesterol	Vegetables+PUFA	Glucose+Omega-3
Triglycerides	Height+LDL	TG+Vegetables
BMI	Vegetables+Alcohol	Weight+Waist
HOMA-IR	Weight+Glucose	Waist+Insulin
WHI depression score	LDL+Sodium	Fruit+Nuts/legumes
CARDIA depression score	Height+Insulin	Vegetables+Red/processed meat

Figure 12: Selected pair interactions for clinical time function 2. All 14 outcomes are shown. Values are descriptive coefficients of the fitted capacity.



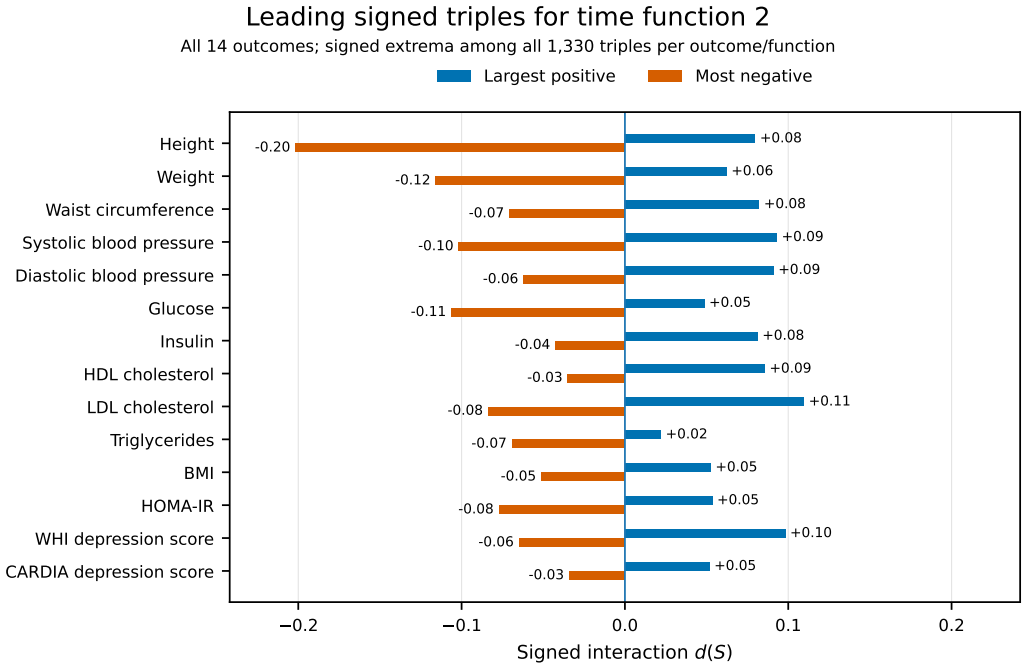
Outcome	Selected negative set	Selected positive set
Height	none < 0	Height+Alcohol
Weight	none < 0	Weight+Waist
Waist circumference	Height+Vegetables	Weight+Waist
Systolic blood pressure	Glucose+Sodium	Systolic BP+Sodium
Diastolic blood pressure	Vegetables+Whole grains	Height+Diastolic BP
Glucose	LDL+Vegetables	Glucose+Insulin
Insulin	Vegetables+Sodium	Weight+Waist
HDL cholesterol	Vegetables+Alcohol	Weight+Vegetables
LDL cholesterol	Insulin+Omega-3	Glucose+Omega-3
Triglycerides	Height+LDL	Diastolic BP+TG
BMI	Vegetables+Alcohol	Weight+Waist
HOMA-IR	Glucose+Insulin	Weight+Insulin
WHI depression score	Height+Trans fats	Trans fats+Omega-3
CARDIA depression score	Height+Insulin	Vegetables+Red/processed meat

Figure 13: Selected pair interactions for clinical time function 3. The selection rule is the same as in Figure 11.



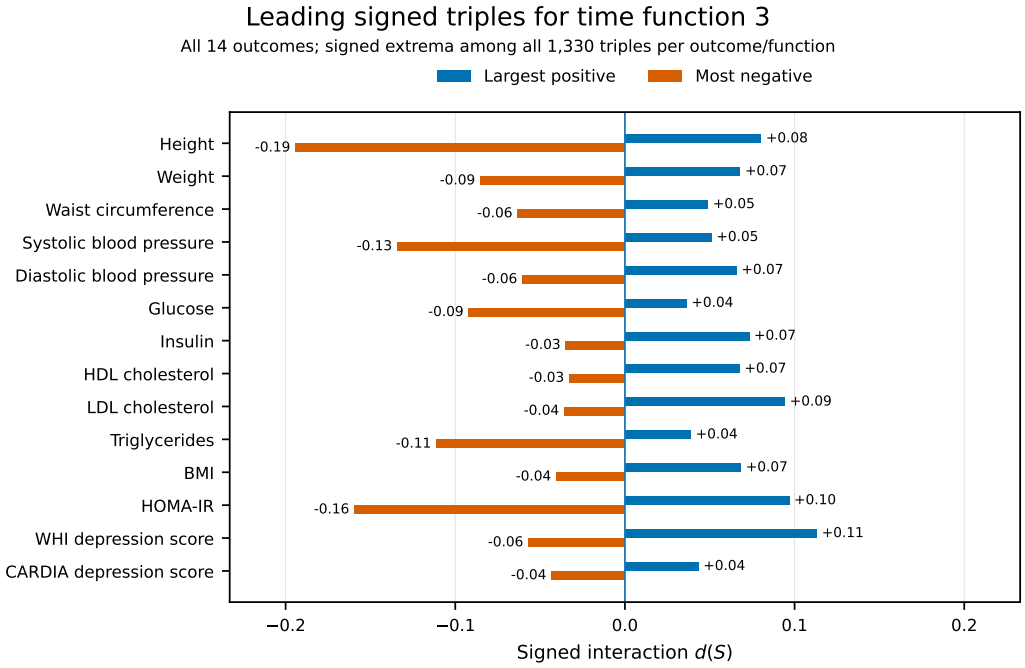
Outcome	Selected negative set	Selected positive set
Height	Height+LDL+TG	Height+Vegetables+Nuts/legumes
Weight	Height+Weight+Waist	Weight+Waist+Systolic BP
Waist circumference	Weight+Diastolic BP+TG	Weight+Glucose+Vegetables
Systolic blood pressure	Systolic BP+Glucose+Sodium	Systolic BP+TG+Sodium
Diastolic blood pressure	Waist+Glucose+Vegetables	Waist+Glucose+Insulin
Glucose	Weight+Glucose+Vegetables	Weight+Waist+Glucose
Insulin	Weight+Waist+Glucose	Weight+Waist+Diastolic BP
HDL cholesterol	Weight+Systolic BP+PUFA	Diastolic BP+TG+Sodium
LDL cholesterol	Glucose+Insulin+PUFA	Trans fats+PUFA+Alcohol
Triglycerides	Glucose+LDL+TG	Waist+Insulin+TG
BMI	Weight+Diastolic BP+Fruit	Height+Weight+Alcohol
HOMA-IR	Weight+Waist+PUFA	Weight+Glucose+Insulin
WHI depression score	Diastolic BP+PUFA+Alcohol	Height+Waist+PUFA
CARDIA depression score	Insulin+LDL+TG	Weight+Insulin+PUFA

Figure 14: Selected triple interactions for clinical time function 1. Each outcome shows its largest positive and most negative $d(S)$ among all 1,330 triples. The table below the bars names the selected negative and positive sets.



Outcome	Selected negative set	Selected positive set
Height	Height+Vegetables+Alcohol	Height+LDL+Red/processed meat
Weight	Height+Weight+Sodium	Weight+Glucose+Vegetables
Waist circumference	Weight+Waist+Nuts/legumes	Weight+Waist+TG
Systolic blood pressure	Systolic BP+Glucose+Sodium	Systolic BP+TG+Sodium
Diastolic blood pressure	Systolic BP+Glucose+Fruit	Waist+Systolic BP+Fruit
Glucose	Glucose+Insulin+Vegetables	Systolic BP+Glucose+Fruit
Insulin	Weight+Waist+Insulin	Weight+Diastolic BP+Insulin
HDL cholesterol	Insulin+HDL+Nuts/legumes	Insulin+Omega-3+PUFA
LDL cholesterol	Waist+Glucose+Omega-3	Glucose+TG+Omega-3
Triglycerides	Weight+TG+Vegetables	TG+Omega-3+Alcohol
BMI	Weight+Waist+Glucose	Weight+Waist+PUFA
HOMA-IR	Weight+Waist+LDL	Height+Waist+PUFA
WHI depression score	TG+Fruit+Omega-3	Vegetables+Fruit+Omega-3
CARDIA depression score	Height+LDL+TG	Whole grains+Sugary drinks+Nuts/legumes

Figure 15: Selected triple interactions for clinical time function 2, using the same selection rule as Figure 14.



Outcome	Selected negative set	Selected positive set
Height	Height+Omega-3+Alcohol	Height+Trans fats+Omega-3
Weight	Height+Weight+Sodium	Weight+Glucose+Vegetables
Waist circumference	Weight+Waist+Diastolic BP	Weight+Waist+TG
Systolic blood pressure	Systolic BP+Glucose+Sodium	Systolic BP+Red/processed meat+Sodium
Diastolic blood pressure	Height+Weight+Diastolic BP	Height+Diastolic BP+PUFA
Glucose	Waist+Systolic BP+Glucose	Height+Glucose+TG
Insulin	Weight+Waist+Diastolic BP	Weight+Systolic BP+Diastolic BP
HDL cholesterol	Systolic BP+Insulin+HDL	Weight+TG+Vegetables
LDL cholesterol	Diastolic BP+Insulin+PUFA	Diastolic BP+Glucose+PUFA
Triglycerides	Diastolic BP+TG+Nuts/legumes	Glucose+Vegetables+Nuts/legumes
BMI	Weight+Waist+Glucose	Weight+Waist+Diastolic BP
HOMA-IR	Weight+Waist+Insulin	Weight+Insulin+Nuts/legumes
WHI depression score	Diastolic BP+Vegetables+Red/processed meat	Vegetables+Trans fats+Omega-3
CARDIA depression score	Insulin+Sodium+Alcohol	Height+PUFA+Sodium

Figure 16: Selected triple interactions for clinical time function 3, using the same selection rule as Figure 14.

Higher orders. The complete set of fuzzy measures is not restricted to pairs and triples. Table 9 summarizes every order from 2 through 21, averaging the absolute Möbius coefficient equally across the 42 outcome/function measures and reporting the largest absolute coefficient observed at that order. The mean absolute coefficient grows strongly at high orders and the maximum reaches 2.100 at order 19. This is not a monotonicity violation: Möbius coefficients are alternating finite differences and need not lie in $[0, 1]$ even though capacities do. It is also a warning against overinterpreting high-order terms from one fit, since Equation (11) shows that finite-difference error sensitivity grows exponentially with order.

Table 9: All-order summary for the selected representative clinical capacity. "Sets / measure" is the number of sets of that order in one 21-feature measure. Mean $|d(S)|$ is first computed within each outcome/function/order summary and then averaged equally across the 42 measures; the maximum is over all measures. Positive/negative percentages pool the corresponding set counts across the 42 measures. Near-zero coefficients at the extraction tolerance are omitted from the displayed percentages.

Order	Sets / measure	Mean $\|d(S)\|$	Max $\|d(S)\|$	Positive / negative
2	210	0.0115	0.238	91.3% / 8.7%
3	1,330	0.0068	0.202	37.3% / 62.7%
4	5,985	0.0069	0.173	45.1% / 54.9%
5	20,349	0.0081	0.146	55.9% / 44.1%
6	54,264	0.0100	0.175	48.5% / 51.5%
7	116,280	0.0127	0.205	47.6% / 52.4%
8	203,490	0.0161	0.285	54.3% / 45.7%
9	293,930	0.0205	0.405	45.6% / 54.4%
10	352,716	0.0260	0.519	53.4% / 46.6%
11	352,716	0.0330	0.685	48.1% / 51.9%
12	293,930	0.0423	0.820	50.4% / 49.6%
13	203,490	0.0549	0.964	50.9% / 49.1%
14	116,280	0.0721	1.053	48.3% / 51.7%
15	54,264	0.0957	1.172	52.2% / 47.8%
16	20,349	0.1281	1.205	48.0% / 52.0%
17	5,985	0.1728	1.298	51.6% / 48.4%
18	1,330	0.2338	1.716	49.1% / 50.9%
19	210	0.3163	2.100	50.2% / 49.8%
20	21	0.4251	2.060	51.2% / 48.8%
21	1	0.5376	1.963	38.1% / 61.9%

The clinical heatmap and selected named-set figures answer different questions. The heatmap summarizes nonnegative immediate additions over many predecessor sets; the pair/triple figures isolate signed Möbius terms for particular named combinations. Neither supplies a clinical effect estimate.

D Ablations, comparisons, and training protocol

D.1 Progressive attention ablation

An attention variant of the model begins with learned feature descriptors e_j for the singleton sets. Shared self-attention produces $h_{\{j\}}$. For every set S of size $k \geq 2$, form all k decompositions

$$\mathcal{D}_S = \{(h_{S\setminus\{j\}}, h_{\{j\}}) : j \in S\}. \tag{19}$$

A size-specific cross-attention operation uses a query from the sum of singleton descriptors in S and keys and values from all members of $\mathcal{D}_S$. Its scalar output is added to an independent parameter b_S in $\theta_S = b_S + a_S$. The recursive map in Equation (3) then produces $v(S)$. The residual coordinate preserves the same ability to represent capacities with strict inequalities as the direct recursive model. Induction shows that all 2^n sets and all $n2^{n-1}$ directed additions are represented; permuting the enumeration of $\mathcal{D}_S$ leaves the attention result unchanged. Details and proof appear in Appendix A.5.

Progressive Choquet uses the same prediction and explanation equations as the main recursive model. Only the capacity parameterization and validation-selected temporal weights differ.

D.2 TIME AGGREGATION AND MODEL SELECTION

The development study compared mean, normalized sigmoid, softmax, and positive power-mean temporal aggregation; scalar sigmoid, probit, and softsign transforms; direct drift and equilibrium trajectories; and additive, max-closure, recursive, positive-mass, and progressive set functions. Successive halving first used a common budget and then extended only configurations that remained competitive or were needed for a structural comparison. Selection used validation participants only.

Softmax time aggregation, sigmoid scaling, recursive capacity, and equilibrium dynamics were selected for the primary clinical model. The progressive comparison used sigmoid time weights. An initial additive baseline achieved validation MSE 0.51533 in one fit, compared with the recursive model's three-seed mean of 0.52795. The later replicated additive comparisons are reported in Appendix E.4.

The clinical validation means for complete three-seed groups were 0.49829 for boosted trees, 0.51336 for GRU-D, 0.51585 for progressive Choquet, 0.52352 for the Transformer, and 0.52795 for recursive Choquet. Model selection preceded the original test evaluation.

D.3 CONSTRAINT ENFORCEMENT

The loss-based ablation defines a violation for adding one feature as

$$e(S, j) = \max\{0, u(S) - u(S \cup \{j\})\}$$

and adds the mean squared violation, the squared largest violation, boundary errors, and range errors to the prediction objective. We compared no penalty, fixed weights 1 and 100, and an adaptive weight that doubles when the largest violation remains above 10^{-5}. Zero violation for every single-feature addition implies monotonicity for all subset relations. A finite penalty does not guarantee zero violation in a nonconvex joint fit.

A six-feature comparison evaluated direct and progressive parameterizations under recursive constraints, no penalty, fixed penalties of 1 and 100, and an adaptive penalty. All ten fits ranked $\{1, 3\}$ and $\{2, 4\}$ as the two largest pair gains. This recovery result does not establish validity of the unpenalized or penalized capacities; recursive constraints guarantee monotonicity.

D.4 COMPARISON-MODEL SETTINGS

Boosted trees receive the same named feature histories through a fixed tabular encoding; depth three was selected by validation error. GRU-D widths 64, 128, and 256, decoder dropout 0.1, 0.3, and 0.5, and learning rates 10^{-3}, 3×10^{-4}, and 10^{-4} were screened. Width 128, dropout 0.5, and learning rate 10^{-4} were selected. The matched causal Transformer has three layers, width 128, eight attention heads, and the same horizon-conditioned outcome decoder. Unlike the Choquet models, GRU-D and the Transformer may use hidden representations internally; they are external comparisons and never supply features to the Choquet integral.

A neural controlled differential equation comparison (Kidger et al., 2020) used rectilinear paths and a four-layer vector field. Its clinical fit reached a best validation MSE 0.52090 at epoch four and did not improve for the next 36 epochs, taking 50.4 hours. It was not replicated or included in the test comparison because its validation error exceeded that of GRU-D and training was substantially slower.

D.5 TRAINING, STOPPING, AND RESERVED DATA

The primary neural clinical models use all eligible training histories, AdamW, gradient clipping at five, validation-based learning-rate reduction, at most 300 epochs, and early stopping after 40 epochs without improvement.

The original prediction comparison was fixed before calibration and test access. It contains 123 fits: 30 each for recursive Choquet, boosted trees, GRU-D, and the Transformer, plus three progressive fits. Participant split-conformal intervals use the calibration participant's largest absolute standardized error across all available future times for an outcome, making the resulting trajectory interval simultaneous within that participant-outcome trajectory. Two thousand participant bootstrap repetitions form the aggregate error intervals.

D.6 HIGHER-ORDER EXPLANATIONS ACROSS PREDICTORS

We compare explanations of the final predicted outcome across recursive Choquet, progressive Choquet, boosted trees, GRU-D, and the Transformer. Individual contributions use Shapley values (Lundberg & Lee, 2017). Synthetic interactions use Faith-Shap (Tsai et al., 2023) and Shapley–Taylor (Sundararajan et al., 2020) through order six; clinical interactions use Faith-Shap through order three. This provides the same explanation definition for every predictor. The interaction coefficients in this comparison differ from the internal Möbius coefficients of a Choquet capacity.

D.6.1 COMMON EXPLANATION DEFINITION

For a participant history x, let $g_x(S)$ be the model's prediction when the features in S retain their available histories and every remaining feature is replaced by its training mean at each available visit. A feature's entire history and latest-value input are changed together. Availability indicators, visit times, outcome anchors, and forecast horizons remain fixed. Thus, the explanation measures the contribution of feature values conditional on the available anchor and observation pattern. It does not explain the anchor itself. Predictions are expressed in standard deviations of training outcomes.

For explanation order $1 \le q \le n$, we use weights $w(S) = (n-1)/[\binom{n}{|S|}|S|(n-|S|)]$ and fit

$$\begin{aligned} \underset{\beta}{\text{minimize}} \quad & \sum_{\varnothing \neq S \subsetneq N} w(S) \left[g_x(S) - g_x(\varnothing) - \sum_{\substack{\varnothing \neq T \subseteq S \\ |T| \le q}} \beta_T \right]^2, \\ \text{subject to} \quad & \sum_{1 \le |T| \le q} \beta_T = g_x(N) - g_x(\varnothing). \end{aligned} \tag{20}$$

Order one gives individual Shapley values. Higher orders give Faith-Shap coefficients for sets of up to q features. We evaluate the terms of order q against the same order of the reference explanation. The endpoint constraint makes the coefficients sum to the prediction difference. All models use the same feature subsets and replacement rule.

For the exact synthetic comparison, define the prediction finite difference

$$m_x(T) = \sum_{U \subseteq T} (-1)^{|T|-|U|} g_x(U).$$

Shapley–Taylor assigns $m_x(T)$ to nonempty sets smaller than q, and assigns

$$\mathrm{ST}_q(T) = \sum_{U \supseteq T} \frac{m_x(U)}{\binom{|U|}{q}}, \qquad |T| = q.$$

Both methods reduce to individual Shapley values at order one and to the full finite-difference decomposition at order six for the six-feature experiments. We compare each method with its own generating-function reference.

D.6.2 SYNTHETIC EXPLANATION ACCURACY

The comparison uses the six-feature longitudinal generators with pair-only and higher-order interactions, mean temporal aggregation, and three seeds for data generation and training. Each dataset has 100,000 training histories, 20,000 validation histories, and 20,000 test histories, with six visits and two outcomes at four forecast horizons. The Choquet models learn their temporal summaries and outcome dynamics jointly with the capacity. Their configurations match the primary recursive and progressive variants: softmax and sigmoid temporal weights, respectively, with sigmoid feature scaling and equilibrium dynamics. The comparison models use their existing validation-selected fits. The generating capacity is $v(S) = 0.1|S|/6 + \sum_{T \subseteq S} w_T$. The pair-only generator assigns weight 0.45 to each of $\{1,3\}$ and $\{2,4\}$. The higher-order generator assigns weights 0.25, 0.25, 0.15, 0.15, and 0.10 to $\{1,3\}$, $\{2,4\}$, $\{1,2,3\}$, $\{1,2,3,4\}$, and the full feature set, respectively.

For each generator and seed, we sample 128 test histories and explain both outcomes at seven years. All 64 feature subsets are evaluated exactly. Reference explanations apply Equation 20 to the known

generating response with the identical history replacements and fixed anchors. Errors therefore compare the same quantity in the same outcome units. We average over participants and the two outcomes within each seed, then report the mean and sample SD across seeds. Sign agreement includes only reference coefficients with magnitude above 10^{-5}.

Table 10: Prediction RMSE on the 128 original test histories used for explanation evaluation, averaged over both outcomes at seven years (mean $\pm$ SD over three seeds). Outcomes use training standard deviations. These errors concern the original inputs, before feature replacement.

Model	Higher-order generator	Pair-only generator
Recursive	$0.000573 \pm 7.1 \times 10^{-5}$	0.00542 ± 0.00087
Progressive	0.000837 ± 0.0006	0.00339 ± 0.0022
Boosted trees	0.137 ± 0.008	0.131 ± 0.013
GRU-D	0.0357 ± 0.0035	0.0213 ± 0.0028
Transformer	0.0212 ± 0.0014	0.0177 ± 0.0022

Table 10 reports prediction accuracy on the original histories used for explanation evaluation. On the higher-order generator, recursive prediction RMSE is 0.00057, compared with 0.00084 for progressive Choquet, 0.13667 for trees, 0.03568 for GRU-D, and 0.02118 for the Transformer. Explanation errors therefore reflect both how well each predictor approximates the generating function and how its interactions behave under feature replacement. This comparison does not isolate an explanation advantage at matched prediction error.

Table 11: Higher-order generator: Faith-Shap attribution MAE at every order ($\times 10^3$; mean $\pm$ SD over three seeds). Each entry averages 128 histories and two outcomes per seed. The explanation includes terms up to the stated order; errors here concern terms of that order. Lower is better.

Order	Recursive	Progressive	Boosted trees	GRU-D	Transformer
1	0.169 ± 0.015	0.308 ± 0.28	115 ± 3.3	30.5 ± 2.1	22.2 ± 1.4
2	0.0672 ± 0.0063	0.15 ± 0.14	42.9 ± 2.3	15.6 ± 0.56	10.4 ± 0.38
3	0.04 ± 0.0095	0.0929 ± 0.085	11.7 ± 1	7.28 ± 0.36	5.63 ± 0.3
4	0.025 ± 0.0081	0.0636 ± 0.057	4.05 ± 0.44	3.51 ± 0.37	3.62 ± 0.35
5	0.0167 ± 0.0076	0.0492 ± 0.05	1.37 ± 0.19	1.56 ± 0.25	2.36 ± 0.22
6	0.0142 ± 0.011	0.0394 ± 0.047	0.439 ± 0.25	0.937 ± 0.29	1.84 ± 0.29

Table 12: Higher-order generator: Shapley–Taylor attribution MAE at every order ($\times 10^3$; mean $\pm$ SD over three seeds). Each entry averages 128 histories and two outcomes per seed. The explanation includes terms up to the stated order; errors here concern terms of that order. Lower is better.

Order	Recursive	Progressive	Boosted trees	GRU-D	Transformer
1	0.169 ± 0.015	0.308 ± 0.28	115 ± 3.3	30.5 ± 2.1	22.2 ± 1.4
2	0.0688 ± 0.011	0.149 ± 0.14	46.3 ± 2.9	18.1 ± 0.44	12 ± 0.32
3	0.0434 ± 0.012	0.102 ± 0.1	12.5 ± 1.2	8.29 ± 0.44	6.62 ± 0.35
4	0.0287 ± 0.0087	0.0733 ± 0.071	4.18 ± 0.45	3.77 ± 0.41	4.11 ± 0.38
5	0.0182 ± 0.0094	0.0536 ± 0.057	1.35 ± 0.14	1.56 ± 0.21	2.55 ± 0.27
6	0.0142 ± 0.011	0.0394 ± 0.047	0.439 ± 0.25	0.937 ± 0.29	1.84 ± 0.29

For the higher-order generator, recursive Choquet has the lowest mean attribution error at every order under both methods. Faith-Shap MAE is 4.0×10^{-5} for triples and 1.4×10^{-5} for sixth-order terms. Sign agreement is 0.987, 0.976, 0.965, and 0.917 at orders three through six. The reference signals become smaller and less frequent as order increases; relative MAE rises from 0.007 for triples to 0.090 for sixth-order terms. Table 15 reports signal magnitudes and counts alongside recovery.

For the pair-only generator, reference interactions above order two are zero. Recursive Choquet has the smallest errors at orders two through six, indicating fewer spurious higher-order terms; progressive Choquet has lower individual-attribution error. These results directly assess higher-order explanations of final predictions. The generators follow the Choquet family, so the evidence supports controlled recovery within that family.

Table 13: Pair-only generator: Faith-Shap attribution MAE at every order ($\times 10^3$; mean $\pm$ SD over three seeds). Each entry averages 128 histories and two outcomes per seed. The explanation includes terms up to the stated order; errors here concern terms of that order. Lower is better.

Order	Recursive	Progressive	Boosted trees	GRU-D	Transformer
1	1.27 ± 0.16	0.954 ± 0.5	83.9 ± 3.3	12.4 ± 1.7	12.7 ± 0.31
2	0.232 ± 0.1	0.301 ± 0.08	34.1 ± 2.5	6.25 ± 0.66	5.98 ± 0.29
3	0.0382 ± 0.017	0.0507 ± 0.026	7.33 ± 0.55	1.89 ± 0.16	2.22 ± 0.069
4	0.0109 ± 0.0064	0.021 ± 0.019	2.9 ± 0.16	0.947 ± 0.085	1.22 ± 0.072
5	0.00345 ± 0.0018	0.0108 ± 0.015	1.06 ± 0.097	0.317 ± 0.016	0.507 ± 0.058
6	$0.00269 \pm 9 \times 10^{-5}$	0.00621 ± 0.0064	0.359 ± 0.077	0.0692 ± 0.016	0.296 ± 0.13

Table 14: Pair-only generator: Shapley–Taylor attribution MAE at every order ($\times 10^3$; mean $\pm$ SD over three seeds). Each entry averages 128 histories and two outcomes per seed. The explanation includes terms up to the stated order; errors here concern terms of that order. Lower is better.

Order	Recursive	Progressive	Boosted trees	GRU-D	Transformer
1	1.27 ± 0.16	0.954 ± 0.5	83.9 ± 3.3	12.4 ± 1.7	12.7 ± 0.31
2	0.228 ± 0.1	0.299 ± 0.079	34 ± 2.6	6.32 ± 0.64	6.12 ± 0.23
3	0.036 ± 0.017	0.0511 ± 0.027	7.33 ± 0.53	1.94 ± 0.14	2.33 ± 0.052
4	0.0117 ± 0.0065	0.0217 ± 0.02	2.9 ± 0.16	0.971 ± 0.087	1.27 ± 0.062
5	0.00373 ± 0.0018	0.0116 ± 0.016	1.06 ± 0.091	0.321 ± 0.016	0.533 ± 0.076
6	$0.00269 \pm 9 \times 10^{-5}$	0.00621 ± 0.0064	0.359 ± 0.077	0.0692 ± 0.016	0.296 ± 0.13

D.6.3 Clinical explanation stability and sensitivity

We use the same three seeds and first imputed dataset for all five models. A fixed random sample contains 32 test participants from each cohort, using each participant's earliest eligible prediction time. We explain the cohort-specific depression score at five years across all 21 biomarker and dietary inputs. Each explanation uses two independent sets of 16,384 sampled proper subsets. Subset size s is sampled with probability proportional to $1/[s(21-s)]$, followed by a uniformly chosen subset of that size; complementary subsets are paired. Empty and full subsets are included exactly. We solve Equation 20 for each sample and average the two coefficient estimates.

We separate numerical variation between subset samples from variation between trained models. Coefficient agreement uses cosine similarity of the signed attribution vectors; a value of one indicates the same direction. Selection agreement uses the Jaccard overlap (intersection divided by union) of the three largest absolute individual contributions or ten largest absolute pair or triple contributions. Coefficients below 10^{-6} in magnitude are treated as zero for these comparisons. Sets contain only nonzero coefficients; cases where both sets or vectors are empty are omitted. This avoids ranking numerical noise when a model has few nonzero interactions. Results average over participants and the three seed pairs. The number of eligible comparisons is reported with the results because some tree predictions have no nonzero higher-order interactions.

Recursive Choquet has the highest agreement across training seeds for pairs and triples, with cosine similarities of 0.778 and 0.677 and selection overlaps of 0.492 and 0.413. The other models have triple cosine similarities between 0.104 and 0.480. Boosted trees have the highest individual-attribution stability (0.963). Trees contribute 191 individual, 163 pair, and 21 triple comparisons with at least one nonzero vector, compared with 192 at each order for the other models; their higher-order results therefore concern fewer participants.

Numerical agreement between subset samples is stronger than agreement across training seeds. For recursive Choquet, triple cosine similarity is 0.945 and top-ten overlap is 0.900; progressive Choquet gives 0.947 and 0.898. Triple cosine similarity ranges from 0.994 to 1.000 for the other predictors. Sampling therefore adds uncertainty, especially for small higher-order coefficients, even when the leading interactions are consistent.

To assess sensitivity, we replace the three features with the largest absolute Shapley values and measure the absolute prediction change. The reference is the mean change for 32 uniformly sampled

Table 15: Reference interaction strength and recursive recovery for the higher-order generator using Faith-Shap. Counts combine three seeds, 128 histories, and two outcomes. Nonzero means magnitude above 10^{-5}. Relative MAE divides attribution MAE by mean absolute reference attribution. Sign agreement is averaged over outcomes and seeds.

Order	Nonzero	Total	Mean $\lvert\text{reference}\rvert \times 10^3$	Relative MAE	Sign
1	4606	4608	285.5	0.001	1.000
2	5402	11520	47.1	0.001	0.993
3	4129	15360	5.519	0.007	0.987
4	1926	11520	1.226	0.020	0.976
5	494	4608	0.331	0.050	0.965
6	56	768	0.1577	0.090	0.917

Table 16: Clinical explanation stability across three training seeds. Cosine similarity compares signed coefficients. Selection overlap is Jaccard overlap of the three leading individual features or ten leading pairs or triples. Higher is more consistent.

Model	SHAP cos.	Pair cos.	Triple cos.	Top 3	Top 10 pairs	Top 10 triples
Recursive	0.941	0.778	0.677	0.707	0.492	0.413
Progressive	0.821	0.552	0.480	0.570	0.311	0.219
Boosted trees	0.963	0.591	0.450	0.832	0.422	0.367
GRU-D	0.946	0.646	0.248	0.704	0.451	0.226
Transformer	0.667	0.449	0.104	0.426	0.288	0.195

three-feature sets shared across participants and models. For pairs and triples, we compare the highest-ranked Faith-Shap set with the mean over all 210 pairs or 1,330 triples using

$$D_x(T) = \left| \sum_{U \subseteq T} (-1)^{|U|} g_x(N \setminus U) \right|.$$

This measures the joint change beyond lower-order removals at the full history. Faith-Shap uses information across feature subsets, so its ranking and this local sensitivity measure need not coincide.

For every model, selected features, pairs, and triples have greater mean prediction sensitivity than their respective reference sets. For recursive Choquet, the selected triple has mean joint sensitivity 0.0097, compared with 0.0003 across all triples. Sensitivity measures how strongly a fitted model uses a feature set; larger values do not establish more accurate explanations.

Agreement between different predictors falls sharply at higher orders (Figure 17). Across model pairs, triple cosine similarity ranges from -0.036 to 0.066, despite greater individual-attribution agreement for several pairs. Thus, stability within a model does not imply agreement about clinical interactions across models.

This clinical comparison evaluates consistency and prediction sensitivity, not biological correctness. It covers two outcomes, one imputation, and 64 participants. Training-mean replacement can produce combinations that are uncommon under the observed feature correlations. The results depend on the reference and on holding anchors fixed. Extending the clinical explanation study across participants and outcomes will require further validation and human evaluation by clinicians and domain experts. The synthetic experiments provide a direct assessment of recovery against known generating functions.

D.6.4 COMPUTATION

Calculations use four threads on an Intel Xeon Silver 4216 CPU (2.10 GHz). Fitted Choquet capacities are evaluated once and reused because they do not depend on the participant. The synthetic extension reuses all 64 subset predictions and calculates both explanation methods exactly. Clinical timings cover both larger subset samples and removal calculations through order three for 64 participants. The final coefficient calculation is reported separately. Loading and initial capacity construction are

Table 17: Numerical agreement between two independent samples of 16,384 subsets, with the model fixed. Triple MAE is the mean absolute difference between coefficient estimates.

Model	SHAP cosine	Pair cosine	Triple cosine	Triple overlap	Triple MAE
Recursive	1.000	0.994	0.945	0.900	0.000149
Progressive	1.000	0.996	0.947	0.898	0.000169
Boosted trees	1.000	1.000	1.000	1.000	2.94×10^{-10}
GRU-D	1.000	1.000	1.000	0.995	5.1×10^{-6}
Transformer	1.000	1.000	0.994	0.961	0.00013

Table 18: Clinical prediction sensitivity in standardized outcome units. The first two columns compare removal of the three SHAP-selected features with random three-feature sets. Pair and triple columns compare the highest-ranked set with the mean over all sets of that size, using the joint change beyond lower-order removals.

Model	Top 3	Random 3	Top pair	Mean pair	Top triple	Mean triple
Recursive	0.0406	0.0128	0.0115	0.00097	0.00965	0.000297
Progressive	0.0428	0.00595	0.0177	0.000158	0.0136	2.74×10^{-5}
Boosted trees	0.0369	0.0102	0.0155	0.000101	0.0027	2.27×10^{-6}
GRU-D	0.0652	0.0312	0.024	0.00103	0.00419	7.15×10^{-5}
Transformer	0.0749	0.0345	0.0432	0.00209	0.0143	0.000315

reported separately. These measurements describe the common explanation implementation and are not comparisons with specialized TreeSHAP or gradient-based explanation software.

D.7 Examples of clinical interactions

The complete clinical capacity contains 2,097,152 sets for each of 14 outcomes and three time functions. A selected representative recursive fit (imputation 1, seed 1729) was evaluated for all 88,080,384 set values and all 924,844,032 directed additions. This is one of the 30 recursive fits included in the predictive comparison; its interaction values are descriptive.

We selected the largest positive and most negative $d(S)$ separately among all pairs and among all triples across the 42 outcome/function measures. Figure 18 shows the four resulting terms. For Height under function 2, Height+Vegetables has $d = +0.2383$, capacity $v = 0.7131$, and mean immediate gain $G = 0.4757$. For HOMA-IR under function 3, Glucose+Insulin has $d = -0.0942$ despite $v = 0.2237$ and $G = 0.0648$. Among triples, Height+Waist circumference+Polyunsaturated fats for the WHI depression score under function 1 has $d = +0.1145$, while Height+Vegetables+Alcohol for Height under function 2 has $d = -0.2017$; the latter still has positive $v = 0.7704$ and $G = 0.2831$ because monotonicity constrains additions, not the sign of the Möbius term.

The examples were selected from 210 pairs and 1,330 triples per capacity. Their coefficient signs describe the fitted aggregation and do not establish clinical effect directions. The cross-fit results in Section 6.4 assess their broader reliability. Appendix C.4 reports the other outcome/function combinations.

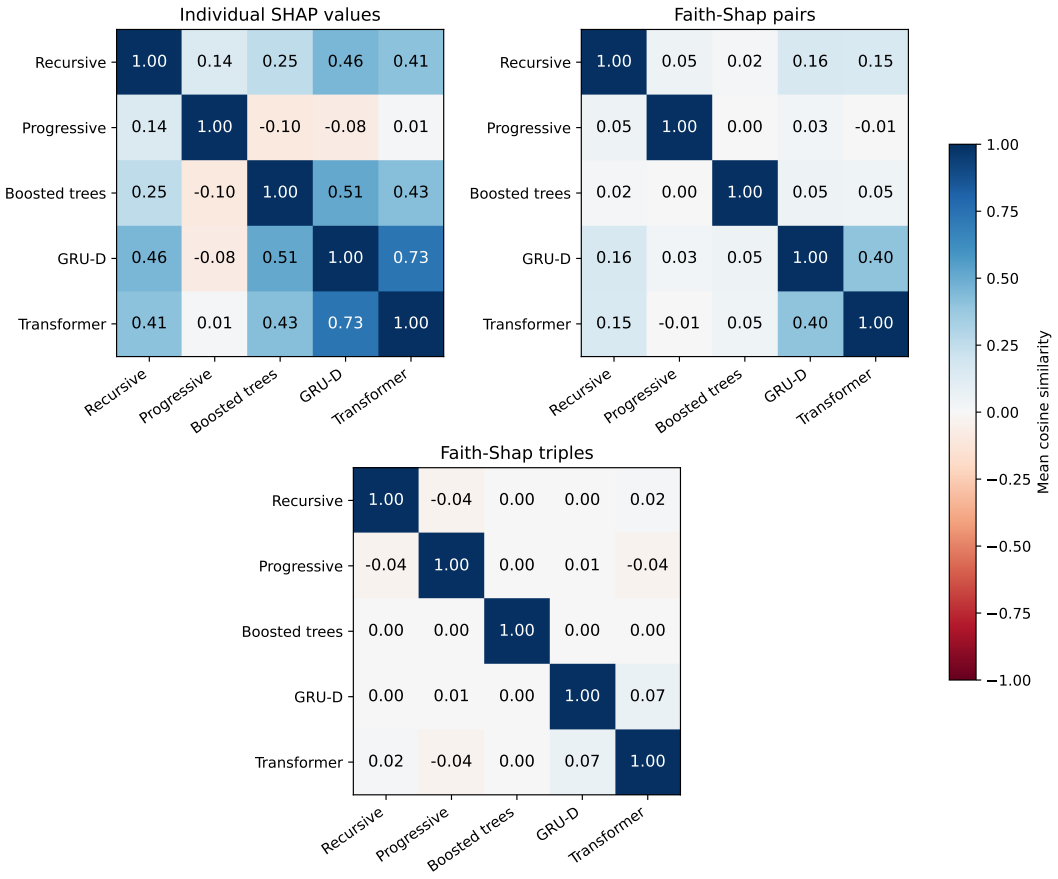


Figure 17: Agreement of clinical prediction explanations between models. Entries are mean cosine similarities of signed individual Shapley values (upper left), Faith-Shap pairs (upper right), and Faith-Shap triples (lower panel), averaged over the same participants and matched training seeds. Agreement measures consistency between fitted models; clinical interaction truth is unavailable.

Table 19: Mean clinical wall-clock seconds on a shared CPU. Prediction time includes subset evaluations, removals, and individual SHAP calculations used to select features. Coefficient time covers the final calculations through order three. Preparation includes loading and initial capacity evaluation.

Model	Prediction evaluations	Coefficient calculation	Preparation
Recursive	30.50	0.24	25.51
Progressive	20.03	0.13	267.77
Boosted trees	16.30	0.17	0.06
GRU-D	110.34	0.15	0.01
Transformer	297.07	0.15	0.02

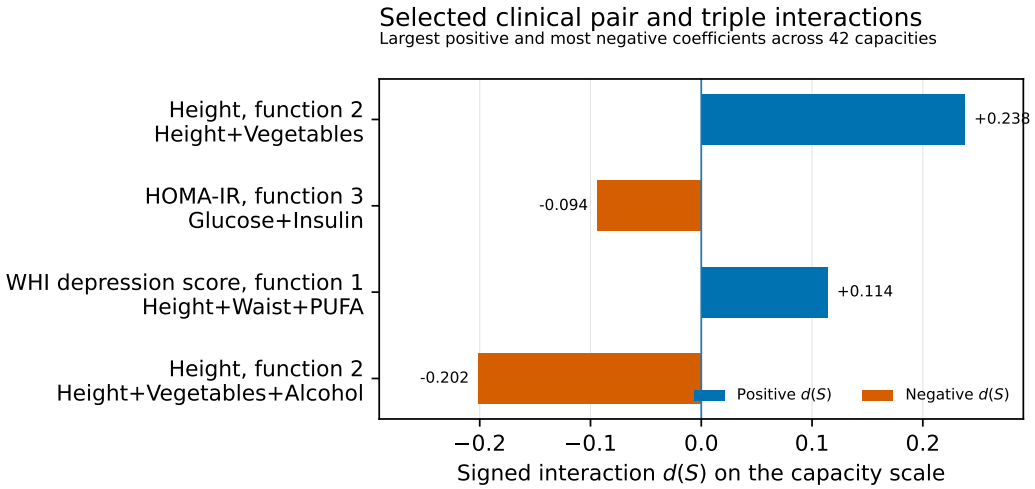


Figure 18: Feature pair and triple terms from the selected representative clinical fit. Selection is deterministic: global positive and negative extrema of the signed Möbius coefficient $d(S)$, separately for pairs and triples, over all 14 outcomes and three time functions. These values are on the capacity scale and are descriptive fitted associations, not changes in outcome units, uncertainty intervals, or causal effects.

E ADDITIONAL RECOVERY AND CLINICAL COMPARISONS

All analyses in this section are secondary. The original clinical test results were known before these extensions were designed. Model settings were chosen using validation participants.

E.1 INCOMPLETE PROBES AND ALTERNATIVE GENERATORS

The incomplete-probe study used the six-feature signed generator, two correlation/noise conditions, two capacity models, four probe designs, and three seeds (48 fits). Ordinary training, validation, and test histories were independently generated. Validation used ordinary histories only. The four training designs used no probes; 31 random proper subsets with eight repeats each; all 62 proper subsets with two repeats each; or only sets of size at most three with eight repeats each. Empty and full values are fixed by normalization.

Table 20: Incomplete-probe recovery under strong correlation and noise. Means over three seeds. MAE is over all 64 signed set interactions.

Model	Training probes	Noise-free RMSE	Interaction MAE
Recursive	None	.0014	.0217
Recursive	Random subsets	.0031	.0559
Recursive	Two per proper subset	.0042	.0382
Recursive	Sizes one to three	.0024	.0305
Progressive	None	.0020	.0536
Progressive	Random subsets	.0038	.0613
Progressive	Two per proper subset	.0041	.0407
Progressive	Sizes one to three	.0025	.0302

All planted pair signs were recovered in the strong condition. The triple sign was recovered in all no-probe, two-repeat, and size-limited fits, but only two of three random-subset fits per model. The complete-probe experiment also used probes for validation, so differences from that experiment reflect both training and selection. They are not an isolated estimate of the benefit of probe coverage.

A further 42 no-probe fits crossed two models, three seeds, and seven generators: the original baseline, a non-Choquet continuous response, unobserved early/late interaction regimes, two irrelevant inputs, noise SD 0.08 instead of 0.03, 2,000 instead of 20,000 training histories, and mixed pair/triple/four-way interactions. Recursive noise-free RMSE increased from 0.0014 at baseline to 0.0250 for the non-Choquet response and 0.0115 for changing regimes. Interaction MAE increased from 0.0216 to 0.0474 with smaller samples and 0.0409 with stronger noise. A non-Choquet response has no generating capacity; its values at binary inputs supply only an attribution reference. For changing regimes, the reference is their mean capacity.

E.2 ATTENTION COMPONENTS AND RESTRICTED CAPACITIES

These matched no-probe synthetic studies used three seeds, 20,000 training histories, 5,000 validation histories, and 20,000 test histories per seed. Selection minimized ordinary validation MSE. The attention comparison preserves recursive monotonicity in every variant. Removing the independent subset parameters produces the largest interaction error in this comparison.

Table 21: Synthetic attention component comparison. Overlap is the fraction of shared top-ten absolute non-singleton interactions across pairs of seeds.

Variant	RMSE	Interaction MAE	Parameters	Overlap
Full progressive	.0018	.0509	7,988	.433
Uniform attention weights	.0020	.0604	7,988	.633
No independent subset parameters	.0072	.1263	7,926	.467
Separate descriptors at each size	.0017	.0423	8,468	.467
Shared attention across sizes	.0022	.0446	2,591	.500

Table 22: Synthetic capacity restrictions on the same three-seed design.

Capacity	RMSE	Interaction MAE	Parameters	Overlap
Full recursive	.0014	.0217	62	.533
Sparse interaction penalty	.0012	.0023	62	.700
Hierarchical interaction penalty	.0063	.0080	62	.900
Rank-two submodular mixture	.0365	.0275	14	1.000
16 reference subsets	.0357	.2211	16	.933
Signed 2-additive	.0190	.0203	21	1.000
Signed 3-additive	.0172	.0191	41	.900

Sparse and hierarchical penalties retained the full recursive capacity and used a fixed penalty scale of 0.002. The generating capacity has few nonzero interactions, favoring sparsity here. The rank-two positive mixture of noisy-OR set functions is submodular; the 16-reference-subset model uses monotone maximum closure. The signed low-order models use sufficient monotonicity constraints that are more restrictive than the full signed k-additive class. These experiments have different expressiveness and parameter counts and do not establish clinical benefits.

E.3 Clinical selection stability

We fit the recursive model 30 times in total over ten imputations and seeds 1729, 2718, and 31415. Interactions through order five were calculated from capacities of those and smaller sizes. For each outcome, time function, and order, the ten largest absolute coefficients in imputation 1, seed 1729 define the reference sets. We compared these with the other 29 fits. Spearman correlations rank all absolute coefficients within an order. Correlation and top-ten overlap are first averaged over the 29 comparisons and then summarized by their median across the 42 capacities. Sign agreement and selection frequency are summarized by their median across the reference sets.

Table 23: Clinical interaction stability. Entries are medians across the corresponding outcome/function comparisons or reference sets. Selection frequency measures how often a reference set remains in the top ten.

Order	Sets per measure	Rank correlation	Top-ten overlap	Sign agreement	Selection frequency
2	210	.781	.669	1.000	.690
3	1,330	.592	.369	1.000	.310
4	5,985	.431	.203	.966	.138
5	20,349	.393	.150	1.000	.103

Median between-imputation SD was 0.0044 for pairs and 0.0065 for triples; corresponding mean within-imputation seed SD was 0.0109 and 0.0153.

A bootstrap of 15,286 validation participants held the fitted model fixed and ranked mean prediction contributions $d(S)\min_{j\in S} z_j$. Median top-ten selection frequency was 1.00 for both pairs and triples. This measures participant sampling for a fixed model, not uncertainty from retraining. Three full-model refits on independent 70% training-participant subsets gave pair/triple mean overlap 0.579/0.326 and rank correlation 0.636/0.478. The original training standardization was retained, including participants omitted from those subsets; the analysis isolates sensitivity to the data used for fitting coefficients.

The corresponding low-order refits gave mean pair overlap 0.730 for the 2-additive model and triple overlap 0.638 for the 3-additive model. Two 3-additive subset fits reached the 200-epoch limit with best epochs 197 and 199, making convergence less certain. These refits cover only imputation 1.

E.4 Paired prediction and low-order clinical models

The restricted-capacity ablations retain the main model's feature summaries, outcome dynamics, and likelihood, and set $d(T) = 0$ for $|T| > k$, with $k \in \{1, 2, 3\}$. In particular, the 2-additive model contains individual and pair terms and uses no attention. For $2 \le |T| \le k$, let $b_T = \tanh(u_T)$ and

define

$$a_j = \mathrm{softplus}(u_j) + \sum_{T \ni j} |b_T|, \qquad Z = \sum_j a_j + \sum_T b_T.$$

The normalized coefficients are $d(\{j\}) = a_j/Z$ and $d(T) = b_T/Z$. Before normalization, every marginal addition of feature j is at least $\mathrm{softplus}(u_j) > 0$; also $Z > 0$. Thus the capacity is normalized and monotone. This sufficient condition restricts the possible signed coefficients. Prediction evaluates $\sum_j d(\{j\})z_j + \sum_{2 \le |T| \le k} d(T) \min_{j \in T} z_j$ directly. Its restriction concerns internal capacity coefficients, which differ from the replacement-based prediction explanations in Appendix D.6.

The paired comparison averages squared errors over matched fits for the same 15,334 test participants, then resamples participants jointly for both models and all outcomes. Table 24 uses Bonferroni-adjusted intervals for four model comparisons. The intervals condition on the fitted models and do not include training uncertainty or constitute Rubin pooling of imputations.

Table 24: Secondary paired test differences relative to recursive Choquet. Negative values favor the comparison model. Progressive uses only matched imputation-1 fits.

Model	MSE difference	Simultaneous 95% interval
Boosted trees	$-.017201$	$[-.027740, -.007741]$
GRU-D	$-.013873$	$[-.022242, -.005842]$
Transformer	$-.005898$	$[-.014014, .001936]$
Progressive Choquet	$-.007253$	$[-.012649, -.002505]$

Table 25: Secondary imputation-1 clinical comparison, three seeds per model. Training duration is the mean; memory is maximum recorded GPU allocation.

Model	Validation MSE	Test MSE	Parameters	Train (s)	GPU (GiB)
Full recursive	.527952	.521481	88,080,789	3,676	3.64
1-additive	.518163	.517169	1,371	564	.26
2-additive	.514920	.514919	10,191	517	.27
3-additive	.515781	.515578	66,051	707	.46

The 2-additive and 3-additive paired MSE differences are -0.006562 and -0.005903, with adjusted 95% intervals $[-0.012456, -0.001072]$ and $[-0.012082, -0.000153]$. The 1-additive difference is -0.004312 with interval $[-0.009752, 0.000798]$. The four-comparison adjustment also includes the initial one-seed additive baseline, paired with the recursive model using the same seed. The sufficient monotonicity construction limits the signed low-order models; this is not a comparison of all possible k-additive parameterizations.

E.5 Prediction time, extraction time, and original-unit errors

Computational cost was measured for one fit per model (imputation 1, seed 1729) on the same 35,831 test histories. Models ran sequentially without competing study workloads. Neural models used a GPU with batches of 512; boosted trees used a CPU. Input preparation used eight CPU threads. Prediction time was measured after warmup and excludes input preparation, model loading, and evaluation. Pair and triple coefficient computation was measured after one warmup as the median of five repetitions across all 42 capacities. Complete models compute all capacity values, whereas lower-order models directly parameterize their nonzero coefficients.

Errors on the original outcome scales were evaluated for 132 clinical fits: 30 each for trees, GRU-D, Transformer, and recursive Choquet, plus three each for progressive and the three low-order models. Biomarker errors invert both training standardization and the biomarker log transform; depression errors invert standardization only. Mean absolute error for recursive Choquet is 1.763 points on the WHI depression scale and 5.406 points on the CARDIA scale, compared with 1.726 and 5.451 for boosted trees. These outcomes remain imputed; observed-only error cannot be calculated.

For each fit, 400 participant bootstrap resamples estimate uncertainty within horizons $(0, 2]$, $(2, 5]$, $(5, 10]$, and > 10 years, retaining all target pairs of each sampled participant. These intervals

Table 26: Prediction time, memory, and interaction computation for one fit per model. GPU memory includes model parameters and intermediate calculations; CPU memory includes input data. Coefficient computation applies only to capacity models.

Model	Prediction (s)	GPU (GiB)	CPU resident (GiB)	Extraction (s)
Boosted trees	.082	—	5.302	—
GRU-D	1.310	.040	5.302	—
Transformer	.345	.069	5.301	—
Recursive Choquet	4.543	1.577	5.302	.2705
Progressive Choquet	308.859	8.212	5.645	5.0055
1-additive	.203	.019	5.301	.0001
2-additive	.206	.019	5.302	.0002
3-additive	.209	.025	5.302	.0002

condition on the fitted model. Figure 9 shows mean errors across fits without uncertainty intervals. The longest-horizon intervals are generally widest.

E.6 Sensitivity to temporal summaries and input changes

Nine imputation-1 fits evaluated only on validation participants used the same 21 inputs and 2-additive model. Three seeds each evaluated last value and last value plus slope, variability, and recency; individual additions used one seed each. The softmax summary had validation MSE 0.514920, compared with 0.522768 for last value and 0.515045 for all additions. Mean top-ten pair overlap with softmax was 0.635 and 0.741, respectively. Pair rank correlations were 0.422 and 0.435. Similar error can therefore conceal changes in fitted interactions. The single-addition results are preliminary.

A separate synthetic evaluation perturbed the aggregated inputs of already fitted models without retraining. Feature-specific masking, value-dependent dropout, adding 0.4 then clipping to $[0, 1]$, and reducing history correlation from 0.85 to 0.2 increased recursive RMSE from 0.0014 to 0.0451, 0.0241, 0.3574, and 0.0037. Targets used the uncorrupted inputs except in the correlation-change condition. These perturbations do not reconstruct original missingness or measure transfer between real cohorts.

E.7 Comparison of interaction definitions

With the 42-fit generator study, we also calculated exact classical Shapley, Shapley-Taylor, Faith-Shap, and Möbius pair/triple indices. Independent-reference functional ANOVA, Sobol pair fractions, and Friedman-style H-statistics describe variance rather than signed capacity coefficients. For the recursive baseline, pair MAE was 0.0126 for Möbius coefficients, 0.0081 for order-two Shapley-Taylor, and 0.0091 for order-two Faith-Shap. Each error uses that method's own generating reference; the magnitudes do not rank methods on a common estimand. The non-Choquet response has no true generating capacity, and its binary-input reference can omit continuous interactions.

For the two depression outcomes, the representative model's first time function was evaluated over all 21 features: 210 pairs and 1,330 triples per capacity. Pairwise variance calculations used independent uniform inputs, seven-point quadrature for the focal pair, 128 background draws for other features, 20,000 draws for total variance, and three background repetitions. This reference distribution differs from observed clinical dependence. A separate six-feature comparison sets the other 15 inputs absent and therefore evaluates a different function. These descriptions concern one fitted capacity, not causal effects or stability across refits.