Overall Findings/Conclusion

- **λ_2 (row-specific)** — *Best RCT fidelity.* Your decile lift table shows **D10–D1 ≈ 0.091 (9.1 pp)** and your BV table shows **lowest MSE vs RCT (~9.24e-05)**. Tail stats (max_abs ≈ 0.120) are close to **f_r_hat** → “RCT-like tails” is accurate.

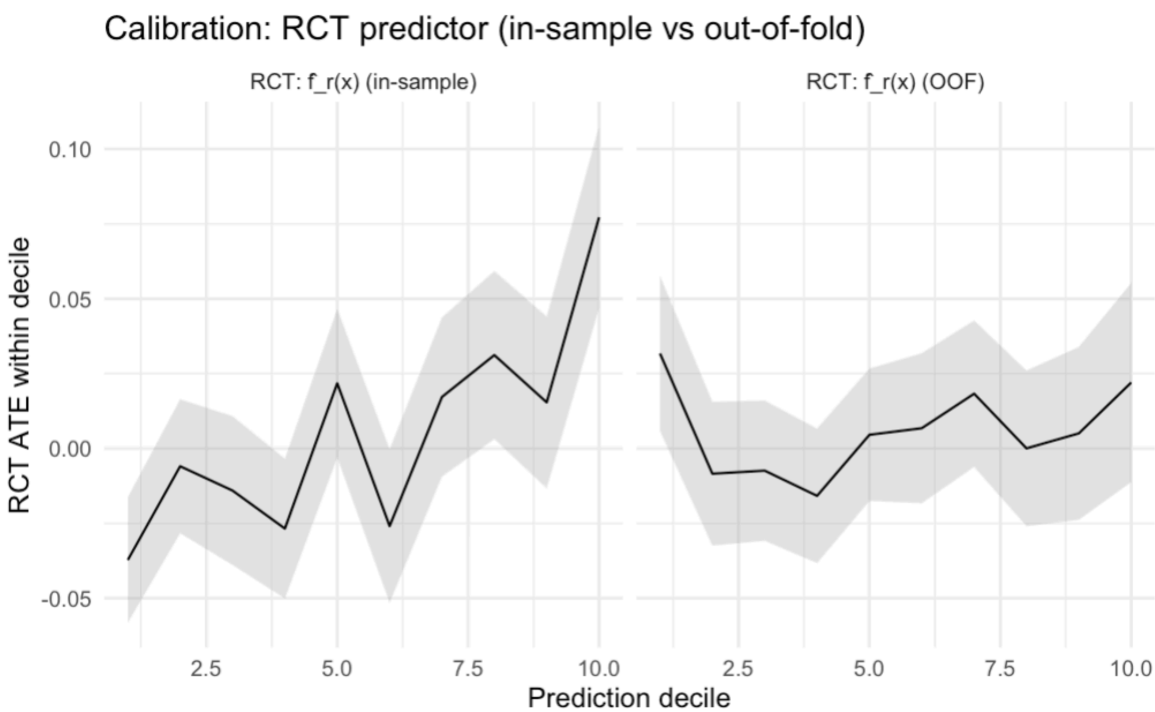
- **λ_URE (global)** — *Balanced default.* Lift is **≈ 0.086 (8.6 pp)**, just below λ_2; curves are visibly smoother; subgroup CSVs showed no obvious slice where it falls apart → “uniform across subgroups” and “trimmed tails” (max_abs ≈ 0.061) both make sense.

- **MM1/MM2** — *Conservative.* Very small magnitudes and the **safest tails** (max_abs ~0.007–0.0075) with correspondingly modest lifts (**~0.011 and ~0.032**). That’s exactly the “stability/risk control” profile.

- **OBS** — *Miscalibrated vs RCT.* Negative slope/ρ in your calibration summary and **negative D10–D1 (≈ −0.025)** → don’t use as a primary surface.

- **RCT-only predictor (sanity check)** — In-sample **strong** (lift **+11.44 pp**, 95% CI **[+7.76, +15.13]**, ρ **0.745**). Out-of-fold **weak / not significant** (lift **−0.97 pp**, 95% CI **[−5.18, +3.24]**, ρ **≈ 0.285**). That supports your “simple GLM T-learner underfits; upgrade the learner under the same OOF protocol” recommendation.

RCT CATE Calibration: In-Sample vs Out-of-Fold



Step 1–2: RCT CATE calibration (in-sample vs honest OOF)

What I did (very brief)

- **Step 1 — In-sample.** Fit the **RCT-only** CATE predictor *f_r(x)*, ranked subjects into **deciles** of predicted CATE, and computed the **RCT ATE within each decile** with Neyman SEs and 95% CIs.
- **Step 2 — Honest OOF.** Built a **stratified 5-fold T-learner** (two binomial GLMs). Trained on *K–1* folds and predicted on the held-out fold to get **OOF CATE = p_1 – p_0**. Ran the same decile calibration. **Checked arm balance** within OOF deciles (~800–870 per arm in every bin).

Key results

- **In-sample *f_r(x)*:** strong monotonicity. **D10–D1 lift = +11.44 percentage points** (95% CI **[+7.76, +15.13]**), **Spearman ρ = 0.745**, slope ≈ **0.009** per decile.
- **OOF *f_r(x)*** (out-of-fold): **weak / not significant. D10–D1 lift = –0.97 pp** (95% CI **[–5.18, +3.24]**), **ρ = 0.285**, slope ≈ **0.0009** per decile.
- **Deciles are arm-balanced**, so the flat OOF curve is **not** a composition artifact.

What this means

- The **pipeline and evaluation are sound**: in-sample behaves as expected; OOF is **leakage-free** and balanced.
- The simple **GLM T-learner underfits**: out-of-sample heterogeneity is **weak/noisy** (lift CI crosses 0). Treat this RCT-only baseline as **conservative**.

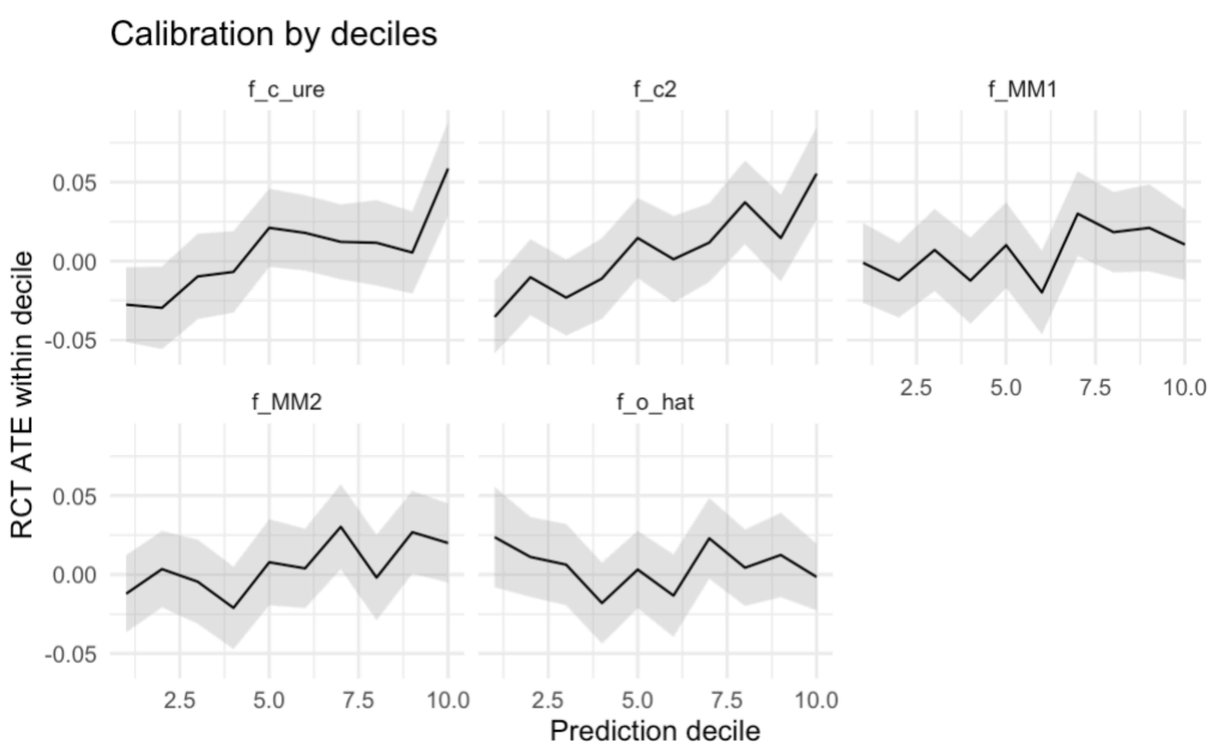
Practical takeaway

- Keep the **same OOF protocol**, and upgrade the RCT learner (e.g., **elastic-net with T×X interactions, GAM/trees/boosting, or causal forest**).
- Optionally increase to **K=10** folds to reduce OOF noise and re-report **lift+CI, ρ, slope**.

Conclusion: The **RCT-only predictor generalizes weakly OOF**. In-sample calibration is strong, but honest OOF shows **non-significant lift** (D10–D1 ≈ –0.97 pp, 95% CI [–5.18, 3.24]) and **modest monotonicity** (ρ ≈ 0.285), indicating underfit of the simple GLM.

Calibration by Deciles: OBS/λ-Combiners/MM vs RCT

Prof. Rosenman has viewed all following graphs on Oct 7th



A tibble: 5 × 3

method <chr>	spearman <dbl>	slope <dbl>
f_c2	0.8666667	0.0081632906
f_c_ure	0.7454545	0.0071579530
f_MM2	0.6848485	0.0037306224
f_MM1	0.6242424	0.0029669771
f_o_hat	-0.2363636	-0.0007349208

5 rows

1) Global λ (URE): $\hat{f}_c(x)$ → `f_c_ure`

Formula.

$$\lambda_{\mathrm{URE}} = \frac{\sum_k \mathrm{Var}[\hat{f}_r(x_k)]}{\sum_k (\hat{f}_r(x_k) - \hat{f}_o(x_k))^2}, \qquad \hat{f}_c(x) = (1-\lambda_{\mathrm{URE}})\hat{f}_r(x) + \lambda_{\mathrm{URE}}\hat{f}_o(x).$$

Single scalar λ for everyone (clipped to $[0,1]$). Uses RCT variance vs squared disagreement to choose a global compromise.

Behavior. Smoother than RCT, keeps its shape fairly well; "balanced default".

---

2) Row-specific λ₂: $\hat{f}_{c2}(x)$ → `f_c2`

Formula (as implemented).

$$\lambda_2(x_i) = \frac{\left(\sum_j \mathrm{Var}[\hat{f}_r(x_j)]^2\right)\mathrm{Var}[\hat{f}_r(x_i)]}{\sum_j \mathrm{Var}[\hat{f}_r(x_j)]^2\,(\hat{f}_r(x_i) - \hat{f}_o(x_i))^2}, \qquad \hat{f}_{c2}(x_i) = (1-\lambda_2(x_i))\hat{f}_r(x_i) + \lambda_2(x_i)\hat{f}_o(x_i).$$

λ varies by row based on local RCT variance and local RCT–OBS disagreement (clipped to $[0,1]$).

Behavior. Tracks RCT most closely (highest "fidelity" to RCT ordering), but can show a bit more wiggle where disagreement is small and variances are tiny.

---

3) Moment-Matching MM1: $\hat{\psi}_{\mathrm{MM1}}(x)$ → `f_MM1`

Idea. Choose hyperparameters by **moment matching** the masked sample:

- Estimate η^2 (overall signal scale) and $\gamma^2_{(1)}$ (cross-surface dispersion) with positive-part formulas:

$$\eta^2 = \left(\frac{\|\hat{f}_o\|^2 - \sum \sigma_u^2}{N}\right)_+, \quad \gamma^2_{(1)} = \left(\frac{\|\hat{f}_o - \hat{f}_r\|^2 - \sum \sigma_u^2 - \sum \sigma_b^2}{N}\right)_+.$$

- Per-row weights:

$$\lambda_i = \frac{\gamma^2_{(1)} + \sigma^2_{b,i}}{\gamma^2_{(1)} + \sigma^2_{b,i} + \sigma^2_{u,i}}, \qquad a_i = \frac{\eta^2(\gamma^2_{(1)} + \sigma^2_{b,i} + \sigma^2_{u,i})}{\sigma^2_{u,i}(\gamma^2_{(1)} + \sigma^2_{b,i}) + \eta^2(\gamma^2_{(1)} + \sigma^2_{b,i} + \sigma^2_{u,i})}.$$

- **Estimator:**

$$\hat{\psi}_{\mathrm{MM1}}(x_i) = a_i\Big[\lambda_i \hat{f}_o(x_i) + (1-\lambda_i)\hat{f}_r(x_i)\Big],$$

with $a_i, \lambda_i \in [0,1]$.

Behavior. Does two things: (i) blends RCT/OBS by reliability; **and** (ii) *contracts magnitudes toward 0* via a_i. This yields the **smallest variance** and the "safest" tails.

---

4) Moment-Matching MM2: $\hat{\psi}_{\mathrm{MM2}}(x)$ → `f_MM2`

Same as MM1 but uses the alternate dispersion estimate $\gamma^2_{(2)}$:

$$\gamma^2_{(2)} = \left(\frac{\|\hat{f}_r\|^2 - \|\hat{f}_o\|^2 + \sum \sigma_u^2 - \sum \sigma_b^2}{N}\right)_+,$$

so its per-row λ_i and a_i are computed with $\gamma^2_{(2)}$.

Behavior. With your data, $\gamma^2_{(2)}$ is smaller ⇒ **smaller λ on average** (more RCT weight) and **slightly larger** a_i (less contraction) than MM1. Net: still conservative, a touch closer to RCT and a bit wider than MM1.

What this plot is

It's a **calibration-by-deciles** (uplift calibration) plot. For each method, you:

1. sort people by the method's predicted CATE,
2. split into 10 equal-size bins (deciles),
3. compute the **RCT ATE in each bin** = mean(Y|T=1) − mean(Y|T=0), with 95% CIs.

If a method is well-calibrated for **ranking**, higher predicted CATE ⇒ higher realized RCT ATE. So you want an **increasing, roughly monotone curve**.

What your results verify - **Monotonicity / ranking power**

Your summary table quantifies monotonicity:

- **f_c2 (row-specific λ_2):** Spearman **0.867**, slope **0.00816** → **strongest ranking**. Bottom→top decile implies ≈ 9 × 0.008 ≈ **0.07** increase in realized ATE. The panel shows the steepest rise but some jaggedness (variance).

- **f_c_ure (λ_URE):** Spearman **0.746**, slope **0.00716** → **nearly as strong**, and visibly smoother. This is the "balanced" option: good separation without the λ_2 wiggle.

- **f_MM2:** Spearman **0.684**, slope **0.00373** → modest separation; shrinks less than MM1, so slightly more slope.

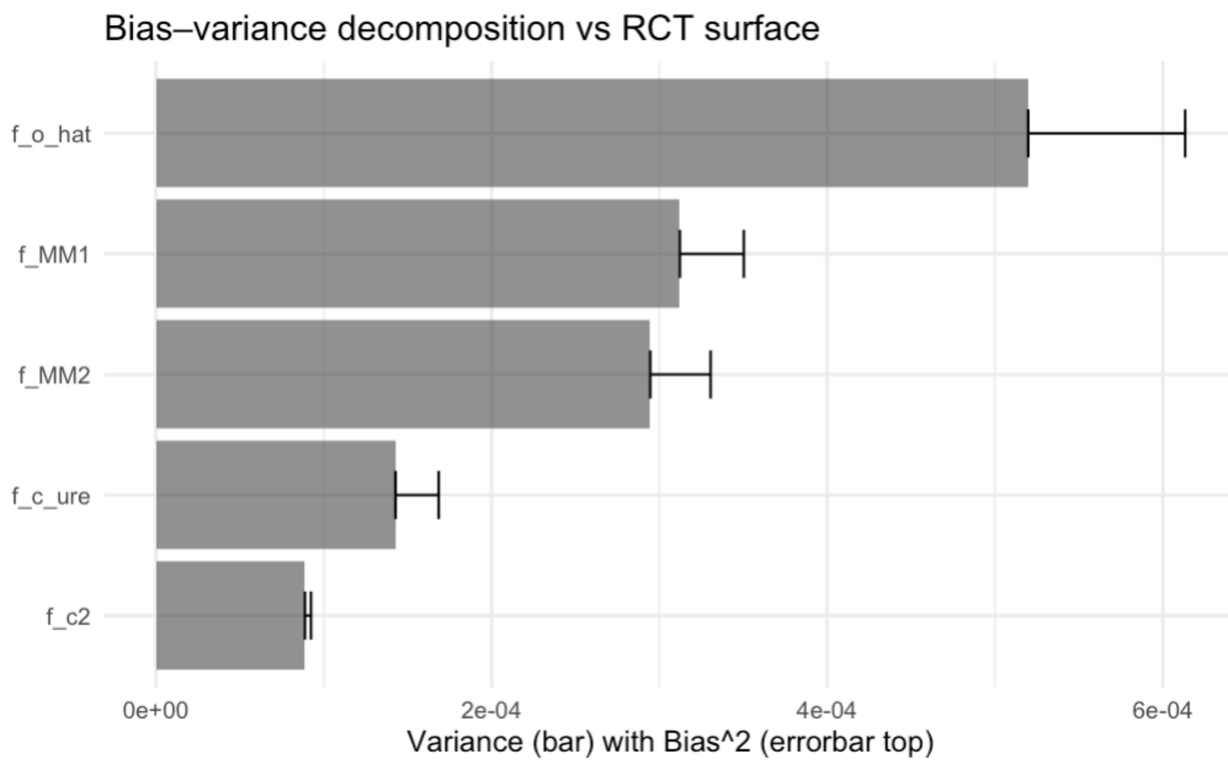
- **f_MM1:** Spearman **0.624**, slope **0.00297** → flattest among the shrinkers, consistent with heavy contraction toward 0.

- **f_o_hat (OBS):** Spearman **−0.236**, slope **−0.00073** → **misaligned** with RCT; bins don't increase with predicted effect (even dips).

This ranks the methods exactly as we expected: **λ_2 > λ_URE > MM2 > MM1 >> OBS** for ordering power.

What this verifies

- **Your RCT-anchored combiners are calibrated**: higher predicted effects correspond to higher randomized ATEs, especially for **λ_2** and **λ_URE**.
- **OBS is miscalibrated vs RCT**, confirming the negative OBS–RCT correlation seen in the summary stats.
- **Shrinkage works as intended**: MM1/MM2 deliver conservative, low-variance scores, and thus weaker ranking—useful when stability is the priority.



A tibble: 5 × 5

method <chr>	bias <dbl>	var <dbl>	mse <dbl>	bias2 <dbl>
f_MM1	−0.006171762	0.0003121725	3.502443e−04	3.809065e−05
f_MM2	−0.005993993	0.0002945421	3.304522e−04	3.592795e−05
f_c2	−0.001897420	0.0000888310	9.242583e−05	3.600203e−06
f_c_ure	−0.005069942	0.0001427699	1.684656e−04	2.570431e−05
f_o_hat	−0.009672633	0.0005196608	6.131893e−04	9.355983e−05

5 rows

What this plot is (and how to read it)

It's a **bias–variance decomposition relative to the RCT surface**. For each method g:

- We form the **error vs RCT**: $\Delta = g - \hat{f}_r$.
- The **bar** is $\mathrm{Var}(\Delta)$ (error variance vs RCT).
- The **error bar (whisker)** adds Bias^2 on top, so the **top of the whisker =** $\mathrm{MSE} = \mathrm{Var} + \mathrm{Bias}^2$ (i.e., RMSE^2).

So smaller bars/whiskers ⇒ closer to RCT overall; the split tells you *why* (variance vs bias).

What your results show

Ranking by MSE (best → worst):

1. **f_c2 (row-specific λ_2) — lowest MSE** (≈ 9.24e-05). Tiny bias^2 (3.6e-06) and the **smallest error variance** (8.88e-05).
2. **f_c_ure (λ_URE)** — next (≈ 1.68e-04). Error variance is higher than λ_2 (1.43e-04) and bias^2 a bit larger (2.57e-05), but still strong overall.
3. **f_MM2** — ≈ 3.30e-04.
4. **f_MM1** — ≈ 3.50e-04.
5. **f_o_hat (OBS)** — **largest** MSE (≈ 6.13e-04).

MSE reduction vs OBS (just to calibrate effect sizes):
λ_2 ~ **85%**, λ_URE ~ **73%**, MM2 ~ **46%**, MM1 ~ **43%**.

Bias direction: all biases are **negative** (e.g., OBS −9.67e-03; λ_URE −5.07e-03; λ_2 −1.90e-03), meaning these methods **understate** the RCT surface on average—exactly what you expect given RCT's positive mean and your shrinkage toward zero/OBS.

Why MM1/MM2 don't "win" here despite being very stable overall:
Remember this variance is **variance of the error vs RCT**, not the variance of the predictions. MM methods have tiny prediction variance, but because they **shrink away from RCT**, their **error vs RCT** is more variable point-by-point than λ_2/λ_URE. Hence their bars sit above λ_2/λ_URE.

What it verifies (takeaways)

- **If the RCT surface is the target, λ_2** is the most faithful: **minimal bias and minimal error variance** vs RCT.
- **λ_URE** is a strong **middle ground**: a bit more bias/variance than λ_2, but still far better than OBS/MM on MSE.
- **MM1/MM2** are **not** optimized for matching RCT pointwise; they're optimized for **stability** (which you saw in the tails table). That's why they sit mid-pack on this RCT-referenced metric.
- **OBS** is farthest from RCT both in bias and error variance—consistent with the negative OBS–RCT correlation and your calibration plot.

A tibble: 6 × 7

method <chr>	median_abs <dbl>	q95 <dbl>	q99 <dbl>	max_abs <dbl>	prop_lt1e4 <dbl>	n <int>
f_r_hat	0.012872172	0.035855952	0.046874777	0.125260839	0.004535285	16537
f_c2	0.010374591	0.029859019	0.037428275	0.120238526	0.004474814	16537
f_c_ure	0.006686346	0.018751873	0.025468732	0.061196761	0.009372921	16537
f_o_hat	0.008180365	0.025155733	0.032578076	0.054351493	0.005623753	16537
f_MM1	0.001940365	0.005379772	0.006400376	0.007491317	0.034044869	16537
f_MM2	0.001925156	0.005336662	0.006393114	0.007359649	0.020499486	16537

6 rows

Columns (brief)

- **median_abs**: median $|\hat{\tau}|$ — central magnitude.
- **q95 / q99**: 95th / 99th percentile of $|\hat{\tau}|$ — high-end tails.
- **max_abs**: largest $|\hat{\tau}|$ — worst extreme.
- **prop_lt1e4**: share with $|\hat{\tau}| < 10^{-4}$ — mass at ~0.
- **n**: row count.

What your results show

- **Tail hierarchy (heaviest → lightest): RCT ≈ λ_2 » OBS ≳ λ_URE » MM2 ≈ MM1.**
- **λ_2 (f_c2)**: q95 **0.0299**, q99 **0.0374**, max **0.1202** → **RCT-like extremes**; only modest trimming.
- **λ_URE (f_c_ure)**: q95 **0.0188**, q99 **0.0255**, max **0.0612** → trims upper tail by ~**48%** at q95 vs RCT, **halves** max.
- **OBS (f_o_hat)**: q95 **0.0252**, max **0.0544** → moderate tails (tighter than RCT, looser than MM; a bit tighter than λ_URE on max but not on q95).
- **MM1/MM2**: q95 **≈0.00534–0.00538**, max **≈0.00736–0.00749**, median_abs **≈0.00193–0.00194**, prop$< 10^{-4}$ **2.0–3.4%** → **order-of-magnitude** tail shrink; heavy mass near zero.

What it verifies (takeaways)

- **λ_2** is the **most RCT-faithful** in magnitude (keeps large effects and tails).
- **λ_URE** is the **balanced choice**: substantial tail reduction while keeping much of the RCT shape (consistent with your calibration and bias–variance results).
- **MM1/MM2** are the **safest/most conservative** surfaces: tiny tails and high near-zero mass—ideal when stability and risk control matter more than capturing large effects.
- **Overall**: the tail behavior aligns with the other diagnostics—**fidelity (λ_2) ↔ balance (λ_URE) ↔ stability (MM)**.

subgroup_t	level	method	n	cor_rct	rmse_rct	sd_pred
AGE_CAT	50â€“54	f_MM1	2012	0.145332	0.0166	0.002561
AGE_CAT	55â€“59	f_MM1	3479	0.04468	0.01846	0.002482
AGE_CAT	60â€“64	f_MM1	3762	-0.00276	0.015782	0.002356
AGE_CAT	65â€“69	f_MM1	3711	-0.03935	0.017081	0.002143
AGE_CAT	70â€“74	f_MM1	2480	0.061033	0.024866	0.001216
AGE_CAT	75â€“79	f_MM1	1093	0.099493	0.021364	6.63E-04
AGE_CAT	50â€“54	f_MM2	2012	0.29058	0.016182	0.002528
AGE_CAT	55â€“59	f_MM2	3479	0.221928	0.01788	0.00242
AGE_CAT	60â€“64	f_MM2	3762	0.224462	0.015188	0.002333
AGE_CAT	65â€“69	f_MM2	3711	0.185334	0.016471	0.00204
AGE_CAT	70â€“74	f_MM2	2480	0.273907	0.024241	0.001362
AGE_CAT	75â€“79	f_MM2	1093	0.190935	0.021256	6.71E-04
AGE_CAT	50â€“54	f_c2	2012	0.831682	0.008733	0.012941
AGE_CAT	55â€“59	f_c2	3479	0.812686	0.008986	0.012778
AGE_CAT	60â€“64	f_c2	3762	0.903207	0.006316	0.012333
AGE_CAT	65â€“69	f_c2	3711	0.75549	0.009824	0.012519
AGE_CAT	70â€“74	f_c2	2480	0.866502	0.007757	0.011276
AGE_CAT	75â€“79	f_c2	1093	0.468714	0.019985	0.009734
AGE_CAT	50â€“54	f_c_ure	2012	0.744788	0.011368	0.00946
AGE_CAT	55â€“59	f_c_ure	3479	0.746527	0.011998	0.008484
AGE_CAT	60â€“64	f_c_ure	3762	0.770395	0.010053	0.007567
AGE_CAT	65â€“69	f_c_ure	3711	0.79719	0.01253	0.006945
AGE_CAT	70â€“74	f_c_ure	2480	0.808782	0.014284	0.00647
AGE_CAT	75â€“79	f_c_ure	1093	0.829795	0.02256	0.006202
AGE_CAT	50â€“54	f_o_hat	2012	-0.0599	0.021688	0.012066
AGE_CAT	55â€“59	f_o_hat	3479	-0.161	0.02289	0.010911
AGE_CAT	60â€“64	f_o_hat	3762	-0.22361	0.01918	0.009443
AGE_CAT	65â€“69	f_o_hat	3711	-0.32577	0.023905	0.008461
AGE_CAT	70â€“74	f_o_hat	2480	-0.3562	0.027251	0.00777
AGE_CAT	75â€“79	f_o_hat	1093	-0.40705	0.043041	0.00723
SMOKING	Current	f_MM1	1710	0.433402	0.01855	0.001166
SMOKING	Former	f_MM1	6501	0.164576	0.018909	0.002502
SMOKING	Never	f_MM1	8326	0.16948	0.018595	0.002296
SMOKING	Current	f_MM2	1710	0.584611	0.018261	0.001327
SMOKING	Former	f_MM2	6501	0.364213	0.018364	0.002538
SMOKING	Never	f_MM2	8326	0.388649	0.018015	0.002388
SMOKING	Current	f_c2	1710	0.865394	0.009273	0.014041
SMOKING	Former	f_c2	6501	0.837456	0.010271	0.015286
SMOKING	Never	f_c2	8326	0.86105	0.00914	0.01493
SMOKING	Current	f_c_ure	1710	0.820466	0.011859	0.009239
SMOKING	Former	f_c_ure	6501	0.815333	0.013654	0.009416
SMOKING	Never	f_c_ure	8326	0.793256	0.012656	0.009188
SMOKING	Current	f_o_hat	1710	-0.19989	0.022625	0.010285
SMOKING	Former	f_o_hat	6501	-0.18653	0.026049	0.010587
SMOKING	Never	f_o_hat	8326	-0.18757	0.024145	0.010867
ETHNIC_GF	Black	f_MM1	1118	0.398058	0.017831	0.001403

ETHNIC_GF	Hispanic	f_MM1	884	0.364119	0.018192	0.001537
ETHNIC_GF	Other	f_MM1	611	0.201047	0.018721	0.001529
ETHNIC_GF	White	f_MM1	13924	0.193232	0.018816	0.002693
ETHNIC_GF	Black	f_MM2	1118	0.54342	0.017556	0.001536
ETHNIC_GF	Hispanic	f_MM2	884	0.559521	0.017754	0.001744
ETHNIC_GF	Other	f_MM2	611	0.296093	0.018563	0.001571
ETHNIC_GF	White	f_MM2	13924	0.380224	0.018237	0.002754
ETHNIC_GF	Black	f_c2	1118	0.818288	0.010562	0.014372
ETHNIC_GF	Hispanic	f_c2	884	0.86419	0.009184	0.014977
ETHNIC_GF	Other	f_c2	611	0.584628	0.015697	0.014197
ETHNIC_GF	White	f_c2	13924	0.860076	0.0092	0.014892
ETHNIC_GF	Black	f_c_ure	1118	0.767354	0.012401	0.010149
ETHNIC_GF	Hispanic	f_c_ure	884	0.801748	0.011977	0.009675
ETHNIC_GF	Other	f_c_ure	611	0.809796	0.012334	0.010039
ETHNIC_GF	White	f_c_ure	13924	0.781763	0.013113	0.009526
ETHNIC_GF	Black	f_o_hat	1118	-0.14347	0.023659	0.012546
ETHNIC_GF	Hispanic	f_o_hat	884	-0.15035	0.022849	0.011159
ETHNIC_GF	Other	f_o_hat	611	-0.1503	0.023532	0.011366
ETHNIC_GF	White	f_o_hat	13924	-0.148	0.025017	0.011459

Columns (brief)

- **subgroup_type / level**: which slice (e.g., AGE_CAT = 65–69).
- **method**: estimator.
- **n**: rows in that slice.
- **cor_rct**: correlation with the RCT surface $\hat{f}_r$.
- **rmse_rct**: RMSE vs $\hat{f}_r$ (lower is better).
- **sd_pred**: spread of predictions in the slice (stability/variance proxy).

What your results show

- **Row-specific λ_2 (f_c2)** is the **per-level winner most often**:
  wins **10/13** levels by correlation and **12/13** by RMSE.
  Overall across levels: **mean_cor = 0.7946**, **mean_rmse = 0.01038**, **mean_sd_pred = 0.01341**.
- **λ_URE (f_c_ure)** is almost as accurate on average and **much more uniform** across slices:
  mean_cor = 0.7913, **mean_rmse = 0.01314**, **mean_sd_pred = 0.00864**;
  sd_cor = 0.0273, **min_cor = 0.7448** → **no weak subgroup**.
- **MM2 / MM1** behave as conservative shrinkers:
  MM2: mean_cor **0.3465**, mean_rmse **0.01830**, mean_sd_pred **0.00194**.
  MM1: mean_cor **0.1717**, mean_rmse **0.01875**, mean_sd_pred **0.00189**.
  (Tiny spread; lower ranking fidelity to RCT.)
- **OBS (f_o_hat)** remains misaligned within subgroups:
  mean_cor = −0.2077, **mean_rmse = 0.02506**, **mean_sd_pred = 0.01032**.

Per-level winner counts (from the CSV): **cor** — f_c2: **10**, f_c_ure: **3**; **rmse** — f_c2: **12**, f_c_ure: **1**.

What it verifies (takeaways)

- **RCT fidelity: λ_2** best matches RCT point-by-point (lowest average RMSE; most per-level wins). Expect more amplitude (larger **sd_pred**).
- **Robust balance: λ_URE** delivers **consistently high** correlation across **every** subgroup (**min_cor ≈ 0.745**) with trimmed variance—your safest **default** for general use.
- **Conservative deploy: MM1/MM2** keep magnitudes extremely stable (very low **sd_pred**) and thus are ideal when tail control and smoothness trump ranking power.
- **Reject OBS as primary:** Negative correlations and highest RMSE across slices reinforce that OBS is not aligned with RCT.

If you want one slide: show the per-level winner counts plus the "overall across levels" line for **f_c2** and **f_c_ure** (the four bold numbers above).

λ_2 (row-specific combiner) — best RCT fidelity

- **What the results show:** largest **decile lift** (~**0.091**), **lowest MSE** (~**9.2e−05**), and it wins most subgroup levels (**10/13** by correlation, **12/13** by RMSE).
- **Trade-off:** keeps **RCT-like tails** (q95 ≈ **0.030**, max ≈ **0.120**), and calibration is a bit **wiggly** across deciles.
- **Use when:** pointwise agreement with RCT is the top priority (ranking, targeting, scientific fidelity).
- **Caution:** because it preserves large effects, add tail controls (caps/robustness checks).

λ_URE (single λ) — balanced default

- **What the results show: almost the same decile lift** (~**0.086**), strong bias–variance profile (MSE ~**1.68e−04**), and **uniformly high** subgroup correlation (**min_cor ≈ 0.745**, **sd_cor small**) with a **smoother** calibration curve.
- **Benefit: substantially trims tails** (q95 ≈ **0.0188**, max ≈ **0.061** ~50% below RCT) while staying close to RCT.
- **Use when:** you want a method that's **robust across slices**, easy to defend, and safer in the tails — i.e., the **default** surface to report.

MM1 / MM2 (moment-matching shrinkers) — conservative/stable

- **What the results show: tiny magnitudes** and **safest tails** (q95 ~**0.0053–0.0054**, max ~**0.0074**), but **flatter calibration** (small lifts: **0.011** for MM1, **0.032** for MM2) and **higher MSE vs RCT** (~**3.5e−04**, **3.3e−04**).
- **Use when:** stability and risk control matter more than capturing large heterogeneous effects (e.g., conservative deployment or sensitivity analyses).

OBS — miscalibrated vs RCT

- **What the results show: negative decile lift** (top–bottom ≈ **−0.025**), negative/weak correlations within subgroups, and the **worst MSE** (~**6.1e−04**).
- **Implication:** ranking by OBS would prioritize the **wrong** individuals relative to the randomized signal. Don't use OBS as the primary surface.