

EXPLORATORY — DOES NOT ALTER T-002'S REGISTERED EVIDENCE CATEGORY

Exploratory T-002 Follow-up Report

Prepared 24 September 2026. Diagnostic and successor-test design only. Original registration: <https://osf.io/gafmy/> ; DOI: 10.17605/OSF.IO/GAFMY. The authoritative original specification and 23 September OSF update remain unchanged.

1. What this follow-up establishes

T-002 remains Invalid / Uninterpretable. Its 543 eligible primary landmarks and 53 distinct qualifying onsets failed the mandatory floors of 600 and 80. All model results below were produced only after that verdict was frozen, with the user's explicit exploratory authorization. Numerical estimability here does not mean registered feasibility, and no evidence-category decision tree is reapplied.

The exploratory institutional model improves pooled grouped-CV log loss by 7.26%, with a high-Mass Direction contrast of -39.35 percentage points. However, the 95% full-refit bootstrap interval for improvement runs from -52.68% to +23.14%. Forward temporal log loss worsens by 209.56%, and its pooled contrast reverses to +4.74 points. These results show an in-sample-era cross-polity pattern that does not demonstrate reliable prospective transport. They provide no confirmatory support, causal conclusion, or revised verdict for PrF.

The largest sequential observation losses are material-capability/history coverage (814 landmarks) and incomplete follow-up without a qualifying onset (714). A shorter horizon or annual sampling increases rows but does not increase the 53 represented distinct onsets. The main successor-design problem is information and follow-up, not simply the number of rows.

2. Scope, implementation and interpretation

MMP is the frozen five-year CINC-based Material Mass Proxy, not ontological PrF Mass. IDP is the Institutional Direction Proxy, not ontological PrF Direction. Its frozen formula is $IDP = ((2L_{recent} - 1) + (L_{recent} - L_{prior})) / 2$: present institutional orientation conditioned by recent trajectory. O* and E* retain that same architecture. Existing predictor transformations, outcome construction and entity lineages were reused; no component weights or predictors were tuned.

The baseline is unpenalized maximum-likelihood logistic regression with MMP and centered linear calendar time. The augmented model adds IDP and $MMP \times IDP$; diagnostic models substitute O* or E*. Training-only means and sample standard deviations are used. Calendar time is internally divided by 100 solely for numerical conditioning. Both models use identical rows in each comparison. Separation, rank deficiency and convergence failures are reported without penalized fallback.

One five-fold polity partition was fixed before exploratory fitting: stratification by whether a polity has any positive primary landmark, shuffled with seed 2002. No seed search was performed. The same assignment is reused across comparisons. Grouped log loss gives each held-out landmark equal weight; high-Mass contrasts average within polity and then equally across polities. Temporal and regional pooling follow the frozen respective equal-origin and equal-region rules. Numerical feasibility floors are waived only for these exploratory fits.

Define $D = \text{LogLoss}_{\text{model}} - \text{LogLoss}_{\text{baseline}}$; negative D is favorable. Define $R = (\text{LogLoss}_{\text{baseline}} - \text{LogLoss}_{\text{model}}) / \text{LogLoss}_{\text{baseline}}$; positive R is favorable. The contrast ΔP compares $X = +1$ versus -1 training-set SD at $MMP = +1$ SD, using each held-out year and recomputing the interaction. Negative ΔP is the predicted Direction sign. A probability contrast is model-based, not a causal intervention; joint predictor support at imposed values was not separately established.

The analysis scope was recorded before exploratory fitting, after outcome exposure. It is not a new preregistration. One additional model comparison was fixed then: calendar-mature 50-year landmarks with $t + 50 \leq 2018$. Annual-anchor feasibility counts were added after initial results and are explicitly descriptive; no annual-anchor predictive models were fitted. All attempted specifications are reported, regardless of sign.

3. Grouped cross-validation and uncertainty

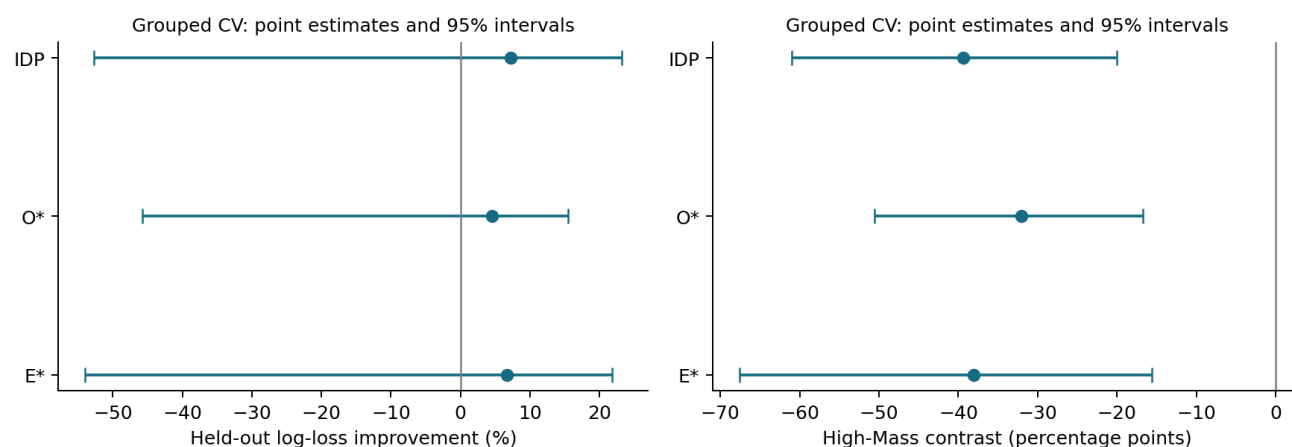
Model	Baseline LL	Model LL	D	R	ΔP (pp)
IDP	0.5112	0.4740	-0.0371	+7.26%	-39.35
O*	0.5112	0.4878	-0.0234	+4.58%	-32.04
E*	0.5112	0.4769	-0.0343	+6.71%	-38.04

All three comparisons use 543 landmarks, 98 polity lineages, 53 distinct onsets and 46 event-bearing polities. Equal-polity rather than equal-landmark loss pooling gives IDP improvement of 8.13%; this descriptive aggregation check does not replace the specified primary pooling.

Model	95% interval: R (%)	95% interval: ΔP (pp)	95% interval: D
IDP	[-52.68, +23.14]	[-61.00, -19.98]	[-0.1396, +0.2868]
O*	[-45.74, +15.42]	[-50.61, -16.73]	[-0.0885, +0.2305]
E*	[-53.99, +21.81]	[-67.62, -15.64]	[-0.1286, +0.3062]

All 2,000 paired full-refit bootstrap attempts succeeded for each model (seed 22002). Polities were resampled with replacement within their original folds; all their landmarks and multiplicities traveled together. Every replicate recalculated training transformations and refitted both models. Intervals are the 2.5th and 97.5th percentiles. They condition on this inherited sample, operationalisation and fold partition; they do not resolve selection bias, historical transport or measurement validity. No temporal or robustness-specific intervals are asserted.

EXPLORATORY — DOES NOT ALTER T-002'S REGISTERED EVIDENCE CATEGORY



Grouped CV uncertainty

Fold	Landmarks	Polities	Onsets	IDP R	IDP ΔP (pp)
1	115	20	10	+19.31%	-36.13
2	83	20	11	+9.19%	-38.88
3	112	20	10	+10.57%	-41.00
4	109	19	12	+18.48%	-34.16
5	124	19	10	-17.22%	-46.70

All five fits are numerically estimable. Four folds improve log loss; the fifth worsens it by 17.22%, despite a negative contrast. Fold onsets are distinct within folds and sum to 53. The exploratory partition satisfies the 10-onset-per-fold numerical floor but cannot repair the failed overall landmark/event floors.

Secondary metric	MMP baseline	IDP model
Brier score	0.1623	0.1443
AUROC	0.7650	0.8133
Average precision	0.7017	0.7084

Average precision is reported under its own name, not as trapezoidal precision-recall AUC. Calibration intercept/slope and additional secondary metrics are not claimed here. The diagnostic focus remains held-out log loss and the registered contrast, rather than selecting a favorable metric.

4. Forward temporal validation and construct discrimination

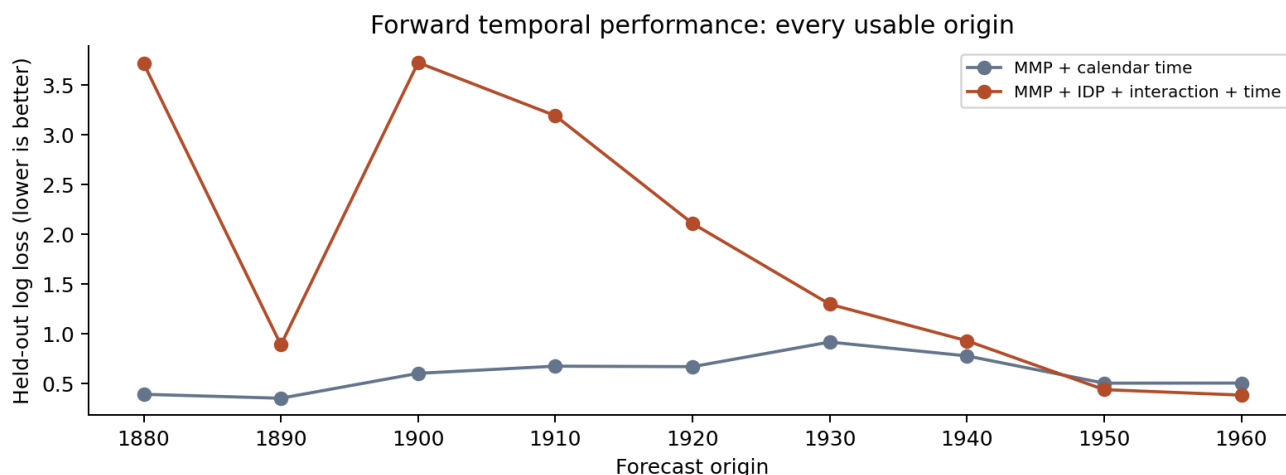
Training at origin C includes only landmarks s with $s + 50 \leq C$. Nominal calendar-mature test origins end in 1960 because $C + 50 \leq 2018$. Every candidate origin from 1820 through 1960 was attempted. Origins 1820–1860 have no training rows; 1870 has only 12 training rows and rank-deficient calendar adjustment. The nine usable primary origins are 1880–1960.

Model	Origins	Test rows	Baseline LL	Model LL	R	ΔP (pp)
IDP	9	367	0.5985	1.8527	-209.56%	+4.74
O*	9	367	0.5985	1.3464	-124.96%	+32.60
E*	Incomplete	—	—	—	Unavailable	Unavailable

The pooled temporal set contains 75 polities, 30 distinct onsets and 24 event-bearing polities. These exploratory counts reach the registered temporal numerical floors of three usable origins and 30 onsets; they do not cure primary invalidity. Loss is averaged equally across polities within each origin and then equally across origins. Relative improvement is calculated from those pooled losses, not by averaging origin-specific percentages. There is one landmark per polity at each origin.

Origin	Train n	Test n	Baseline LL	IDP LL	IDP R	IDP ΔP (pp)
1880	23	28	0.3910	3.7148	-850.08%	+0.07
1890	38	30	0.3514	0.8879	-152.69%	+6.73
1900	57	26	0.6016	3.7255	-519.29%	+10.68
1910	79	27	0.6738	3.1915	-373.63%	+9.37
1920	105	30	0.6688	2.1083	-215.25%	+31.53
1930	133	49	0.9163	1.2966	-41.51%	+6.47
1940	163	49	0.7768	0.9290	-19.60%	+4.17
1950	189	57	0.5030	0.4379	+12.94%	-6.85
1960	216	71	0.5039	0.3829	+24.01%	-19.50

EXPLORATORY — DOES NOT ALTER T-002'S REGISTERED EVIDENCE CATEGORY



Temporal log loss

Early training sets are small and historically narrow; very poor losses show that these estimated probabilities transport badly. Later favorable origins are retained alongside the unfavorable ones and do not replace their pooled result. IDP's pooled loss is 3.096 times baseline loss. No origin was removed because its prediction was unfavorable.

Origin	O* R	O* ΔP (pp)	E* R	E* ΔP (pp)
1880	-318.99%	+28.78	Separation	Unavailable
1890	-122.34%	+22.05	-434.04%	+34.23
1900	-198.21%	+73.95	-1492.02%	+8.47
1910	-226.87%	+50.60	-475.66%	+10.91
1920	-238.72%	+73.80	-154.32%	+12.24
1930	-45.67%	+19.96	-27.36%	+4.27
1940	-43.41%	+8.74	-7.84%	+2.63
1950	-19.21%	+13.65	+16.31%	-13.74
1960	+23.08%	+1.90	+20.10%	-23.95

E* is non-estimable at 1880 because of complete or quasi-separation. Its remaining origins are reported individually, but an eight-origin favorable subset is not substituted for the nine-origin comparison. O* fails the corresponding prospective-signal pattern because pooled temporal R is negative and ΔP is positive. E* cannot establish that pattern across all required origins. These diagnostics do not distinguish the full IDP result as prospective evidence beyond a good-governance alternative. No diagnostic veto or good-governance evidence-category rule is applied to the frozen invalid test.

5. Robustness and one additional fixed comparison

Specification	n	Onsets	Baseline LL	Model LL	R	ΔP (pp)
25-year, primary rows	543	53	0.3740	0.3602	+3.70%	-30.45
100-year	322	53	0.4758	0.5319	-11.80%	-20.00
Six-component MMP	369	48	0.5191	0.4795	+7.62%	-34.15
IDP level only	543	53	0.5112	0.4791	+6.28%	-38.19
IDP trajectory only	543	53	0.5112	0.4987	+2.44%	-20.04
Omit public goods G	543	53	0.5112	0.4769	+6.71%	-37.52
Omit consultation A	543	53	0.5112	0.4810	+5.89%	-37.65
Omit administration I	543	53	0.5112	0.4723	+7.60%	-39.02
Omit stewardship S	543	53	0.5112	0.4760	+6.87%	-38.24
Exclude pre-1900	380	51	0.6037	0.5220	+13.54%	-74.33
Calendar-mature 50-year	472	32	0.4739	0.4671	+1.44%	-23.42

These are numerically estimable exploratory versions of the registered robustness set, plus the separately labeled calendar-mature comparison. Applicable registered secondary feasibility floors remain unmet, particularly the 60-onset requirement; results are not reinstated as verdict-relevant robustness. The 100-year sample also falls below its 400-landmark floor. The pre-1900 exclusion does not create new observations or events. Regional blocking is reported next. Full fold counts and secondary metrics for every comparison are supplied in the machine-readable results.

The calendar-mature restriction retains 472 landmarks, 79 polities, 32 onsets and 26 event-bearing polities. Improvement falls from 7.26% to 1.44%, with a -23.42-point contrast. Among the 71 primary landmarks outside nominal calendar maturity, 70 are positive and one is negative. Incomplete forward windows can be included when a failure is observed but excluded when no failure is observed. This creates an outcome-dependent ascertainment concern. The restriction also changes historical composition, so its result does not establish that ascertainment explains the whole original association.

The inherited outcome audit has two rows beyond the nominal 2018 endpoint, through 2020 for one polity. Saved T-002 rows were not changed. The exploratory scope fixed the nominal 2018 maturity cutoff before fitting; that choice, and the later correction of diagnostic endpoint accounting from global maximum 2020 to nominal 2018, are documented in the scope addendum. This endpoint inconsistency deserves explicit treatment in any successor protocol.

6. Variation across regions and time

All six leave-one-region-out fits are numerically estimable. No lineage has an inconsistent non-missing region in the inherited primary sample. Regions use V-Dem's frozen six-category coding, not a newly selected geographic grouping. Loss averages within polity, then within region, then equally across six held-out regions.

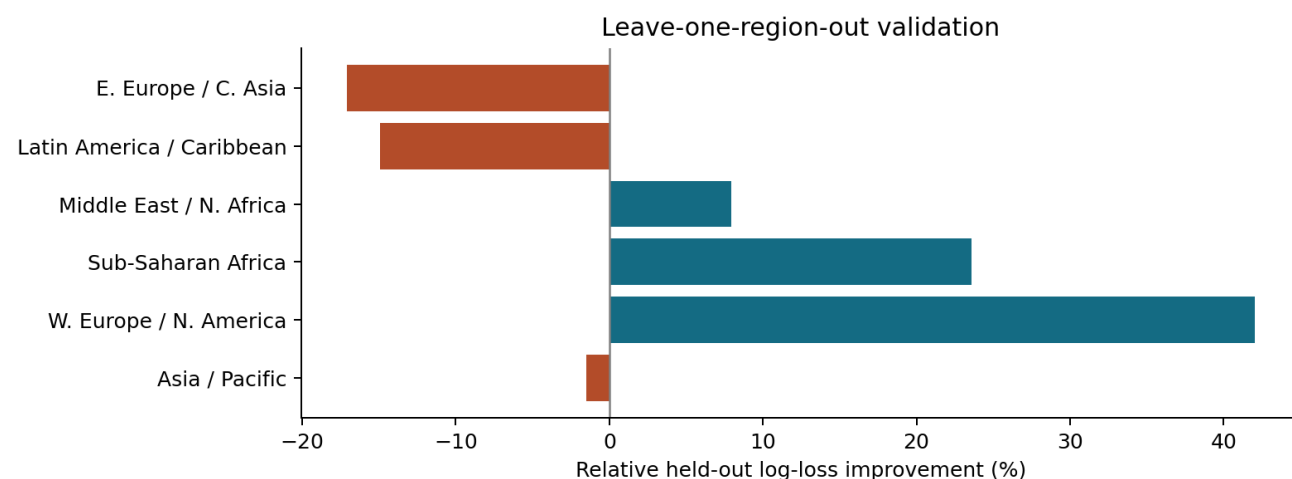
Held-out region	n	Onsets	Baseline LL	Model LL	R	ΔP (pp)
E. Europe / C. Asia	41	8	0.7090	0.8304	-17.12%	-45.75
Latin America / Caribbean	179	9	0.5155	0.5924	-14.91%	-37.26
Middle East / N. Africa	51	7	0.6089	0.5607	+7.92%	-38.78
Sub-Saharan Africa	50	18	0.7305	0.5584	+23.57%	-44.80
W. Europe / N. America	177	4	0.4890	0.2835	+42.04%	-18.97

EXPLORATORY — DOES NOT ALTER T-002'S REGISTERED EVIDENCE CATEGORY

Held-out region	n	Onsets	Baseline LL	Model LL	R	ΔP (pp)
Asia / Pacific	45	7	0.5933	0.6022	-1.50%	-43.81

Equal-region pooled baseline loss is 0.6077, IDP loss 0.5712, R +6.00%, and ΔP -38.23 points. Loss improves in three regions and worsens in three, while all contrasts are negative. The Western Europe/North America category contains only four distinct onsets; its large improvement should not be interpreted as a stable ranking of regions.

EXPLORATORY — DOES NOT ALTER T-002'S REGISTERED EVIDENCE CATEGORY



Regional transfer

The following time summaries reuse the same grouped-CV held-out predictions. They are not separate fitted historical models. Because one event can be inside landmarks from more than one era, event counts across eras must not be summed.

Landmark era	n	Onsets	R	ΔP (pp)
before1900	163	5	-7.94%	-12.20
1900-1949	181	22	+0.00%	-40.37
1950onward	199	38	+21.39%	-49.71

Most grouped-CV improvement is concentrated in later landmarks. The 1900-1949 improvement is only 0.0048%, effectively flat at the displayed precision. The later group also contains the follow-up selection problem described above. Historical and regional heterogeneity therefore limits a broad transport claim.

For completeness, region summaries of the original grouped-CV predictions below differ from leave-one-region-out validation: the latter withholds an entire region from training, whereas grouped CV withholds polities while other polities in that region can remain in training.

Region	Grouped-CV R	Grouped-CV ΔP (pp)
E. Europe / C. Asia	-15.10%	-39.60
Latin America / Caribbean	-1.73%	-36.63
Middle East / N. Africa	+16.09%	-39.92
Sub-Saharan Africa	+21.62%	-47.79
W. Europe / N. America	+13.72%	-32.12
Asia / Pacific	+9.52%	-43.60

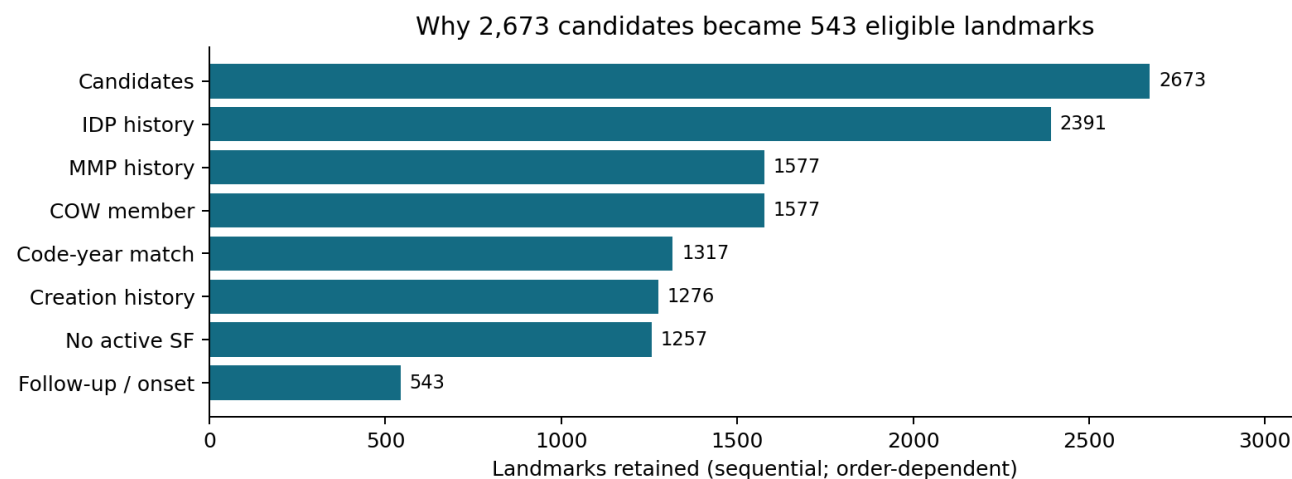
7. Exact feasibility diagnosis

Mandatory primary requirement	Observed	Registered floor
Eligible landmarks	543	600
Distinct polities	98	60
Distinct qualifying onsets	53	80
Polities with onsets	46	30

The shortfalls are 57 landmarks and 27 distinct onsets. They are not repaired by successful model fitting, by multiple landmark windows containing the same event, or by exploratory intervals.

Sequential requirement	Removed	Remaining	Represented onsets
All constructed candidate landmarks	0	2673	57
IDP ten-year complete history	282	2391	57
MMP five-year complete history	814	1577	54
COW member at landmark	0	1577	54
Unique Polity code-year mapping	260	1317	54
CHANGE99 history requirement	41	1276	53
Exclude active SF at landmark	19	1257	53
Incomplete follow-up without onset	714	543	53

EXPLORATORY — DOES NOT ALTER T-002'S REGISTERED EVIDENCE CATEGORY



Feasibility flow

This attribution is order-dependent, not a causal decomposition. All 814 sequential MMP losses include a missing required CINC year outside COW membership or coverage. This does not imply that all could be recovered by imputing CINC: many fail the target-state coverage requirement itself. The COW-membership stage removes zero additional rows because of the preceding history filter.

Reason for incomplete non-event follow-up	Landmarks lost
Nominal source endpoint	594
COW membership gap	53
Polity or lineage gap before endpoint	67

Reason for incomplete non-event follow-up	Landmarks lost
Total	714

The 714 losses remove non-onset windows by construction, so they remove no distinct represented events. They nevertheless substantially change the balance and timing of observable negative windows. Their row-level audit is supplied as EXPLORATORY_excluded_followup_landmarks.csv.

Standalone failure (overlapping)	Landmarks
IDP_history_missing	282
MMP_history_missing	993
not_COW_member	905
no_Polity_mapping	1058
creation_history_insufficient	126

Standalone failures overlap and cannot be added. Ten-year component-history failures are G: 252, A: 249, I: 252 and S: 271; these overlap within the 282 IDP-history failures. Shortening an institutional window would be a new construct decision, and was not explored to seek better predictions.

Under the inherited outcome and crosswalk rules, 59 observable onsets occur from 1821 onward. Two are unreachable from any constructed qualifying candidate window, three more lose all candidate support at the MMP-history stage, and one at the creation-history stage, leaving 53. No additional distinct onset is lost solely at the IDP-history stage in this sequential ordering. The event-retention audit gives the identifier and exclusion stage for each of the 59; these are counts within the inherited constructed universe, not a claim to enumerate every historical breakdown.

The entity crosswalk was not broadened. The inherited pipeline has no explicit supplied predecessor/successor link table and uses code-year matching and its existing exit-censoring fallback; it does not manually merge historically related entities. Thus both the 59-event ceiling and loss attribution are conditional on that mapping. The follow-up has not re-audited every historical identity or silently repaired continuity.

8. Feasibility of genuinely different prospective designs

The next table changes only the candidate horizon or anchor for descriptive feasibility accounting. It is not a predictive specification search. In particular, the all-eligible 25-year sample below differs from the registered 25-year robustness comparison above, which was confined to the 543 primary rows.

Anchor	Horizon	Landmarks	Polities	Onsets	Calendar-mature rows
Decennial	25	885	154	53	871
Decennial	50	543	98	53	472
Decennial	100	322	70	53	185
Annual	25	8510	154	53	8368
Annual	50	5516	104	53	4906
Annual	100	3206	71	53	1918

A 25-year design materially improves observation coverage (885 decennial landmarks and 154 polities) but still has 53 onsets. Annual 50-year anchors give 5,516 rows and 104 polities, also with 53 onsets. Annual sampling does not generate independent event information and cannot be used to claim an 80-event design. One hundred years reduces usable follow-up further. No alternative horizon or annual-anchor model was optimized for a favorable result.

Successor-design recommendations

1. Register a genuinely new test with a new identifier and disclose that these data have been examined. Reusing this panel is discovery or development work; a new confirmatory claim needs a defensible untouched validation source, future outcomes or other genuinely independent evaluation. T-002's verdict stays fixed.
2. Decide the target population, administrative endpoint, entry, succession and competing exits before fitting. Choose either uniformly mature fixed-horizon landmarks or a fully specified event-history/right-censoring design with appropriate censor-aware evaluation. Do not label unknown future outcomes as non-events. The exact treatment of the two post-2018 records should be frozen prospectively.
3. Treat a 25-year horizon as a possible substantively justified question, not a rescue. It expands coverage but leaves the independent onset shortage unchanged. Increasing event information would require independently justified source expansion, historical coverage or outcome ascertainment; broadening "breakdown" merely to hit a target count would change the question.
4. Base the new sample-size and precision plan on anticipated risk, model complexity, dependence, follow-up and plausible effect sizes, using simulation where appropriate. More repeated landmarks are not a substitute for more independent events. Do not lower an event floor simply because this exposed sample contains 53 onsets.
5. Specify minimum temporal training information and numerical estimability in advance. Early temporal failures suggest a stability problem worth addressing in a new design; a prospectively chosen regularized estimator could be investigated in development, but none was substituted here. Preserve a final chronological evaluation and region transfer checks, reporting all prespecified periods rather than selecting favorable late origins.
6. Audit source coverage and formal entity continuity independently of prediction performance. Verify the support of high-Mass contrasts and retain O*/E* diagnostics, because a full IDP association alone does not establish the broader PrF interpretation. Preserve the distinction between empirical MMP/IDP and ontological Mass/Direction.

These recommendations are methodological proposals, not additional executed specifications. General prediction-model sample-size principles are discussed by Riley et al. (2020), "Calculating the sample size required for developing a clinical prediction model," *BMJ* 368:m441, <https://doi.org/10.1136/bmj.m441>. That clinical-methods reference does not supply a validated numeric threshold for polity data. Censor-aware prediction evaluation is discussed by Gerds and Schumacher (2006), "Consistent estimation of the expected Brier score in general survival models with right-censored event times," <https://pubmed.ncbi.nlm.nih.gov/17240660/>. Its assumptions would require explicit treatment in a successor design.

9. Reproducibility, deviations and limits

The package includes the pre-fit exploratory scope, later scope addendum, code, input hash record, every result table and held-out prediction file, fold assignment, all 2,000 bootstrap estimates, feasibility/event-retention audits, figures and this report. It includes the exact inherited derived inputs and frozen verdict for reproducibility. No new dataset was retrieved for these exploratory analyses. Original source versions remain COW NMC v7.0, COW State System Membership v2024, V-Dem v16 and Polity5 v2018; their existing provenance record is included unchanged.

Exploratory departures are explicit: fitting despite failed registered feasibility; an exploratory fold assignment fixed once; the calendar-mature subset; descriptive annual/all-eligible-horizon counts; and post-fit time/region summaries fixed in the exploratory scope. The scope addendum records the source-end accounting correction. No predictor tuning, coefficient-sign search, alternative penalty, favorable-fold selection or confirmatory rerun occurred. During final packaging, one saved IDP held-out prediction CSV was found empty; only its five baseline/full-IDP fold fits were regenerated using unchanged code, data and assignments. All recovered point estimates agreed with the completed results within 10^{-12} . The bootstrap and other analyses were not rerun; the recovery is logged. All inherited input hashes were verified before analysis and again before packaging.

Independent SciPy BFGS fits reproduce the custom Newton solver's primary training probabilities within 3.4×10^{-9} . All primary region values are present. This is a numerical check, not external validation of the historical measurements. Bootstrap uncertainty remains conditional on inherited operationalisations, sampling and fold choices. No correction for a family of confirmatory hypothesis tests is claimed because this entire follow-up is exploratory.

The most defensible narrow conclusion is that this exposed historical panel contains a negative high-Mass institutional contrast in grouped CV, with uncertain predictive gain and poor forward transport. It motivates better follow-up and independent validation design; it does not establish or refute an ontological PrF Mass/Direction law. **T-002's registered evidence category remains Invalid / Uninterpretable.**