

Temporally non-local harmonisation and cardiovascular model transport between England and China: a measurement-methods study

Han Liu

University College London

Qun Liang

liangqun1@sina.com

The First Affiliated Hospital of Heilongjiang University of Chinese Medicine

Research Article

Keywords: cardiovascular disease, data harmonisation, external validation, longitudinal studies, measurement error, prediction models, sex interaction

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10821704/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Harmonised disease histories can incorporate later reports, altering eligibility, outcomes, and model transport.

Methods

Reciprocal external validation compared released and time-anchored histories among CHARLS and ELSA participants aged 60 years or older. Analyses used full and common risk sets, prespecified sex interactions, multiple imputation, and 500 paired household bootstraps.

Results

Histories disagreed in 8.95% of CHARLS and 6.11% of ELSA participants; gross churn was 10.57% and 7.91%. Released and time-anchored calibration slopes were 0.173 (95% CI – 0.121 to 0.468) and 0.473 (0.189 to 0.758) for China-to-England transport, and 0.299 (0.002 to 0.596) and 0.478 (0.208 to 0.748) for England-to-China transport. AUC and Brier error both increased; IPA remained negative. Common-risk-set calibration-slope, AUC, and IPA estimates and all sex contrasts had confidence intervals spanning zero. England-to-China common-risk-set Brier error increased.

Conclusions

In harmonised cohorts, incorporating diagnoses reported at later waves into earlier disease histories changed cardiovascular disease-free eligibility and interval-specific outcome ascertainment among adults aged 60 years or older, and changed the calibration slope for China-to-England transport. The difference between full-pipeline and common-risk-set results indicated that this effect operated through cohort composition rather than predictor recoding and did not constitute an improvement in overall predictive performance.

Introduction

Cardiovascular disease is a consequential outcome in ageing cohorts because its burden differs across populations and historical periods (Li et al., 2025). Longitudinal comparison requires a disease-free baseline and a consistently defined incident outcome because external performance depends on the target population and outcome definition (Debray et al., 2015). If information elicited at a later interview enters an earlier disease history, the distributed harmonised variables can change both eligibility and event timing before any prediction model is fitted. The CHARLS documentation records these revisions

through wave-specific dispute fields (Phillips et al., 2021). The ELSA G.3 documentation contains analogous temporal information (Wilkins et al., 2023).

The English Longitudinal Study of Ageing repeatedly measures health, behaviour, and socioeconomic conditions in adults living in England (Stephens et al., 2013). The China Health and Retirement Longitudinal Study uses a parallel longitudinal design in Chinese adults (Zhao et al., 2014). Harmonised releases align variable names and coding to support cross-national analysis (Lee et al., 2021). External validation then asks whether a model developed in one cohort retains calibration and discrimination in another population (Debray et al., 2015).

The unresolved problem is that harmonised disease histories can be temporally non-local. In CHARLS, a value stored for one wave may incorporate information disputed at a later interview (Phillips et al., 2021). ELSA G.3 uses corresponding flags to preserve similar revisions (Wilkins et al., 2023). Self-reported cardiovascular events can differ from linked hospital records even when the survey variable appears clinically specific (Kivimäki et al., 2017). Agreement also varies across cardiovascular endpoints and populations (Sabanayagam et al., 2024). Existing external-validation frameworks separate target-population and model-performance problems but do not isolate temporally revised eligibility from revision of predictors and outcomes within the same participants (Debray et al., 2015).

This gap creates a testable tension between temporal validity and predictive utility: a phenotype can be more faithful to what was observable at baseline yet transport no better if apparent improvement arises mainly from changing who enters the risk set. Diagnosis reporting is shaped by access to care and diagnostic opportunity (Andersen, 1995). Survey wording and retrieval conditions can also change self-reported histories without changing disease at the earlier time point (Schwarz, 1999). External validation therefore needs to distinguish composition, predictor-history, and outcome-history pathways rather than describe only their combined performance change (Debray et al., 2015).

This study aimed to determine whether temporally non-local harmonisation changed cardiovascular disease-free eligibility, interval-specific event ascertainment, and reciprocal model transport between CHARLS and ELSA G.3 among adults aged 60 years or older. The primary comparison contrasted released with time-anchored histories across the complete analysis pipeline. A fixed common-risk-set factorial analysis separated predictor-history, outcome-history, interaction, and sample-composition pathways. Sensitivity analyses examined heart disease, stroke, and alternative event-construction rules; subgroup analyses formally tested modification by recorded sex.

Methods

Study design and data sources

This observational measurement-methods study combined a longitudinal phenotype audit with reciprocal external validation. Observational reporting followed the STROBE statement (von Elm et al.,

2008). Prediction-model reporting followed TRIPOD (Collins et al., 2015). Analyses by recorded sex followed the SAGER guidance (Heidari et al., 2016).

ELSA is a multidisciplinary panel study of a representative cohort of men and women aged 50 years or older living in England (Steptoe et al., 2013). CHARLS is a nationally representative longitudinal survey of Chinese adults aged 45 years or older (Zhao et al., 2014). Harmonized CHARLS version D provided wave 1 (2011–2012) as baseline and waves 2 (2013) and 3 (2015) as follow-up (Phillips et al., 2021). Harmonized ELSA version G.3 provided wave 4 (2008–2009) as baseline and waves 5 (2010–2011) and 6 (2012–2013) as follow-up (Wilkens et al., 2023). The baseline waves were aligned by analytic role rather than calendar year: each was the first wave of a three-wave sequence with all required disease histories, dispute flags, predictors, and interview-status variables.

Exposure of interest

The analytical exposure was history definition, not a participant characteristic. CHARLS released histories used the distributed wave-specific ever-diagnosed variables for heart disease, stroke, hypertension, and diabetes (Phillips et al., 2021). ELSA released histories used the corresponding G.3 variables (Wilkens et al., 2023). CHARLS time-anchored histories reconstructed the state observable at each interview from its dispute flags (Phillips et al., 2021). ELSA time-anchored histories applied the analogous G.3 flags (Wilkens et al., 2023). CHARLS flag values 2, 3, 4, 5, and 6 mapped to yes, no, yes, no, and yes, respectively (Phillips et al., 2021). ELSA flag values 2, 3, and 4 mapped to no, while value 6 mapped to yes (Wilkens et al., 2023). Within CHARLS, flag 0 retained the released value and non-substantive values were treated as missing (Phillips et al., 2021). The same rule was applied to the corresponding ELSA values (Wilkens et al., 2023). The reconstruction tested temporal admissibility within the distributed files and did not assert diagnostic truth.

Covariates

The transported model included baseline age, recorded sex, current smoking, education, diagnosed hypertension, and diagnosed diabetes. CHARLS covariates followed the harmonised version D definitions (Phillips et al., 2021). ELSA covariates followed the corresponding G.3 definitions (Wilkens et al., 2023). Age was centred and scaled in the development cohort and entered as linear and quadratic terms. Education was grouped as lower than upper secondary, upper secondary or vocational, and tertiary. Non-substantive CHARLS education codes were treated as missing (Phillips et al., 2021). The same rule was applied to the corresponding ELSA codes (Wilkens et al., 2023). Hypertension and diabetes followed the predictor-history definition assigned to each analytical regime. Prespecified terms linked recorded sex with age, smoking, hypertension, and diabetes.

Outcome

The primary outcome was incident composite cardiovascular disease, defined from the harmonised heart-disease and stroke histories over two follow-up intervals. CHARLS component histories followed

the version D definitions (Phillips et al., 2021). ELSA component histories followed the G.3 definitions (Wilkins et al., 2023). A composite history was positive when either component was positive, negative only when both components were observed negative, and missing otherwise. Heart disease and stroke were analysed separately as endpoint-specific sensitivity analyses. Death before an observed cardiovascular event was the competing event. Alive non-response, unknown vital status, dropped status, or missing disease history censored follow-up at the first unobserved interval; return after a missed interview was not used because event order could not be established.

Study population and sample selection

Participants were eligible when they were interviewed at baseline, aged 60 years or older, and had recorded sex. Baseline cardiovascular status was then evaluated separately under the released and time-anchored definitions. Participants with missing baseline cardiovascular status or prevalent cardiovascular disease were excluded from the corresponding native risk set. The common risk set retained participants classified as disease-free under both definitions. Predictor missingness did not trigger exclusion. Figure 1 records every sequential exclusion and the definition-specific risk-set counts.

Missing values

Missing age, recorded sex, or baseline cardiovascular status prevented eligibility assessment. Missing smoking, education, hypertension, or diabetes did not remove an otherwise eligible participant. Predictors were multiply imputed within each cohort, endpoint, scope, and predictor-history specification after the risk set was fixed. Chained equations used logistic regression for binary variables and polytomous regression for education, with 20 imputations and 10 iterations. Age and sex were complete in the eligible samples and were not imputed. Outcomes, death, censoring, later-wave predictors, and all post-baseline information were excluded from the imputation model; outcomes were never imputed.

Statistical analysis

The phenotype audit estimated released-versus-time-anchored disagreement with counts, percentages, transition direction, and Wilson 95% confidence intervals. Binomial logit models with household-clustered HC0 covariance tested cohort-by-sex and cohort-by-sex-by-endpoint interactions. These models require variation in the binary outcome within analytical cells and independent clusters; cross-tabulations confirmed both revised and unrevised histories in every cohort–sex–endpoint cell, and household identifiers defined the independent clusters.

Reciprocal transport used pooled multinomial logistic regression for discrete-time competing risks, with no cardiovascular event as the reference outcome and interval-specific intercepts. The model included interval, standardised age, age squared, recorded sex, smoking, education, hypertension, diabetes, and the prespecified sex interactions. Source-cohort centring constants and coefficients were transported without refitting. Predicted two-interval cardiovascular risk equalled interval-1 cardiovascular risk plus interval-1 event-free probability multiplied by interval-2 cardiovascular risk. Performance was evaluated

by calibration-in-the-large, calibration slope, Brier score, IPA, and AUC among participants with an observed horizon outcome.

Multinomial fitting assumes mutually exclusive interval outcomes, adequate support for each outcome category, and a common coefficient vector across transported settings. The event algorithm enforced mutually exclusive cardiovascular disease, death, no-event, and censoring states. Cross-tabulations confirmed cardiovascular disease, death, and no-event observations in every development dataset. Calibration and discrimination were reported together because no single metric can establish transportability (Debray et al., 2015).

Point estimates averaged 20 multiply imputed analyses. Uncertainty used 500 paired household bootstrap replicates that independently resampled households in both cohorts; two stochastic regression imputations were generated within each replicate before model refitting and transport. Percentile 95% confidence intervals and two-sided bootstrap P values were calculated from completed replicates. Nesting imputation within bootstrap resampling propagates uncertainty from both missing predictors and sampling (Bartlett & Hughes, 2020).

Primary full-pipeline and common-risk-set joint contrasts used two-sided $P < .05$ without multiplicity adjustment. Benjamini–Hochberg q values accompanied the exploratory endpoint-specific and sex-interaction families. Effect estimates and confidence intervals governed interpretation; a confidence interval spanning zero was treated as insufficient evidence of a contrast, not as equivalence.

Sensitivity analyses

The primary sensitivity analysis fixed eligibility to the common risk set and crossed released and time-anchored predictor and outcome histories. RR denoted released predictors and outcomes, AR time-anchored predictors with released outcomes, RA released predictors with time-anchored outcomes, and AA time-anchored predictors and outcomes. For each metric M, predictor-history, outcome-history, joint, and interaction effects were $M(\text{AR}) - M(\text{RR})$, $M(\text{RA}) - M(\text{RR})$, $M(\text{AA}) - M(\text{RR})$, and $M(\text{AA}) - M(\text{RA}) - M(\text{AR}) + M(\text{RR})$, respectively. Endpoint sensitivity repeated the complete analysis for heart disease and stroke. Event-construction sensitivity evaluated 16 plausible combinations of interview requirements, death codes, and event-versus-death precedence before applying the selected rule to ELSA G.3.

Subgroup analyses

Recorded sex was the prespecified subgroup variable. Models were fitted in the full sample, and performance was evaluated separately in men and women without refitting the transported coefficients. Sex modification was the paired women-minus-men difference in the AA-minus-RR transport effect, with its bootstrap confidence interval. Differences between subgroup-specific P values were not treated as interaction evidence. Heart-disease-versus-stroke contrasts were estimated analogously as formal endpoint interaction contrasts.

Mechanism analysis

The mechanism analysis tested why temporal anchoring changed transport performance. The fixed-risk-set predictor-history, outcome-history, joint, and interaction contrasts quantified changes attributable to recoding within the same participants. The risk-set composition pathway was defined as the full-pipeline AA-minus-RR effect minus the fixed-risk-set joint effect. A non-zero composition contrast therefore indicated that rebuilding eligibility, rather than predictor or outcome recoding alone, accounted for part of the overall performance change. This was a statistical decomposition of the measurement process, not a biological mediation analysis.

Software and reproducibility

Analyses used R 4.5.1 with haven, data.table, mice, nnet, and sandwich. Script 05_full_transport_g3.R generated all transport, sensitivity, subgroup, and mechanism estimates; script 01_phenotype_audit_and_subgroups.R generated phenotype interactions; script 03_event_timing_audit.R generated event-rule sensitivity; and scripts 06–08 generated tables, figures, and the manuscript. Aggregate outputs and session information are supplied with the analysis package.

Ethics

This secondary analysis used de-identified data under the applicable ELSA and CHARLS data-use conditions. CHARLS was approved by the Biomedical Ethics Review Committee of Peking University (IRB00001052-11015). ELSA waves 4, 5, and 6 were approved under references 07/H0716/48, 09/H0505/124, and 11/SC/0374, respectively. Both cohorts obtained participant consent; no new data were collected.

Results

Study population

The sample-selection analysis retained 7,290 of 25,586 CHARLS records and 7,321 of 19,802 ELSA G.3 records after baseline interview, age, and recorded-sex criteria. Released histories yielded 5,778 CHARLS and 5,578 ELSA risk-set participants; time anchoring yielded 5,741 and 5,292. The common risk sets contained 5,438 and 5,270 participants, respectively (Fig. 1). Predictor missingness did not reduce any eligible risk set (Table 1).

Table 1

Baseline characteristics and predictor missingness in the native composite cardiovascular disease-free risk sets

Characteristic	CHARLS released	CHARLS time-anchored	ELSA G.3 released	ELSA G.3 time-anchored
n	5,778	5,741	5,578	5,292
Age, years	68.2 (6.9)	68.2 (6.9)	69.5 (7.6)	69.4 (7.6)
Women	2,836/5,778 (49.1%)	2,792/5,741 (48.6%)	3,061/5,578 (54.9%)	2,918/5,292 (55.1%)
Current smoking	1,721/5,624 (30.6%)	1,705/5,578 (30.6%)	643/5,482 (11.7%)	615/5,197 (11.8%)
Lower education	5,408/5,777 (93.6%)	5,381/5,740 (93.7%)	2,013/5,095 (39.5%)	1,902/4,826 (39.4%)
Upper/vocational education	268/5,777 (4.6%)	263/5,740 (4.6%)	2,317/5,095 (45.5%)	2,205/4,826 (45.7%)
Tertiary education	101/5,777 (1.7%)	96/5,740 (1.7%)	765/5,095 (15.0%)	719/4,826 (14.9%)
Hypertension	1,632/5,754 (28.4%)	1,658/5,718 (29.0%)	2,355/5,578 (42.2%)	2,331/5,292 (44.0%)
Diabetes	334/5,743 (5.8%)	347/5,707 (6.1%)	540/5,578 (9.7%)	537/5,292 (10.1%)
Smoking missing	154 (2.7%)	163 (2.8%)	96 (1.7%)	95 (1.8%)
Education missing	1 (0.0%)	1 (0.0%)	483 (8.7%)	466 (8.8%)

Values are mean (SD), n/observed n (%), or missing n (%). Missing predictor values were retained and multiply imputed; denominators for observed percentages exclude missing values.

Phenotype revision

The phenotype audit found released-versus-time-anchored composite disagreement in 643/7,188 CHARLS participants (8.95%, 95% CI 8.31% to 9.63%) and 330/7,318 ELSA G.3 participants (4.51%, 4.06% to 5.01%). Gross risk-set churn was 10.57% in CHARLS and 5.89% in ELSA G.3, compared with net displacement of -0.64% and - 5.13% (Table 2; Figs. 2 and 3).

In the binomial logit phenotype model with household-clustered HC0 covariance (n = 14,506), women had greater revision odds than men in CHARLS (OR 1.25, 95% CI 1.06 to 1.46; P=.007) but not in ELSA G.3 (OR 0.85, 0.68 to 1.05; P=.135). The cohort-by-sex interaction OR was 0.68 (0.52 to 0.89; P=.005).

Table 2
Baseline phenotype revision and risk-set turnover by cohort and endpoint

Measure	CHARLS composite	CHARLS heart	CHARLS stroke	ELSA G.3 composite	ELSA G.3 heart	ELSA G.3 stroke
Paired states, n	7,188	7,192	7,218	7,318	7,318	7,318
Revised, n (%; 95% CI)	643 (8.95%; 8.31–9.63)	537 (7.47%; 6.88–8.10)	181 (2.51%; 2.17–2.89)	330 (4.51%; 4.06–5.01)	356 (4.86%; 4.40–5.38)	39 (0.53%; 0.39–0.73)
R→A+ / R+→A-	340 / 303	298 / 239	78 / 103	308 / 22	329 / 27	39 / 0
Released risk set, n	5,778	6,012	6,907	5,578	5,837	6,890
Time-anchored risk set, n	5,741	5,953	6,932	5,292	5,535	6,851
Common risk set, n	5,438	5,714	6,829	5,270	5,508	6,851
Gross churn, %	10.57	8.59	2.58	5.89	6.07	0.57

R→A+ denotes released negative to time-anchored positive; R+→A- denotes the reverse transition. Wilson intervals describe revision percentages.

To determine whether baseline revision was concentrated in one endpoint, Fig. 2 compared endpoint-specific disagreement across cohorts. Heart-disease histories changed more often than stroke histories in CHARLS (7.47% versus 2.51%) and ELSA G.3 (4.86% versus 0.53%).

Endpoint-specific revision percentages do not show whether the same people remained eligible. Figure 3 therefore decomposed risk-set membership and showed that gross churn exceeded net displacement in CHARLS (10.57% versus – 0.64%) and ELSA G.3 (5.89% versus – 5.13%).

Primary reciprocal transport

To test the main research question, the 20-imputation multinomial logit model was transported in the China-to-England direction. The released pipeline evaluated n = 4,717 participants with 488 cardiovascular events and yielded a calibration slope of 0.206 (bootstrap 95% CI -0.131 to 0.479), AUC 0.530, Brier score 0.098, and IPA – 0.055; the time-anchored pipeline evaluated n = 4,443 with 474 events and yielded a calibration slope of 0.573 (0.225 to 0.808), AUC 0.582, Brier score 0.102, and IPA – 0.072. Observed and mean predicted risks were 0.103 and 0.154 under RR and 0.107 and 0.181 under AA, with calibration-in-the-large values of -0.467 and – 0.633, respectively. AA minus RR was 0.367 (95% CI 0.053

to 0.594; P=.008) for calibration slope, 0.0517 (0.0116 to 0.0779; P=.008) for AUC, 0.0044 (-0.0010 to 0.0108; P=.108) for Brier score, and - 0.0177 (-0.0834 to 0.0394; P=.468) for IPA.

To test the main research question, the 20-imputation multinomial logit model was transported in the England-to-China direction. The released pipeline evaluated n = 4,895 participants with 549 cardiovascular events and yielded a calibration slope of 0.371 (bootstrap 95% CI 0.021 to 0.656), AUC 0.546, Brier score 0.100, and IPA - 0.007; the time-anchored pipeline evaluated n = 4,917 with 674 events and yielded a calibration slope of 0.586 (0.244 to 0.845), AUC 0.575, Brier score 0.120, and IPA - 0.012. Observed and mean predicted risks were 0.112 and 0.087 under RR and 0.137 and 0.088 under AA, with calibration-in-the-large values of 0.281 and 0.510, respectively. AA minus RR was 0.214 (95% CI -0.006 to 0.448; P=.056) for calibration slope, 0.0291 (0.0037 to 0.0574; P=.020) for AUC, 0.0195 (0.0131 to 0.0260; P=.004) for Brier score, and - 0.0054 (-0.0163 to 0.0045; P=.296) for IPA.

Table 3
Full-pipeline effect of time anchoring on reciprocal transport performance

Direction	Metric	AA – RR	Bootstrap 95% CI	P	Successful replicates
China → England	Calibration slope	0.367	0.053 to 0.594	.008	500
China → England	AUC	0.0517	0.0116 to 0.0779	.008	500
China → England	Brier score	0.0044	-0.0010 to 0.0108	.108	500
China → England	IPA	-0.0177	-0.0834 to 0.0394	.468	500
England → China	Calibration slope	0.214	-0.006 to 0.448	.056	500
England → China	AUC	0.0291	0.0037 to 0.0574	.020	500
England → China	Brier score	0.0195	0.0131 to 0.0260	.004	500
England → China	IPA	-0.0054	-0.0163 to 0.0045	.296	500

Positive values indicate that the metric was higher under time-anchored than released histories. Higher AUC and IPA are preferable; lower Brier score is preferable; calibration slope is interpreted relative to 1.

The paired changes reported in Table 3 do not show whether absolute calibration became adequate. Figure 4 therefore located each calibration slope against the ideal value of 1; all four point estimates remained below that benchmark.

Sensitivity analyses

The event-construction sensitivity analysis evaluated 16 plausible rule combinations. All 16 reproduced the reported CHARLS cardiovascular counts because event and no-event histories were concordant with interview-status ordering in the analysed records. Under the selected rule, released and time-anchored horizon counts were 549/4,895 and 674/4,917 in CHARLS, compared with 488/4,717 and 474/4,443 in ELSA G.3.

In the common-risk-set sensitivity analysis (China-to-England; $n = 5,270$ target participants for England and $n = 5,438$ for China before horizon censoring), the joint AA-minus-RR effect was 0.109 (95% CI -0.067 to 0.260; $P=.228$) for calibration slope, 0.0162 (-0.0079 to 0.0342; $P=.188$) for AUC, 0.0011 (-0.0004 to 0.0028; $P=.140$) for Brier score, and 0.0003 (-0.0166 to 0.0129; $P=.908$) for IPA.

In the common-risk-set sensitivity analysis (England-to-China; $n = 5,270$ target participants for England and $n = 5,438$ for China before horizon censoring), the joint AA-minus-RR effect was 0.085 (95% CI -0.060 to 0.210; $P=.260$) for calibration slope, 0.0103 (-0.0086 to 0.0284; $P=.224$) for AUC, 0.0049 (0.0031 to 0.0070; $P=.004$) for Brier score, and 0.0047 (-0.0011 to 0.0122; $P=.104$) for IPA.

Endpoint-specific full-pipeline sensitivity analyses used the same 20-imputation multinomial logit specification and 500 paired household bootstrap procedure as the composite outcome.

For Heart disease in the China → England direction, RR included 4,932/427 participants/events and AA included 4,643/423; AA minus RR was Calibration slope 0.313 (95% CI 0.033 to 0.519; $P=.008$; $q=.043$); AUC 0.0397 (95% CI 0.0047 to 0.0698; $P=.012$; $q=.048$); Brier 0.0039 (95% CI -0.0014 to 0.0098; $P=.136$; $q=.242$); IPA 0.0010 (95% CI -0.0731 to 0.0647; $P=.976$; $q=.992$).

For Heart disease in the England → China direction, RR included 5,107/472 participants/events and AA included 5,094/554; AA minus RR was Calibration slope 0.240 (95% CI -0.020 to 0.492; $P=.072$; $q=.144$); AUC 0.0304 (95% CI -0.0001 to 0.0632; $P=.048$; $q=.128$); Brier 0.0135 (95% CI 0.0073 to 0.0196; $P=.004$; $q=.032$); IPA - 0.0036 (95% CI -0.0107 to 0.0052; $P=.428$; $q=.527$).

For Stroke in the China → England direction, RR included 5,773/145 participants/events and AA included 5,734/145; AA minus RR was Calibration slope 0.295 (95% CI -0.004 to 0.608; $P=.060$; $q=.137$); AUC 0.0537 (95% CI 0.0037 to 0.0984; $P=.036$; $q=.115$); Brier 0.0001 (95% CI -0.0001 to 0.0006; $P=.244$; $q=.390$); IPA 0.0016 (95% CI -0.0175 to 0.0118; $P=.992$; $q=.992$).

For Stroke in the England → China direction, RR included 5,854/147 participants/events and AA included 5,894/211; AA minus RR was Calibration slope 0.188 (95% CI -0.150 to 0.474; $P=.320$; $q=.437$); AUC 0.0267 (95% CI -0.0249 to 0.0791; $P=.328$; $q=.437$); Brier 0.0100 (95% CI 0.0061 to 0.0135; $P=.004$; $q=.032$); IPA 0.0007 (95% CI -0.0064 to 0.0129; $P=.748$; $q=.855$).

Table 4
Fixed-risk-set factorial sensitivity analysis

Direction	Contrast	Metric	Estimate	Bootstrap 95% CI	P
China → England	Predictor-history AR – RR	Calibration slope	0.061	-0.063 to 0.190	.348
China → England	Predictor-history AR – RR	AUC	0.0106	-0.0068 to 0.0259	.232
China → England	Predictor-history AR – RR	Brier	0.0000	-0.0008 to 0.0010	.908
China → England	Predictor-history AR – RR	IPA	-0.0005	-0.0114 to 0.0087	.908
China → England	Outcome-history RA – RR	Calibration slope	0.029	-0.070 to 0.129	.540
China → England	Outcome-history RA – RR	AUC	0.0034	-0.0112 to 0.0169	.604
China → England	Outcome-history RA – RR	Brier	0.0011	-0.0001 to 0.0026	.100
China → England	Outcome-history RA – RR	IPA	0.0001	-0.0123 to 0.0096	.732
China → England	Joint AA – RR	Calibration slope	0.109	-0.067 to 0.260	.228
China → England	Joint AA – RR	AUC	0.0162	-0.0079 to 0.0342	.188
China → England	Joint AA – RR	Brier	0.0011	-0.0004 to 0.0028	.140
China → England	Joint AA – RR	IPA	0.0003	-0.0166 to 0.0129	.908
China → England	Interaction AA – RA-AR + RR	Calibration slope	0.020	-0.053 to 0.082	.776
China → England	Interaction AA – RA-AR + RR	AUC	0.0021	-0.0076 to 0.0107	.852
China → England	Interaction AA – RA-AR + RR	Brier	-0.0001	-0.0004 to 0.0004	.840
China → England	Interaction AA – RA-AR + RR	IPA	0.0007	-0.0042 to 0.0048	.840
England → China	Predictor-history AR – RR	Calibration slope	0.062	-0.068 to 0.185	.296

Direction	Contrast	Metric	Estimate	Bootstrap 95% CI	P
England → China	Predictor-history AR – RR	AUC	0.0077	-0.0090 to 0.0233	.308
England → China	Predictor-history AR – RR	Brier	-0.0002	-0.0006 to 0.0001	.248
England → China	Predictor-history AR – RR	IPA	0.0023	-0.0019 to 0.0079	.248
England → China	Outcome-history RA – RR	Calibration slope	0.033	-0.048 to 0.114	.468
England → China	Outcome-history RA – RR	AUC	0.0031	-0.0067 to 0.0153	.484
England → China	Outcome-history RA – RR	Brier	0.0051	0.0033 to 0.0071	.004
England → China	Outcome-history RA – RR	IPA	0.0027	-0.0012 to 0.0084	.168
England → China	Joint AA – RR	Calibration slope	0.085	-0.060 to 0.210	.260
England → China	Joint AA – RR	AUC	0.0103	-0.0086 to 0.0284	.224
England → China	Joint AA – RR	Brier	0.0049	0.0031 to 0.0070	.004
England → China	Joint AA – RR	IPA	0.0047	-0.0011 to 0.0122	.104
England → China	Interaction AA – RA–AR + RR	Calibration slope	-0.010	-0.054 to 0.036	.756
England → China	Interaction AA – RA–AR + RR	AUC	-0.0004	-0.0073 to 0.0050	.692
England → China	Interaction AA – RA–AR + RR	Brier	0.0000	-0.0002 to 0.0003	.872
England → China	Interaction AA – RA–AR + RR	IPA	-0.0003	-0.0039 to 0.0026	.796

RR, released predictors and outcomes; AR, time-anchored predictors and released outcomes; RA, released predictors and time-anchored outcomes; AA, time-anchored predictors and outcomes.

Subgroup analyses

In the same 20-imputation multinomial logit model, the China-to-England sex analysis included 2,129/1,993 men and 2,588/2,450 women with observed horizons under RR/AA, respectively.

In the same 20-imputation multinomial logit model, the England-to-China sex analysis included 2,507/2,540 men and 2,388/2,377 women with observed horizons under RR/AA, respectively.

The formal sex: women – men interaction in the China → England full-pipeline AA-minus-RR effect was Calibration slope – 0.018 (95% CI -0.463 to 0.425; P=.928; q=.988); AUC 0.0006 (95% CI -0.0637 to 0.0552; P=.824; q=.988); Brier – 0.0069 (95% CI -0.0150 to 0.0009; P=.072; q=.988); IPA 0.0152 (95% CI -0.0520 to 0.1019; P=.664; q=.988). Interaction contrasts, rather than differences between subgroup-specific P values, defined effect modification.

The formal sex: women – men interaction in the England → China full-pipeline AA-minus-RR effect was Calibration slope – 0.140 (95% CI -0.907 to 0.426; P=.548; q=.988); AUC – 0.0163 (95% CI -0.0800 to 0.0346; P=.528; q=.988); Brier – 0.0070 (95% CI -0.0191 to 0.0044; P=.280; q=.988); IPA – 0.0092 (95% CI -0.0303 to 0.0087; P=.320; q=.988). Interaction contrasts, rather than differences between subgroup-specific P values, defined effect modification.

The formal endpoint: heart–stroke interaction in the China → England full-pipeline AA-minus-RR effect was Calibration slope 0.018 (95% CI -0.439 to 0.434; P=.952; q=.988); AUC – 0.0140 (95% CI -0.0696 to 0.0442; P=.748; q=.988); Brier 0.0038 (95% CI -0.0016 to 0.0096; P=.152; q=.988); IPA – 0.0006 (95% CI -0.0692 to 0.0709; P=.988; q=.988). Interaction contrasts, rather than differences between subgroup-specific P values, defined effect modification.

The formal endpoint: heart–stroke interaction in the England → China full-pipeline AA-minus-RR effect was Calibration slope 0.052 (95% CI -0.386 to 0.513; P=.764; q=.988); AUC 0.0037 (95% CI -0.0582 to 0.0586; P=.904; q=.988); Brier 0.0035 (95% CI -0.0036 to 0.0103; P=.312; q=.988); IPA – 0.0043 (95% CI -0.0174 to 0.0059; P=.452; q=.988). Interaction contrasts, rather than differences between subgroup-specific P values, defined effect modification.

Mechanism analysis

The mechanism decomposition showed that the China-to-England composition pathway contributed 0.258 (95% CI -0.044 to 0.548; P=.100) to the calibration-slope change, 0.0355 (0.0007 to 0.0668; P=.044) to AUC, 0.0033 (-0.0021 to 0.0095; P=.204) to Brier score, and – 0.0181 (-0.0812 to 0.0357; P=.452) to IPA. On the fixed risk set, the separate predictor-history and outcome-history calibration-slope effects were 0.061 and 0.029, respectively.

The mechanism decomposition showed that the England-to-China composition pathway contributed 0.129 (95% CI -0.065 to 0.330; P=.216) to the calibration-slope change, 0.0188 (-0.0058 to 0.0417; P=.136) to AUC, 0.0146 (0.0079 to 0.0207; P=.004) to Brier score, and – 0.0101 (-0.0196 to -0.0022; P=.020) to IPA. On the fixed risk set, the separate predictor-history and outcome-history calibration-slope effects were 0.062 and 0.033, respectively.

Discussion

Main findings

Among adults aged 60 years or older, time anchoring improved cross-country risk ranking but did not consistently improve absolute cardiovascular risk prediction. AUC increased in both transport directions, whereas Brier error increased materially in the England-to-China direction and IPA did not improve in either direction. Holding participants fixed removed the AUC gains, while changing the outcome history within the England-to-China common risk set still increased Brier error. The narrowest interpretation is that temporal reconstruction changed who entered validation more than it repaired the transported predictor–outcome relation.

The most informative result is therefore a divergence, not a uniformly better model. In the China-to-England direction, calibration slope and AUC improved without clear improvement in Brier score or IPA. In the England-to-China direction, the AUC gain coexisted with worse Brier error, showing that better ordering of participants did not yield better risk probabilities. Time anchoring strengthened temporal admissibility but did not by itself establish transportability.

Relation to previous evidence

Cross-national ageing studies depend on harmonised variables to make cohort comparisons operationally possible (Lee et al., 2021). External validation theory nevertheless separates case mix, baseline risk, and predictor effects as distinct sources of transport failure (Debray et al., 2015). The present analysis adds temporal provenance to that framework. A wave label described where a value was stored, but not necessarily when its information became observable.

The endpoint pattern also matters. Heart-disease histories changed more often than stroke histories, consistent with greater heterogeneity in what respondents and surveys classify as a heart problem. Self-reported cardiovascular ascertainment varies by endpoint even when the same questionnaire architecture is used (Sabanayagam et al., 2024). The endpoint interaction intervals, however, determine the inferential boundary: differences in revision frequency did not establish different transport effects with the available precision.

Mechanism: socio-biological selection through recorded disease history

The flip counts explain the cohort contrast before any social mechanism is invoked. Among 7,188 paired CHARLS baseline states, 340 changed from released negative to time-anchored positive and 303 changed in the opposite direction; 10.57% gross churn therefore produced only – 0.64% net displacement. Version D contains dispute pathways in both directions, so restoring the state observable at the earlier interview can both remove and return participants to the disease-free risk set (Phillips et al., 2021). Among 7,318 paired ELSA G.3 states, the corresponding counts were 308 and 22; the imbalance converted 5.89% churn into – 5.13% net displacement. The G.3 later-dispute code can denote a history subsequently set to no, so time anchoring restores the earlier positive state and preferentially removes those participants from disease-free eligibility (Wilkins et al., 2023). The two churn patterns are

therefore the arithmetic consequence of different observed dispute directions, not conflicting estimates of the same revision process.

Holding the participants fixed separated changes associated with within-person recoding from changes associated with sample replacement. In the China-to-England direction, most of the AUC gain disappeared on the common risk set, while predictor-history and outcome-history recoding each produced small, imprecise changes. In the England-to-China direction, risk-set composition accounted for most of the Brier deterioration and generated the negative IPA shift, while the fixed-sample joint effect partly offset that IPA loss. Within the fixed participants, outcome-history recoding explained the increase in Brier error. The transported coefficients therefore changed little for the same people; the larger performance movements arose because temporal anchoring changed the joint distribution of predictors, recorded disease, and subsequent events in the people being evaluated. Conditioning on a disease-free state that is jointly influenced by latent disease and its probability of being recorded can alter predictor–outcome associations through selection (Hernán et al., 2004).

The divergence between AUC and probability accuracy follows from what each metric measures. In China-to-England transport, observed cardiovascular risk changed little while mean predicted risk increased, so participant ordering improved but average overprediction did not. In England-to-China transport, mean predicted risk was almost unchanged while observed risk increased, so the baseline-risk gap widened even as rank ordering improved. AUC measures ordering, whereas the Brier score combines calibration and discrimination in the predicted probabilities (Steyerberg et al., 2010). Calibration slope measures the spread of predicted log odds, while calibration-in-the-large measures their level; improving one does not repair the other (Van Calster et al., 2019). This is why slopes moved toward 1 but Brier error and IPA did not improve, and why the England-to-China model continued to underpredict absolute risk.

The selection variable itself is socio-biological rather than purely administrative. Social conditions become embodied through cumulative material, behavioural, and physiological pathways before cardiovascular disease is named in a survey (Krieger, 2005). Healthcare contact then governs diagnostic opportunity, so similar latent disease can generate different recorded histories when access and recognition differ (Andersen, 1995). Recall and question context add another filter between experienced illness and the harmonised response (Schwarz, 1999). The larger revision of heart-problem histories than stroke histories is consistent with endpoint-specific variability in self-reported cardiovascular ascertainment (Sabanayagam et al., 2024). Time anchoring consequently selects on embodied disease burden after it has passed through healthcare access, diagnostic naming, memory, and release-specific coding.

This explanation remains a mechanism hypothesis rather than demonstrated mediation. The decomposition did not include socioeconomic position, healthcare use, biomarkers, symptom severity, or linked clinical diagnoses, and the cohort contrast also bundles calendar period and questionnaire architecture. A direct test should model baseline-state revision as an outcome of these social, healthcare, and biological variables, then repeat the transport decomposition after standardising

diagnostic opportunity across cohorts. That design would show whether the composition pathway reflects embodied cardiovascular risk, differential recognition, survey-memory processes, or release logic rather than treating all four as one country effect.

Strengths and limitations

The study traced one measurement threat from phenotype construction through eligibility, event timing, model fitting, and reciprocal external validation. The factorial common-risk-set design separated predictor and outcome histories, paired household resampling preserved within-household dependence, multiple imputation retained participants with missing covariates, and formal interaction contrasts avoided comparing significance across subgroups. The sample-composition contrast converted an implicit analytic consequence into an explicit mechanism estimand.

The principal limitation is that temporal states were reconstructed from harmonised values and dispute flags rather than raw question routing and interview timestamps. The analysis establishes which information was admissible within version D and version G.3, not which diagnosis was clinically correct. Estimates are release-specific and should not be combined with counts from a different harmonised release. Censoring began at the first unobserved interval, missed-wave returnees were excluded, measured blood pressure was unavailable on an equivalent schedule, and no adjudicated cardiovascular records were used.

Bootstrap inference incorporated household sampling and stochastic predictor imputation, but 500 replicates do not define equivalence. Confidence intervals spanning zero indicate insufficient evidence of a contrast, not equality. Sex and endpoint analyses had less precision than the overall comparison, and the cohort contrast bundles country, calendar period, healthcare context, survey wording, and release-specific algorithms. It cannot be interpreted as a causal country effect.

Implications and future research

Cross-cohort analyses should audit temporal provenance before declaring a baseline disease-free sample. Reporting only net sample-size change is inadequate because two-way churn can be substantial even when the final counts are similar. Researchers should present released, reconstructed, and common risk sets; show how event intervals change; and evaluate calibration, discrimination, and prediction error together.

The next study should reconstruct diagnosis histories from raw questionnaire modules, retain the interview at which each component became observable, and link survey histories to adjudicated records. Missed-wave returnees should be incorporated with censoring weights based only on prior information. Repeating the decomposition across later ELSA releases would show whether the composition pathway is stable or is itself a version-specific artefact.

Conclusion

Temporal discordance between released and time-anchored harmonised disease histories was empirical rather than hypothetical. Composite cardiovascular status changed for 643/7,188 CHARLS participants (8.95%) and 330/7,318 ELSA G.3 participants (4.51%); the corresponding gross risk-set churn was 10.57% and 5.89% (Table 2). There was no evidence that the transport effect of time anchoring differed by recorded sex in either direction, although phenotype revision itself was associated with sex in CHARLS and the cohort-by-sex contrast was non-null. Released calibration slopes were far below the ideal value of 1: 0.206 (95% CI -0.131 to 0.479, which included zero) for China-to-England and 0.371 (0.021 to 0.656) for England-to-China. Time anchoring moved the point estimates to 0.573 (0.225 to 0.808) and 0.586 (0.244 to 0.845), respectively. Only the China-to-England AA-minus-RR change was clearly supported by its bootstrap interval, and neither absolute slope approached ideal calibration (Table 3; Fig. 4). In England-to-China transport, observed cardiovascular risk exceeded mean predicted risk under both histories (0.112 versus 0.087 under RR and 0.137 versus 0.088 under AA), with positive calibration-in-the-large values of 0.281 and 0.510. This persistent baseline-risk offset indicates systematic underprediction despite better discrimination (Debray et al., 2015). Temporal anchoring therefore corrected part of the temporal-ordering problem without resolving cross-country probability transport; raw-question timing, target-baseline-risk recalibration, and adjudicated outcomes are required before clinical use.

Abbreviations

- AA
time-anchored predictor and outcome histories
- AR
time-anchored predictor and released outcome histories
- AUC
area under the receiver operating characteristic curve
- CHARLS
China Health and Retirement Longitudinal Study
- CI
confidence interval
- CVD
cardiovascular disease
- ELSA
English Longitudinal Study of Ageing
- HCO
heteroskedasticity-consistent covariance estimator
- IPA
index of prediction accuracy
- RA
released predictor and time-anchored outcome histories

RR

released predictor and outcome histories.

Declarations

Ethics approval and consent to participate

CHARLS was approved by the Biomedical Ethics Review Committee of Peking University (IRB00001052-11015), and participants provided written informed consent. ELSA wave 4 was approved by the National Hospital for Neurology and Neurosurgery and Institute of Neurology Joint Research Ethics Committee (07/H0716/48), wave 5 by the Berkshire Research Ethics Committee (09/H0505/124), and wave 6 by the NRES Committee South Central–Berkshire (11/SC/0374). ELSA participants provided informed consent. This secondary analysis used de-identified data under the applicable data-use conditions and involved no new participant contact or data collection.

Consent for publication

Not applicable.

Availability of data and materials

ELSA data are available through the UK Data Service under its registration and licensing terms (Wilkins et al., 2023). CHARLS data are available through the CHARLS portal under its data-use agreement (Phillips et al., 2021). Licensed participant-level data cannot be redistributed; the accompanying package contains code and aggregate outputs only.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' contributions

HL conceived the study, curated the data, developed the methodology and software, performed the analysis, prepared the visualisations, and drafted the manuscript. QL supervised the study, interpreted the findings, and critically revised the manuscript. Both authors read and approved the final manuscript.

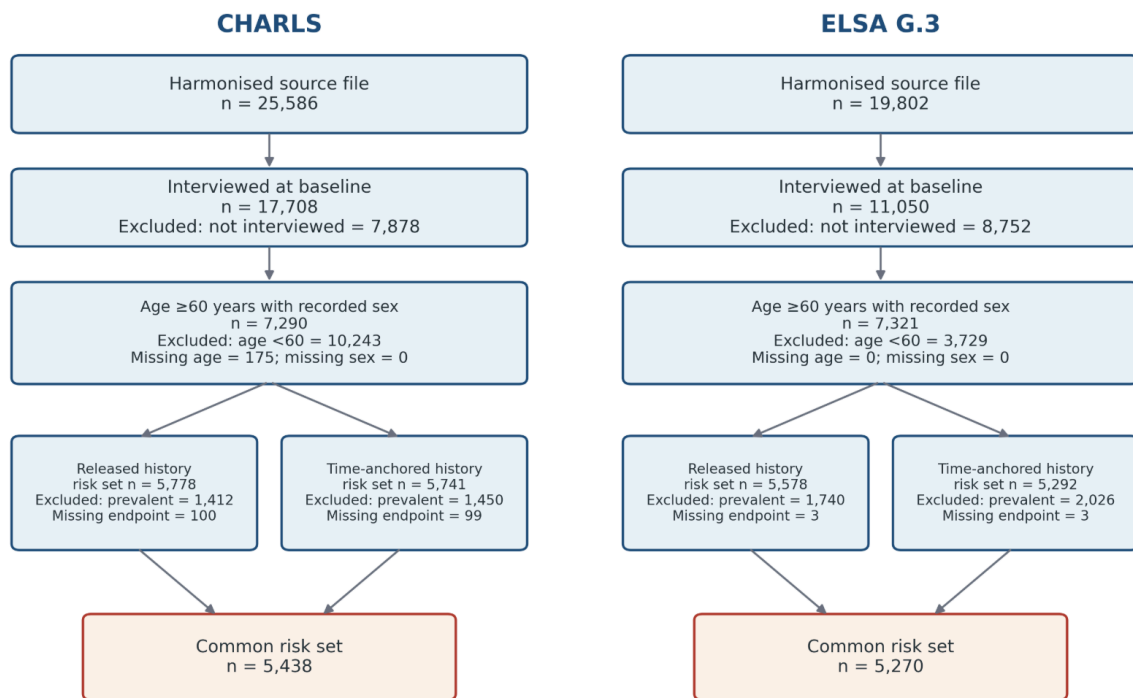
References

1. Andersen RM. Revisiting the behavioral model and access to medical care: Does it matter? *J Health Soc Behav.* 1995;36(1):1–10. <https://doi.org/10.2307/2137284>.
2. Bartlett JW, Hughes RA. Bootstrap inference for multiple imputation under uncongeniality and misspecification. *Stat Methods Med Res.* 2020;29(12):3533–46. <https://doi.org/10.1177/0962280220932189>.
3. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *J Clin Epidemiol.* 2015;68(2):134–43. <https://doi.org/10.1016/j.jclinepi.2014.11.010>.
4. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol.* 2015;68(3):279–89. <https://doi.org/10.1016/j.jclinepi.2014.06.018>.
5. Heidari S, Babor TF, De Castro P, Tort S, Curno M. Sex and Gender Equity in Research: Rationale for the SAGER guidelines and recommended use. *Res Integr Peer Rev.* 2016. <https://doi.org/10.1186/s41073-016-0007-6>, 1, Article 2.
6. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology.* 2004;15(5):615–25. <https://doi.org/10.1097/01.EDE.0000135174.63482.43>.
7. Kivimäki M, Batty GD, Singh-Manoux A, Britton A, Brunner EJ, Shipley MJ. Validity of cardiovascular disease event ascertainment using linkage to UK hospital records. *Epidemiology.* 2017;28(5):735–9. <https://doi.org/10.1097/EDE.0000000000000688>.
8. Krieger N. Embodiment: A conceptual glossary for epidemiology. *J Epidemiol Commun Health.* 2005;59(5):350–5. <https://doi.org/10.1136/jech.2004.024562>.
9. Lee J, Phillips D, Wilkens J, Gateway to Global Aging Data Team. Gateway to Global Aging Data: Resources for cross-national comparisons of family, social environment, and healthy aging. *Journals Gerontology: Ser B.* 2021;76(Suppl):S5–16. <https://doi.org/10.1093/geronb/gbab050>.
10. Li J, Lin S, Pei H, Xie G, Pei L, Chen G. Age-period-cohort analysis of cardiovascular disease trends in middle-aged and older adults: Cross-country comparison across HRS, ELSA, SHARE, and CHARLS. *J Global Health.* 2025;15:04260. <https://doi.org/10.7189/jogh.15.04260>.
11. Phillips D, Green H, Petrosyan S, Shao K, Wilkens J, Lee J. (2021). Harmonized CHARLS documentation: Version D (2011–2018). Gateway to Global Aging Data. <https://g2aging.org/>.
12. Sabanayagam C, He F, Chee ML, Cheng CY. Validity and reliability of self-reported prevalent and incident cardiovascular disease among Asian adults. *J Cardiovasc Dev Disease.* 2024;11(11):350. <https://doi.org/10.3390/jcdd11110350>.
13. Schwarz N. Self-reports: How the questions shape the answers. *Am Psychol.* 1999;54(2):93–105. <https://doi.org/10.1037/0003-066X.54.2.93>.
14. Steptoe A, Breeze E, Banks J, Nazroo J. Cohort profile: The English Longitudinal Study of Ageing. *Int J Epidemiol.* 2013;42(6):1640–8. <https://doi.org/10.1093/ije/dys168>.
15. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: A framework for traditional and novel measures.

Epidemiology. 2010;21(1):128–38. <https://doi.org/10.1097/EDE.0b013e3181c30fb2>.

16. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group Evaluating Diagnostic Tests and Prediction Models of the STRATOS Initiative. Calibration: The Achilles heel of predictive analytics. BMC Med. 2019;17., Article 230. <https://doi.org/10.1186/s12916-019-1466-7>.
17. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. J Clin Epidemiol. 2008;61(4):344–9. <https://doi.org/10.1016/j.jclinepi.2007.11.008>.
18. Wilkens J, Wang Y, Rebellato G, Oh Y, Lee J. (2023). Harmonized ELSA documentation: Version G.3 (2002–2019). Gateway to Global Aging Data. https://doc.ukdataservice.ac.uk/doc/5050/mrdoc/pdf/5050_harmonized_elsa_g3_2002-2019.pdf
19. Zhao Y, Hu Y, Smith JP, Strauss J, Yang G. Cohort profile: The China Health and Retirement Longitudinal Study (CHARLS). Int J Epidemiol. 2014;43(1):61–8. <https://doi.org/10.1093/ije/dys203>.

Figures



Predictor missingness did not trigger exclusion; smoking, education, hypertension, and diabetes were multiply imputed within each cohort and history-defined risk set.

Figure 1

Selection of the released, time-anchored, and common cardiovascular disease-free risk sets in CHARLS and ELSA G.3. Boxes show sequential eligibility criteria, exclusions, definition-specific native risk sets,

and the common risk set used in fixed-sample sensitivity analyses. Predictor missingness did not trigger exclusion.

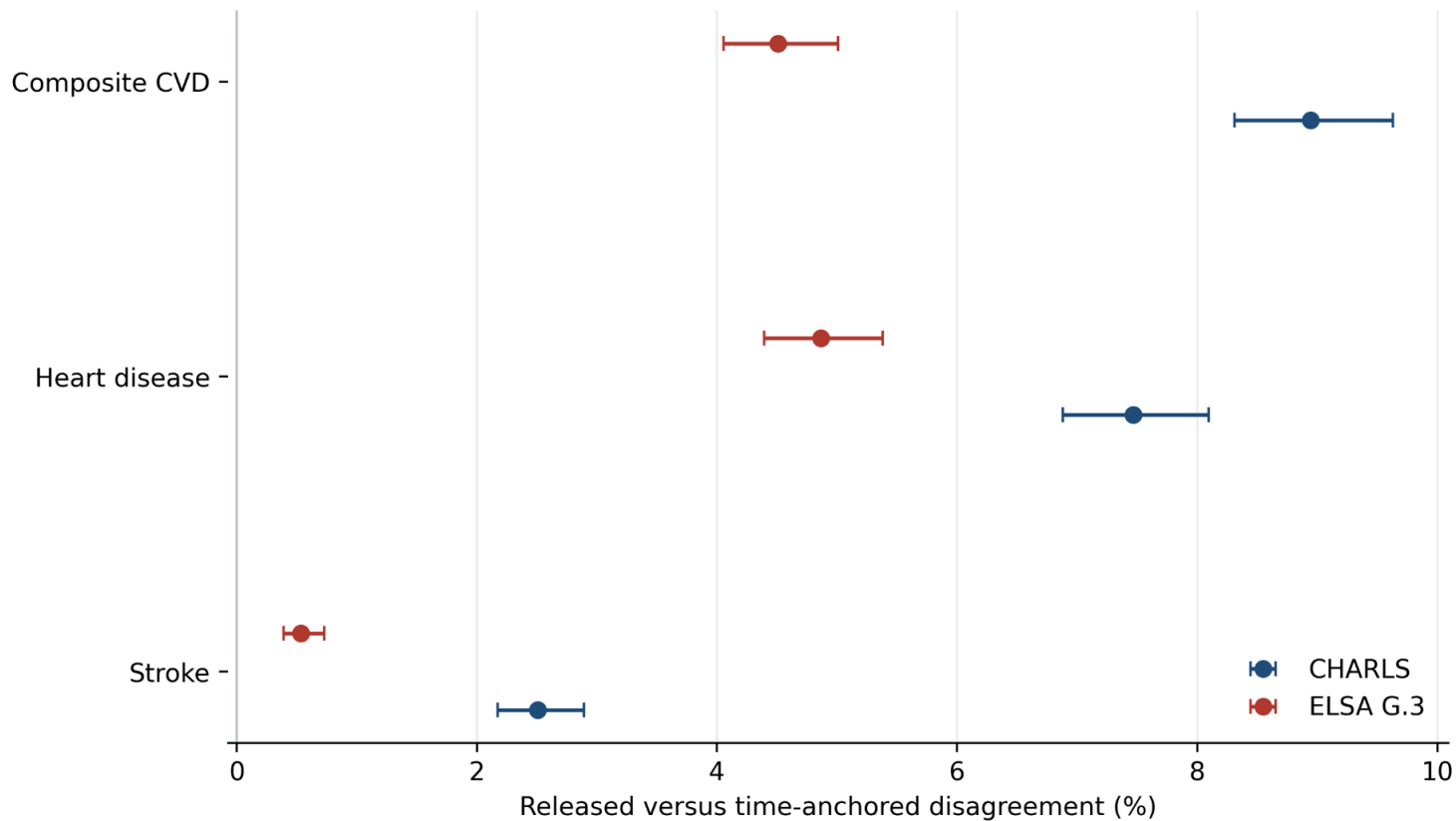


Figure 2

Baseline-state revision differed by cohort and endpoint. Points denote the percentage of participants whose released and time-anchored baseline states differed; horizontal bars denote Wilson 95% confidence intervals. Denominators ranged from 7,178 to 7,218 in CHARLS and were 7,318 in ELSA G.3. CVD indicates cardiovascular disease.

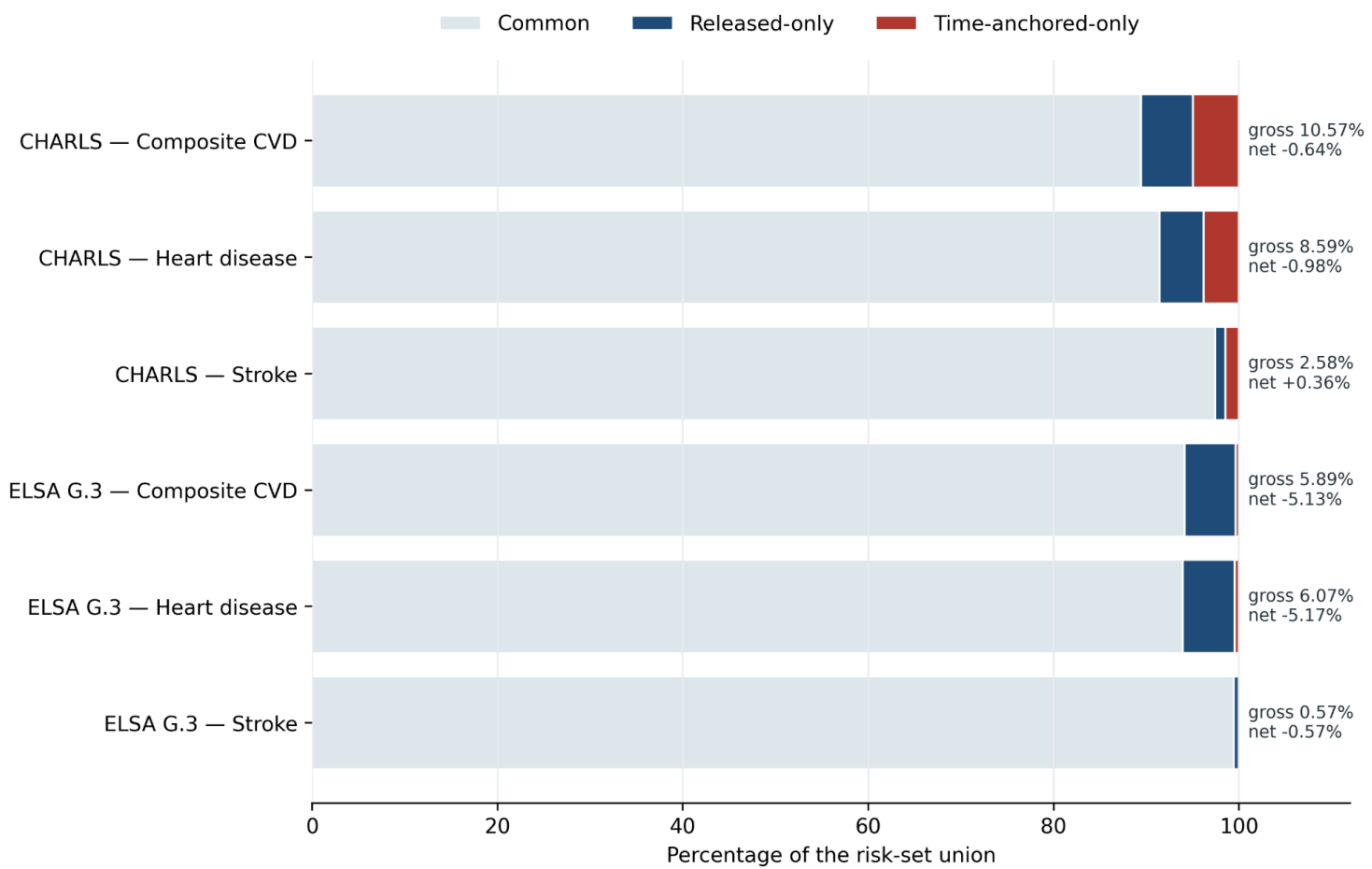


Figure 3

Gross risk-set turnover exceeded net displacement. Bars partition the risk-set union into common, released-only, and time-anchored-only groups. Gross churn divides the symmetric difference by the union. Net displacement divides the risk-set size difference by the released risk-set size.

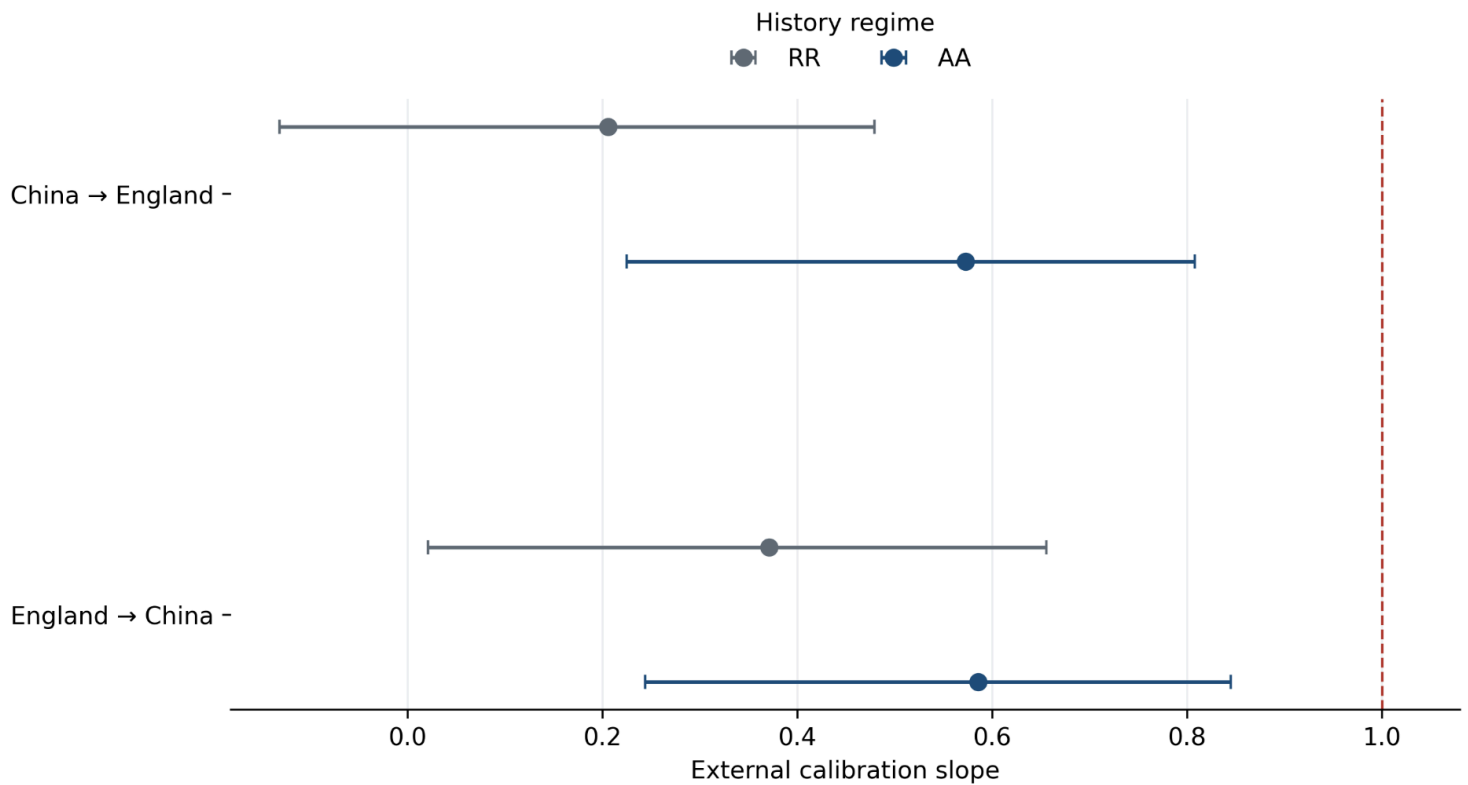


Figure 4

Full-pipeline external calibration-slope point estimates moved toward 1 but remained below 1. Points show 20-imputation estimates; horizontal bars show percentile 95% confidence intervals from 500 paired household bootstrap replicates. Each replicate independently resampled households in both cohorts, generated two stochastic predictor imputations, refitted the source model, and evaluated the transported model. RR denotes the released pipeline; AA denotes the time-anchored pipeline. The dashed line marks the ideal slope of 1.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [additionalfile1.zip](#)