Rentosertib and geroprotective endpoints - an AI-assisted reanalysis and critical assessment

Sándor Szalma

Abstract

Two linked publications have proposed that proteomic aging clocks can serve as geroprotective endpoints in conventional drug trials. A phase 2a randomised trial of rentosertib, a TNIK inhibitor, in idiopathic pulmonary fibrosis reported a mean 12-week forced vital capacity (FVC) gain of +98.4 ml in its highest-dose arm2; a companion analysis then applied six published proteomic clocks to serum from 42 of the same participants and reported biological-age reductions of 2.7 to 3.5 years, proposing that aging and disease endpoints can be assessed simultaneously1. We recomputed every quantity we cite from the deposited supplementary tables and source data. Published headline values reproduce exactly, including the 21 of 54 significant clock comparisons, the contrast of 326 treated against 2 placebo proteins, and the four protein–FVC correlations. Six findings do not survive reanalysis. First, no between-arm contrast is reported for the trial's lead efficacy endpoint; forming it from the published intervals gives +118.7 ml (95% CI -3.4 to 240.8, P = 0.056), and the arms differ at baseline by 0.88 standardized units of FVC in the direction of the claim. Second, all 54 clock comparisons are one-sided despite a Methods statement that tests were two-sided; the two-sided recomputation leaves 13 of 54 significant. Third, the 326-versus-2 contrast compares an interaction term with a main effect and depends on a per-term false-discovery partition; pooling the four compared terms gives 222 versus 12. The associated "mirror-image" enrichment between placebo and treated arms is weaker than a no-effect simulation of the same coding predicts (observed r = -0.622, -0.617, -0.563; simulated null median -0.690). Fourth, all six clocks were trained on UK Biobank EDTA plasma and applied to serum with plate-median normalization and, in the authors' own words, "no bridging or anchoring controls", so absolute biological ages are not on a calibrated scale. Fifth, a sensitivity analysis removes one placebo patient whose 12-week FVC gain was about +526 ml and one 30 mg once-daily patient, on a baseline-quality criterion stated only in a figure legend, which moves the placebo comparator down by 42 ml and grows the headline difference by 35 to 45% while leaving the highest-dose arm untouched. Sixth, none of the six clocks has a published test–retest reliability; at n = 10 the smallest detectable shift is almost exactly the test–retest standard deviation, so the reported effects are interpretable only if that unmeasured quantity is below about 2.7 to 3.5 years. We set out separately what the two studies do establish, and what a defensible geroprotective-endpoint trial would require.

ark:24975/genrxiv-2026-00006

Rentosertib and geroprotective endpoints - an AI-assisted reanalysis and critical assessment

Preprint · not peer reviewed · submitted to GenRxiv

Sándor Szalma (ORCID 0000-0001-7591-3902) Inxyte Informatics, Carlsbad, CA, USA Correspondence: sszalma@earthlink.net

17 September 2026

Abstract

Two linked publications have proposed that proteomic aging clocks can serve as geroprotective endpoints in conventional drug trials. A phase 2a randomised trial of rentosertib, a TNIK inhibitor, in idiopathic pulmonary fibrosis reported a mean 12-week forced vital capacity (FVC) gain of +98.4 ml in its highest-dose arm[2]; a companion analysis then applied six published proteomic clocks to serum from 42 of the same participants and reported biological-age reductions of 2.7 to 3.5 years, proposing that aging and disease endpoints can be assessed simultaneously[1]. We recomputed every quantity we cite from the deposited supplementary tables and source data. Published headline values reproduce exactly, including the 21 of 54 significant clock comparisons, the contrast of 326 treated against 2 placebo proteins, and the four protein–FVC correlations. Six findings do not survive reanalysis. First, no between-arm contrast is reported for the trial's lead efficacy endpoint; forming it from the published intervals gives +118.7 ml (95% CI −3.4 to 240.8, P = 0.056), and the arms differ at baseline by 0.88 standardized units of FVC in the direction of the claim. Second, all 54 clock comparisons are one-sided despite a Methods statement that tests were two-sided; the two-sided recomputation leaves 13 of 54 significant. Third, the 326-versus-2 contrast compares an interaction term with a main effect and depends on a per-term false-discovery partition; pooling the four compared terms gives 222 versus 12. The associated "mirror-image" enrichment between placebo and treated arms is weaker than a no-effect simulation of the same coding predicts (observed = −0.622, −0.617, −0.563; simulated null median −0.690). Fourth, all six clocks were trained on UK Biobank EDTA plasma and applied to serum with plate-median normalization and, in the authors' own words, "no bridging or anchoring controls", so absolute biological ages are not on a calibrated scale. Fifth, a sensitivity analysis removes one placebo patient whose 12-week FVC gain was about +526 ml and one 30 mg once-daily patient, on a baseline-quality criterion stated only in a figure legend, which moves the placebo comparator down by 42 ml and grows the headline difference by 35 to 45% while leaving the highest-dose arm untouched. Sixth, none of the six clocks has a published test–retest reliability; at n = 10 the smallest detectable shift is almost exactly the test–retest standard deviation, so the reported effects are interpretable only if that unmeasured quantity is below about 2.7 to 3.5 years. We set out separately what the two studies do establish, and what a defensible geroprotective-endpoint

trial would require.

1. Introduction

Drugs developed for age-related disease might, in principle, modify aging itself, and conventional trial endpoints are not designed to detect that. Aging clocks — models that map a molecular profile onto a predicted age or mortality risk — are the natural candidate instrument, and proteomic clocks are an attractive subclass because plasma proteins are proximal effectors rather than regulatory marks. The proposal that such clocks can be read out alongside ordinary efficacy endpoints in an ordinary trial is therefore consequential for how geroprotector development is designed and funded.

That proposal was recently made concrete. A randomised, double-blind, placebo-controlled phase 2a trial of rentosertib (ISM001-055), a generative-AI-discovered TNIK inhibitor, enrolled 71 patients with idiopathic pulmonary fibrosis across four arms and reported safety as its primary endpoint, with FVC among its secondary endpoints[2]. A companion analysis applied six published proteomic aging clocks — ProtAge[3], PAC[4], OrganAge in its chronological and mortality-trained forms[5], ipfP3GPT[6] and PAOPAC[7] — to serum from 42 consenting participants at four visits, and reported reductions in predicted biological age of 2.7 to 3.5 years under treatment[1].

We appraised both papers. Our purpose is not to dispute that the trial measured something real, nor that proteomic age acceleration is a valid construct in the cohorts where it was developed; section 10 sets out what we think the two papers establish, and it is not a short list. Our purpose is to ask whether the specific inferences drawn survive recomputation from the deposited data, because the proposal at issue — that a clock shift in a small trial can be read as evidence of geroprotection — will be imitated, and the methodological standard set here will be inherited.

We lead with the problems, because that is where the information is. Six of them are, in our view, material: they change a reported number, a reported direction, or the class of claim the data can support. We then separate the findings by kind, since "this is wrong", "this is undisclosed" and "this cannot be checked from what was supplied" carry different weight and conflating them weakens the argument.

2. Methods

2.1 Data provenance

All recomputations use only material published with the two papers. For the clocks analysis[1] we used the Supplementary Tables workbook (41587_2026_3286_MOESM3_ESM.xlsx), specifically S4 (Mann–Whitney comparisons of ΔBioAge between each treated arm and placebo, 54 rows), S6 (continuous-time linear mixed-effects model coefficients for 2,836 proteins), S7

(the categorical-time version of the same model) and S10–S13 (the preranked lists used for gene-set enrichment, one per arm). For the trial[2] we used the Source Data workbook (41591_2025_3743_MOESM3_ESM.xlsx), specifically the sheet underlying Fig. 4d, which carries per-patient change in NPX and change in FVC for COL1A1, FAP, FN1 and MMP10 with treatment-arm labels ($n = 41$ with both values present). Sheet header rows are preceded by comment lines; the header was located per sheet rather than assumed. Where a quantity could only be obtained from the printed text we say so explicitly and label it approximate.

Analyses were run in Python 3.11 with pandas 2.3.3, NumPy, SciPy 1.17.1 and openpyxl 3.1.5. The complete recomputation script and its tabular output are available on request; every number quoted below is read from that output rather than transcribed.

2.2 Reproduction checks

Before testing any criticism we reproduced the published headline values, on the principle that a criticism is only as good as the analyst's demonstrated ability to recover the original result. The following reproduce: the count of 21 significant clock-level comparisons out of 54; the 326 proteins with treatment-altered trajectories against 2 in placebo; the 2,841 assays retained after Olink quality-control exclusion; and all four pooled protein–FVC correlations in the trial (COL1A1 $r = -0.42$, $P = 0.0061$; FAP $r = -0.41$, $P = 0.0070$; FN1 $r = -0.24$, $P = 0.1383$; MMP10 $r = -0.52$, $P = 0.0005$).

2.3 Recovering the multiple-comparison family and threshold

Neither paper states the significance threshold used for the clock analyses. We recovered it by bracketing: the largest q value flagged significant in S4 is 0.0848 and the smallest unflagged value is 0.1065, so the threshold is $q < 0.10$. The same bracketing on S6 gives 0.0997 and 0.1033, again $q < 0.10$. Elsewhere the enrichment analyses use the $q < 0.25$ default of the GSEA implementation. We identified the false-discovery partition by recomputing Benjamini–Hochberg q values[12] within each candidate partition of the 54 tests — all pooled, by clock, by timepoint, by arm, and by clock × timepoint — and comparing each against the shipped adjusted values. A single pooled family over all 54 tests reproduces the shipped values to 1.7×10^{1}; no other partition comes close. In the protein-level models, by contrast, each model term carries its own adjusted column, so BH was applied within each term separately.

2.4 Testing the sidedness of the reported comparisons

S4 reports the Mann–Whitney U statistic, both group sizes and the P value for every comparison, which makes the sidedness decidable rather than a matter of interpretation. For each row we reconstructed the normal-approximation one- and two-sided P from U, n and n with a tie correction and compared

both against the reported value. All 54 reported P values match the one-sided reconstruction, with a maximum absolute deviation of 6.8×10 ; the two-sided value is exactly twice the reported one in every case. To recompute the result set two-sided we doubled each P, re-applied BH within the same pooled family, and applied the same $q < 0.10$ threshold, changing nothing else.

2.5 The null correlation induced by a shared reference group

Both papers rest an argument on placebo and treated arms showing opposite patterns. In a model with a reference group, each treated estimate is an interaction term — already a difference from placebo — while the placebo estimate is the corresponding main effect. A contrast written as (arm − reference) shares −Var(reference) with the reference term, so the two are negatively correlated even when the treatment has no effect whatever. Analytically, (null) $= -\sqrt{}$(n(arm) / (n(arm) + n(ref))), which for arm sizes of 8 to 14 against a reference of 11 gives −0.65 to −0.75.

Because the quantity actually correlated in the paper is not a raw coefficient but a ranking metric, we also simulated the null for the exact statistic. We first identified the metric used to build the preranked lists: among the candidate constructions we tested, coefficient × −log (P) × 1000 reproduces the shipped lists at Spearman 1.0000 with a median ratio of exactly 1000. This is a nonstandard choice — the conventional preranking statistic is the signed test statistic or the log fold change — and it lets significance enter the ranking twice, once through the coefficient's precision and once through the explicit P-value factor. We then drew, for each of 400 replicates, independent standard-normal placebo and treated slopes for 2,836 proteins with no treatment effect present, formed the interaction as (treated − placebo) with standard error $\sqrt{2}$, applied the same coefficient × −log (P) metric, and recomputed the Spearman correlation (seed 20260917). The resulting null has median $= -0.690$ with a 2.5th to 97.5th percentile range of −0.710 to −0.672.

2.6 The between-arm FVC contrast

The trial reports a mean 12-week FVC change of +98.4 ml (95% CI 10.9 to 185.9) in the 60 mg once-daily arm and −20.3 ml (95% CI −116.1 to 75.6) in placebo, with no contrast between them. Per-patient FVC is not deposited, so we back-calculated each arm's standard error from its published interval and arm size (SE = (upper − lower) / 2t(0.975,n−1)), combined them as an unpaired difference, and referred the result to a t distribution on n + n − 2 degrees of freedom. This is an approximation and we label it as such: it assumes the intervals are symmetric t intervals on the arm means, which is what the figure legend states. We did not digitize the ANCOVA-with-imputation panel, so we form no contrast for it.

2.7 Partial correlations for the protein–FVC association

The trial reports four protein–FVC correlations computed across all arms pooled. Because treatment simultaneously lowers these proteins and is the arm in which FVC rises, a pooled correlation is a between-arm contrast rather than a within-person association. We therefore recomputed each correlation three ways from the deposited per-patient data: as published (Pearson, pooled across arms, $n = 41$); adjusted for arm, as the partial correlation after residualising both variables on treatment-arm indicator variables, with the P value referred to $n - k - 1$ degrees of freedom; and within the placebo arm alone ($n = 10$), which is the only stratum in which the association can be observed free of treatment.

2.8 Detectability of the reported clock shifts

To express what the absence of reliability data costs, we asked what within-person variability would still permit the reported effects to reach significance. For a paired comparison in n participants, a mean shift is significant at two-sided $= 0.05$ when $|\text{mean}| \,/\, (\text{SD(change)}/\sqrt{n})$ exceeds $t(0.975,n-1)$; with independent measurement error, $\text{SD(change)} = \sqrt{2} \times \text{SD(test–retest)}$. At $n = 10$ this makes the smallest detectable shift $1.01 \times \text{SD(test–retest)}$ — that is, almost exactly equal to the test–retest standard deviation. We report this as a paired-t approximation to the Wilcoxon tests actually used, for scale rather than as a restatement of the published analysis.

2.9 What we did not do

We did not digitize any figure; every value we cite is either recomputed from a deposited table or quoted from the printed text with its provenance stated. We did not refit the linear mixed-effects models, because the per-protein data are not deposited — our criticisms of those models concern the coding of their coefficients and the multiplicity partition applied to them, both of which are decidable from the shipped coefficient tables. We did not attempt to obtain unpublished plate layouts, clock weights or protocol documents. We assessed no model assumptions beyond those our specific arguments require.

3. Design and inferential validity

3.1 The clock substudy is a post hoc, consent-based, completers-only subset

The clock analysis was performed on 42 of the trial's 71 randomised patients, recruited by consent obtained after randomisation, and restricted to those with samples at the relevant visits. Randomisation guarantees exchangeability of the arms as randomised; it guarantees nothing about a subset selected afterwards on a post-randomisation variable. No comparison of the consenting and non-consenting participants is provided, and no prespecified hypothesis, threshold or analysis plan for the clock endpoints is cited. The analysis is exploratory in

construction, and we think that is the strongest defence available for several of the results below. It is not, however, how the abstract reads: clock reductions are reported as detected, and a simultaneous-assessment capability is proposed on their basis.

3.2 The lead efficacy claim has no between-arm contrast, and baseline favours the treated arm

The trial's FVC result is presented as two per-arm intervals side by side: +98.4 ml in the 60 mg once-daily arm and −20.3 ml in placebo. Two intervals that do not overlap zero in opposite directions are not a treatment effect, and the contrast is absent from the abstract, the Results, the figures and the supplement. Forming it from the published intervals gives +118.7 ml (95% CI −3.4 to 240.8), P = 0.056 (Fig. 1a). We do not claim this refutes the drug's effect on FVC; a 12-week phase 2a trial is not powered for it. We claim that the quantity the reader is invited to infer is not significant at the conventional level, and that this is the quantity the paper should have reported. It is also worth setting the effect against the measurement scale: back-calculating from the per-arm intervals, the between-patient standard deviation of the 12-week FVC change is about 141 ml in the 60 mg once-daily arm and 183 ml in placebo, against an ATS/ERS within-session repeatability tolerance for FVC of 150 ml[13]. A 98 ml mean shift remains detectable by averaging across patients, but it is smaller than the visit-to-visit dispersion of a single acceptable spirometry session.

The comparison is further complicated at entry. The 60 mg once-daily arm began with a mean FVC of 2,762 ml against 2,246 ml in placebo, a standardized difference of 0.88, and with 50.0% versus 17.6% of patients on background nintedanib (Fig. 1b). Both differences run in the direction of the efficacy claim, and the clock analyses apply no baseline adjustment.

Absence of baseline adjustment matters more for the clock endpoints than for FVC, because the clock arms are compared on a change score in a small sample. When groups differ at baseline on the measured quantity, change scores are subject to regression to the mean in the direction of the higher-baseline group, and the paper reports that IPF patients start with elevated biological age. Neither an adjusted analysis nor an argument for why one is unnecessary appears.

3.3 A post hoc exclusion moves the comparator and not the treated arm

Two patients were removed from a sensitivity analysis of the FVC endpoint: one from placebo and one from the 30 mg once-daily arm, on the grounds that they "exhibited >600 mL difference between screening and baseline FVC measurements, making uncertain the baseline FVC values in those patients". Excluding patients whose baseline is unreliable is defensible, and presenting the analysis alongside the primary one is the right instinct. Three features of how it was done are not.

The criterion appears only in a figure legend. It is stated in the legend to Extended Data Fig. 2 and nowhere else — not in the Methods, not in the statistical-analysis section, and with no indication that the rule or the threshold was set before the data were seen. A reader of the main text encounters the citation "(Fig. 2 and Extended Data Fig. 2)" with no signal that the two figures analyse different sets of patients.

The exclusion is confined to the comparator and to the weakest dose arm. The 30 mg twice-daily and 60 mg once-daily arms are unchanged at +19.7 ml and +98.4 ml, because no patient was removed from either. The placebo mean moves from −20.3 ml (n = 14) to −62.3 ml (n = 13), and the 30 mg once-daily mean from −27.0 ml (n = 13) to −37.5 ml (n = 12). The apparent difference between the highest dose and placebo therefore grows from +118.7 ml to +160.7 ml on observed cases, and from +97.4 ml to +141.0 ml under ANCOVA with imputation — a 35 to 45% increase in the headline quantity produced by removing two of 71 patients.

The excluded placebo patient was the placebo arm's largest gain. This follows from the two printed means without any digitization: a 14-patient mean of −20.3 ml and a 13-patient mean of −62.3 ml imply that the removed patient had a 12-week FVC change of about +526 ml. The patient removed from the 30 mg once-daily arm had about +99 ml. So the rule, as applied, deleted a placebo responder five times larger than the treatment effect the abstract reports, and nothing comparable from the arms that carry the efficacy claim.

We are not suggesting the exclusion was chosen for its effect; a >600 ml discrepancy between screening and baseline is a real data-quality problem and the patients were identified by a criterion that has nothing to do with outcome. But the criterion is applied to a *baseline* discrepancy and its consequence falls on the *outcome* in one direction, which is exactly the situation that prespecification exists to protect against. What the paper should carry is the rule in the Methods, a statement of when it was fixed, and a note that the excluded analysis differs from the primary one only in the comparator.

There is a further reading of the same fact that has nothing to do with the exclusion. Two of 71 patients showed screening-to-baseline FVC discrepancies exceeding 600 ml — four times the ATS/ERS within-session repeatability tolerance[13] — and one placebo patient gained about 526 ml over 12 weeks. Both observations say that the between-visit dispersion of this endpoint is large relative to a 98 ml mean treatment effect, which is the point of section 3.2 made from the data rather than from the intervals.

3.4 Two mechanistic claims dissolve when treatment arm is accounted for

The trial reports that changes in fibrosis-associated serum proteins correlate with change in FVC, and that drug exposure correlates with response. Both are computed with the arms pooled. Because rentosertib lowers these proteins

and the treated arms are where FVC rose, a pooled correlation measures the between-arm difference, not a within-person relationship between protein change and lung-function change.

Recomputing from the deposited per-patient data shows this directly (Fig. 1c). COL1A1 falls from r = −0.42 (P = 0.0061) to r = −0.17 (P = 0.30) after adjustment for arm, and to r = −0.26 (P = 0.46) in placebo alone. MMP10, the strongest of the four as published, falls from r = −0.52 (P = 0.0005) to r = −0.28 (P = 0.09) and then to r = +0.01 (P = 0.97). FAP falls from r = −0.41 to r = −0.10 and then to r = −0.01. FN1 changes sign. None of the four associations is significant within the placebo arm, and none is significant after adjustment for arm. The same pooling drives the exposure–response claim: drug exposure is zero in placebo by construction, so any endpoint difference between arms will correlate with exposure across the pooled sample.

Fig. 1 | The trial's efficacy and mechanistic claims under reanalysis. a, The paper reports per-arm 12-week FVC changes with 95% confidence intervals but no between-arm contrast; the contrast formed from those intervals (dark) is +118.7 ml (95% CI −3.4 to 240.8), P = 0.056. Arm sizes are the observed-case numbers. **b**, Baseline standardized differences, 60 mg once daily versus placebo; all three run in the direction of the efficacy claim. **c**, The four published protein–FVC correlations recomputed as published (pooled across arms), adjusted for treatment arm, and within the placebo arm alone. Asterisks mark $P < 0.05$. Panels a and b use values printed in the trial paper; panel c is recomputed from the deposited per-patient source data (n = 41, of whom 10 received placebo).

4. Statistical specification

4.1 All 54 clock comparisons are one-sided, against the paper's own Methods

The clocks paper states that tests were two-sided unless otherwise specified. They were not. Every one of the 54 reported P values in Supplementary Table S4 matches the one-sided Mann–Whitney reconstruction from the shipped U statistic and group sizes, to within 6.8×10 ; the two-sided value is exactly double the reported one throughout. This is not a matter of statistical taste. A one-sided test in this setting encodes the assumption that treatment cannot increase biological age, which is precisely the question at issue, and it is applied to an exploratory analysis of a drug with no prior geroprotective evidence.

Substituting the two-sided value and changing nothing else — same pooled BH family, same $q < 0.10$ threshold — leaves 13 of 54 comparisons significant rather than 21 (Fig. 2a). The abstract's sub-claims move with it: the week-4 arm in which five of six clocks detected a reduction becomes three of six, and the week-12 results fall from 6 of 18 to 2 of 18. Eight of the 21 published positives are one-sided artefacts, and they are not concentrated in one clock or arm.

4.2 The false-discovery family is partitioned by model term

In the protein-level models, Benjamini–Hochberg correction is applied within each model term separately, so each of the terms gets its own family of 2,836 tests. This looks conservative and is not, because BH is not monotone in family composition: what a test's q value becomes depends on which other tests it is pooled with. When the four quantities the manuscript compares against one another — the Time main effect and the three Group × Time interactions — are placed in one family of 11,344 tests, the treated-versus-placebo contrast falls from 326 versus 2 to 222 versus 12 (Fig. 2b), or from 163-fold to 18-fold. The same asymmetry is present in the categorical-time model, where the corresponding counts are 310 versus 1.

We are not asserting that the pooled family is the correct one; the choice is genuinely arguable. We are asserting that the 163-fold figure is a property of the partition as much as of the biology, that the partition is not stated as a choice, and that a reader cannot tell from the paper that a defensible alternative reduces the contrast by an order of magnitude.

4.3 Six correlated clocks are treated as independent corroboration

The paper's recurring argument is agreement: a reduction "detected by five of six clocks" reads as five independent instruments concurring. They are not independent. All six were trained on UK Biobank Olink data, they share large fractions of their input features from the same 2,841 assays, and they are weighted sums over those features. The paper itself reports that the chronological clocks agree with each other at $r = 0.78$ to 0.89 while correlating with the mortality-trained clocks at only $r = 0.25$ to 0.61 — which is an argument against averaging the two families, not for it. Counting concordant clocks is closer to counting correlated tests of one quantity than to independent replication.

Two further specification issues belong here briefly. Effect sizes are estimated as Cohen's d from roughly ten observations per arm, where the sampling distribution of d is wide and no interval is reported. And non-significant comparisons are read as evidence that no effect is present — in particular to argue that placebo did not move — when at these sample sizes the analysis has little power to detect a shift of the size claimed for treatment.

4.4 An undisclosed sensitivity analysis is shipped with the paper

Supplementary Table S4 contains a complete parallel analysis in columns prefixed *excl_*, together with a *concordance* column classifying each of the 54 comparisons. Under it the placebo arm drops from 11 to 9 participants and the treated arms from 9–11 to 8–10, 19 of 54 comparisons are significant rather than 21, and four comparisons change status: ProtAge at week 2 in 30 mg twice daily (lost), PAC at week 2 in 30 mg once daily (gained), OrganAge(mortality) at week 12 in 30 mg once daily (lost), and PAOPAC at week 4 in 30 mg twice

daily (lost). The last of these sits in the week-4 cell that carries the abstract's five-of-six claim.

No exclusion criterion, and no mention that this analysis exists, appears in the manuscript, the Methods, the figure legends or the table legends. We stress the category of this finding: it is not evidence of a wrong result. Nineteen against 21 is a modest difference and the analysis may well have been a routine robustness check. It is an undisclosed analytic choice affecting a headline number, which a reader can only discover by opening the workbook.

Fig. 2 | The clock-level statistics under reanalysis. a, All 54 arm × time-point × clock comparisons. Filled markers remain significant when the reported one-sided P is replaced by its two-sided value within the same pooled Benjamini–Hochberg family at $q < 0.10$; open red markers are significant only one-sided. **b**, Proteins with treatment-altered trajectories against the placebo comparator, under the published per-term false-discovery partition and under a single pooled family over the four compared terms. **c**, Spearman correlation between the placebo and treated arms' preranked ranking metrics (red lines), against the distribution obtained from 400 simulated replicates in which no treatment effect is present and only the shared-reference coding operates (histogram). All three observed values are weaker than the null median of −0.690.

5. A reference-coding artefact under two headline claims

5.1 The 326-versus-2 contrast is not a like-for-like comparison

The linear mixed-effects models are specified with placebo as the reference group and Group × Time interaction terms for the treated arms. Under that coding each treated coefficient estimates how that arm's trajectory differs from placebo, while the Time main effect estimates the trajectory in placebo itself. The 326 proteins are counted from the interaction terms and the 2 from the main effect. These are different estimands: the first is a contrast, the second is a level. A protein that moves identically in all four arms contributes to the main effect and not to any interaction; a protein that moves only relative to placebo contributes to the interactions and not to the main effect. There is no *Group_placebo:Time* column in the shipped table because the coding cannot produce one — which is the diagnostic feature of the problem rather than an omission.

The sentence the reader takes away — rentosertib altered 326 proteins "compared to only 2 in the placebo group" — therefore compares a quantity that is by construction a difference from placebo against a quantity that is by construction placebo's own movement. Combined with the per-term multiplicity partition of section 4.2, the 163-fold contrast is best read as a property of the model parameterisation and the FDR bookkeeping rather than as a biological ratio.

5.2 The mirror-image enrichment is what the coding predicts

The same coding carries a second, more striking claim. The enrichment analyses use preranked lists built per arm, and placebo and treated arms show approximately opposite enrichment patterns — a mirror image that the paper reads as treatment reversing a placebo-arm aging signature. Because the treated lists are built from interaction coefficients and the placebo list from the main effect, the two are negatively correlated by construction, with no treatment effect required.

We quantified this two ways. Analytically, the induced null correlation for these arm sizes is -0.65 to -0.75. By simulation of the exact statistic under no treatment effect, the null has median $= -0.690$ (2.5th to 97.5th percentile -0.710 to -0.672). The observed correlations are $= -0.622$ (30 mg once daily), -0.617 (30 mg twice daily) and -0.563 (60 mg once daily) (Fig. 2c). All three are *weaker* in magnitude than the coding artefact alone predicts. On this evidence the mirror image requires no biological explanation at all: it is the expected consequence of how the ranking metric is constructed, and the observed values are, if anything, slightly less extreme than pure noise through the same pipeline would give.

5.3 The ranking metric compounds the problem

The preranked lists are not built on a conventional statistic. We identified the metric as coefficient $\times$ $-$log (P) $\times$ 1000, which reproduces the shipped lists at Spearman 1.0000. Standard practice preranks on the signed test statistic or the log fold change, precisely so that significance enters the ordering once. Multiplying an effect estimate by the negative logarithm of its own P value lets precision enter twice and gives the ranking a heavy tail driven by the best-powered assays. Since enrichment scores depend on the ordering, this choice propagates into every enrichment claim, and it is not described or justified in the Methods.

6. Assay and clock transfer

A fixed-weight model applied outside the conditions in which its weights were estimated returns a number whose scale is not guaranteed. Four features of the assay chain, all taken from the clocks paper's own Methods, bear on whether the reported biological ages are on the training scale (Fig. 4a).

6.1 The specimen matrix changes between training and application

All six clocks were trained on UK Biobank EDTA plasma. This is verifiable in the source publications rather than assumed: PAC describes 2,923 plasma proteins from the UK Biobank Pharma Proteomics Project[4], and ProtAge's cohorts are EDTA-plasma aliquots[3]. The trial substudy measured serum: "Serum samples were collected from all participants at baseline, week 2, week 4 and week 12."

Serum and plasma are not interchangeable inputs to a fixed linear model. Coagulation consumes fibrinogen and related factors while platelet activation during clotting releases granule contents into the supernatant, so the difference between matrices is protein-specific in both sign and magnitude rather than a constant offset. A clock is a weighted sum over hundreds to thousands of such proteins with weights estimated in plasma; applying it to serum reweights a systematically altered input vector.

The paper acknowledges the *population* transfer explicitly: the models "were trained originally with UK Biobank data composed of mostly healthy people" and the trial is described as an "out-of-scope case". It nowhere acknowledges the *matrix* transfer. A keyword scan of the full text finds no sentence in which serum and plasma are discussed together, and no adjustment, bridging study or sensitivity analysis for it.

6.2 The scale is anchored to the trial cohort, without bridging controls

NPX values were produced by the standard two-step Olink procedure: extension normalization against each sample's extension control, then intensity normalization in which, for each assay on each plate, the median across the samples on that plate is subtracted. The Methods then state, in one sentence: **"No bridging or anchoring controls were applied."**

That sentence is the crux. Intensity normalization sets the zero point of every assay to the median of whichever samples sit on the plate — here, 42 IPF patients at up to four visits. The clocks' weights and scalers were estimated against a different zero point, and two of the six implementations then apply externally fitted transformations on top: ProtAge inputs are min–max scaled with scaler objects fitted on UK Biobank reference data and centred on population medians, and the organ-specific models rescale by standard deviations taken from their source publication's supplementary table. An externally fitted scaler applied to cohort-internally centred data is a mismatch: the affine map the model expects is not the one the data have been put through. Bridging or anchoring samples are the standard remedy and were not used. The consequence is that absolute biological ages, including the headline 2.7- to 3.5-year reductions, are not interpretable on the training-population year scale.

A second consequence is specific to treatment studies. Because the subtracted quantity is the median across samples for each assay, it is not a fixed constant: if a drug moves an assay in a substantial share of the samples on a plate, the plate median moves with it and the normalization removes part of the effect being estimated. With 168 samples (42 participants at four visits) and 326 proteins reported as shifting under treatment, this is not a negligible configuration. Whether it matters depends entirely on the plate layout — how many plates, whether a participant's four visits were kept together, and whether treatment arms were balanced across plates. None of this is reported, and it is information the authors hold.

6.3 About 91 training assays are absent, with no imputation rule

After removing assays annotated EXCLUDED by Olink quality control, 2,841 proteins remained. Elsewhere the Methods state that the enrichment background "consisted of 2,832 proteins shared by both UK Biobank and phase 2a trial datasets". Against the 2,923 proteins of the panel on which the clocks were trained, about 91 training assays are therefore unavailable. The arithmetic uses only numbers the authors themselves print.

For a fixed-weight model this is a live question rather than a technicality: a missing feature must be imputed, dropped with the remaining weights rescaled, or set to a reference value, and each choice shifts the output differently. The word "impute" does not appear in the paper and no feature-matching procedure is described for any of the six clocks. Whether a given clock lost inputs, and how many, cannot be determined from the published material.

6.4 What this does and does not undermine

Fairness requires separating four quantities, because they are not equally exposed.

- **Within-person change is partly protected.** Any per-assay offset constant across a participant's four visits — which includes a pure serum-versus-plasma shift, and a plate offset if all four visits sat on one plate — cancels in the ΔBioAge difference.
- **Absolute biological age is not protected.** The 2.7- to 3.5-year reductions and the baseline finding that IPF patients are biologically older than their chronological age are read off a scale whose zero was set by this cohort, through an externally fitted scaler, from a different specimen matrix.
- **Cross-matrix comparisons are not protected.** The paper compares the trial's serum coefficients against UK Biobank plasma aging trajectories, which is how its aging-versus-disease question is adjudicated. That comparison runs directly across the matrix boundary with no bridging.
- **The cancellation fails wherever plate composition tracks visit or arm.** If visits were distributed across plates, or arms unbalanced across them, the offsets differ between a participant's timepoints and do not cancel. The unreported plate layout is what turns this from a resolved issue into an open one.

7. Serial use of clocks without a reliability envelope

This is, in our view, the deepest of the five material concerns, because it bears not on one analysis but on the paper's central proposal. A within-person change is interpretable only against the variation the score shows when nothing has changed.

7.1 No clock has a published test–retest reliability

None of the six source publications reports a test–retest reliability, an intraclass correlation, or any repeat-measurement study of its score (Fig. 3). ProtAge comes closest with a panel showing that protein–age association coefficients are stable across three UK Biobank visits in 1,085 participants, but that is the stability of a population-level relationship, not the repeatability of an individual's score. Nor has the class been validated across assay platforms: all six are Olink-trained, none demonstrates transfer to SomaScan, and an independent head-to-head of other proteomic clocks found that published weights often cannot be transferred across platforms at all[11].

Before this trial, five of the six clocks had never been applied to repeat samples from the same person. The exception is ProtAge, in a 12-week supervised exercise study in 26 men in which the score fell by the equivalent of about ten months while most constituent proteins were stable, with the authors attributing the change to a small number of proteins[9]. That is a useful demonstration that the score moves on intervention timescales; it is not a reliability estimate.

The cost of this gap can be stated numerically. At $n = 10$, the smallest detectable within-person shift is $1.01 \times$ the test–retest standard deviation, so the reported 2.7- to 3.5-year effects are distinguishable from measurement noise only if the test–retest standard deviation is below about 2.7 to 3.5 years (Fig. 4b). Whether it is, nobody — including the authors — currently knows.

7.2 Pre-analytical variability is neither controlled nor reported

The entire pre-analytical description in the clocks paper is that serum was collected at baseline and weeks 2, 4 and 12, and that "samples were processed according to standardized protocols and stored until analysis". Time of day of the draw, fasting state and posture are not reported, and there is no statement that a participant's repeat visits were drawn at a consistent hour.

The magnitude at stake is measurable. In a controlled study of healthy adults with venous draws every three hours across 24 hours, 138 of 523 quantified plasma proteins — about 26% — showed significant diurnal oscillation, with the rhythmic set enriched for liver- and platelet-derived proteins and for haemostasis, immune-signalling and metabolic pathways[8]. Those are the same acute-phase, coagulation and matrix compartments that carry much of the weight in proteomic clocks, and platelet-derived proteins are also the group most disturbed by the serum-versus-plasma change of section 6.1.

This does not show that the reported shifts are collection-time artefacts. Randomisation protects the between-arm contrast against collection-time variation unless draw times differ systematically by arm, and there is no reason to expect that. What it shows is that the within-person noise floor of a proteomic clock is unknown and plausibly of the same order as the reported effect, and that the trial did not collect the information needed to bound it. Reporting the collec-

tion window, and whether repeat visits were time-matched, would cost nothing and would materially change how a reader should weigh a 3-year shift.

7.3 An independent cohort has documented the relevant failure mode

The concern is not hypothetical. An independent longitudinal study of organ-specific proteomic clocks in paired samples a decade apart reports that age acceleration is only moderately stable over that interval, and — directly relevant here — that initiation of medication moves organ-specific clocks through the proteins the drug targets rather than through generalised organ aging[10]. That is precisely the alternative explanation for the rentosertib result, established in a cohort with no connection to any of the six clocks' developers.

Fig. 3 | Validation evidence for the six clocks, compiled from their own source publications. Filled markers indicate the property is established in the source publication; half-tone markers indicate partial evidence; open markers indicate absence; question marks indicate dimensions that could not be checked because the source is closed access (OrganAge) or the preprint full text was not retrievable (PAOPAC). The shaded block contains the three dimensions on which reading a 12-week within-person change depends, and it is empty for all six clocks. "Predicts a hard outcome" means demonstrated prediction of mortality or incident disease. "Within-person serial use" is scored partial where a source describes longitudinal training data or a single small intervention study, neither of which is a reliability estimate.

Fig. 4 | The transfer chain and the detectability boundary. a, The transformations between the data in which the clocks' weights and scalers were estimated and the quantity reported in the trial, with the four points at which calibration to the training scale is lost. Quoted text is from the clocks paper's Methods. **b**, The paired-comparison detectability boundary at n = 10: a mean within-person shift is significant at two-sided = 0.05 only above the line. Because the smallest detectable shift is 1.01 × the test–retest standard deviation at this sample size, the reported 2.7- to 3.5-year effects (shaded band) require a test–retest standard deviation below 2.7 to 3.5 years — a quantity that has never been measured for any of the six clocks.

8. Aging signal versus disease signal

The interpretive question on which the whole proposal rests is whether the clock movement reflects aging or the resolution of fibrosis. The clocks paper is candid that proteomic clocks alone cannot fully separate the two, and then offers three arguments that the signal is not merely antifibrotic. We do not think any of the three carries the weight placed on it.

8.1 The dose–response argument

The argument is that a purely pharmacodynamic signal should scale with dose and the clock response does not behave that way. But the clock response is not monotone in dose either: the 30 mg twice-daily arm shows the broadest response on several measures, including the protein-level counts, and the 60 mg once-daily arm does not lead. Non-monotonicity is equally compatible with a pharmacodynamic reading — exposure, receptor occupancy and downstream matrix turnover need not be linear in dose — and with noise at ten patients per arm. The observation does not discriminate between the hypotheses.

8.2 The variance-explained argument

The argument is that change in FVC explains little of the variance in change in biological age, so the clock is measuring something else. With roughly ten patients per arm, a low R^2 is what one expects whether or not the two are related: the sampling variability of a correlation at $n = 10$ spans most of the available range. An underpowered null result cannot establish dissociation. It is also the wrong comparator — FVC is a downstream functional endpoint measured with substantial visit-to-visit variability, so weak coupling to a serum protein score is unsurprising under either hypothesis.

8.3 The UK Biobank aging-direction argument

The argument is that the trial's protein changes run opposite to the direction those proteins take with age in UK Biobank. This is the most interesting of the three, and it is the one most exposed to section 6. The comparison places serum coefficients from 42 IPF patients, normalized to their own plate medians without bridging, against plasma aging trajectories from a different assay batch, matrix and population. It also has an unavoidable confound: many of the proteins that rise with age are acute-phase, matrix and fibrosis-associated proteins, which is exactly the set an antifibrotic lowers. Anti-fibrotic efficacy and anti-aging direction are not separable on this axis.

8.4 The clock contributors are the drug's own targets

The affirmative case for the pharmacodynamic reading is straightforward. LTBP2 is the top contributor to all six clocks, and the clock contributors and the top differentially expressed proteins are the same fibrosis and matrix set — SPP1, COL1A1, COL5A1, COMP, MMP10, FAP — which an antifibrotic is expected to lower. The arms also differ at entry along that axis (section 3.2), and the clock-level tests adjust for nothing. A reader is being asked to accept that a drug which demonstrably lowers fibrosis proteins, in patients selected for a fibrotic disease, moved a score whose largest weights are on fibrosis proteins, for reasons other than lowering fibrosis proteins.

One asymmetry in the reporting is worth naming. Extended Data contains organ-specific clock *increases* of +1.0 to +4.2 years, positive in all 18 treated

arm × timepoint cells for the kidney and lung models. These are larger than the reductions the abstract reports, they go unmentioned in the text, and a one-sided test in the direction of benefit cannot detect them by construction. Under a geroprotection reading they are difficult to accommodate; under a pharmacodynamic reading they are unremarkable.

8.5 A construct-validity step in the abstract

The abstract moves from a clock shift to geroprotective assessment. Even granting every statistical point, that step requires the clock to be a measure of aging rate rather than a correlate of aging-associated protein levels. Proteomic age acceleration predicts mortality and incident disease, which establishes that it carries prognostic information; it does not establish that lowering the score lowers risk, and no intervention study has shown that for any proteomic clock. Treating movement of a prognostic index as demonstrated modification of the process it indexes is the same inferential step that has repeatedly failed for surrogate endpoints in other fields, and it deserves explicit defence rather than assumption.

9. Undisclosed analyses and reporting inconsistencies

Grouped here are findings that do not change an estimate but affect what a reader can verify. We separate them from the preceding sections deliberately: they are failures of disclosure rather than of inference.

- **The FVC exclusion criterion of section 3.3 is stated only in an Extended Data figure legend**, with no Methods entry and no statement of when it was fixed.
- **The trial's primary endpoint receives no inferential analysis.** Safety was the primary endpoint, and the dose-graded pattern in serious adverse events of grade 3 or above — 11.1%, 22.2% and 38.9% across ascending regimens against 17.6% in placebo — is reported without any test. We recomputed the comparison and it is not significant at $n = 18$ per arm, which is worth stating: the absence of a test is the finding, not a concealed positive.
- **Secondary endpoints are narrated selectively.** One nominal Leicester Cough Questionnaire result is emphasised while discordances between the observed-case and imputed analyses of DLCO and 6-min walk distance go unaddressed.
- **The significance threshold is undocumented and inconsistent** — $q < 0.10$ for the clock and protein analyses, $q < 0.25$ for enrichment, neither stated as a choice.
- **Internal inconsistencies between Results, Methods and Discussion.** Seven were identified, including an arithmetic error in the stated number of per-arm comparisons.

- **Two clocks cannot be independently checked at all.** OrganAge's source publication is closed access and the PAOPAC preprint full text was not retrievable, so statements about those two rest on their abstracts.

10. What the two studies establish

A critique that never concedes anything is not a critique but an attitude, and several of the things these papers get right are the things a sceptical reader would assume were wrong. We checked them and they hold.

10.1 The reported numbers are what they say they are

Every headline count we tested reproduces exactly from the deposited tables. The 21 of 54 significant clock comparisons, the 326-versus-2 protein contrast, the 2,841 assays after quality control, and all four protein–FVC correlations recompute to the published values. The Benjamini–Hochberg adjustment in the protein-level analyses is exactly what it claims to be: recomputed q values match the shipped adjusted values to machine precision. In the trial, the per-arm differential-expression counts reproduce once duplicated gene rows are collapsed, and the unmentioned placebo comparator turns out to be clean — zero significant proteins at every threshold we tested, so the omission conceals nothing.

10.2 The hardest statistical question was handled correctly

The multiplicity objection to a six-clock, three-arm, three-timepoint design is the obvious one, and the authors addressed it properly with a permutation test over 100,000 relabellings rather than with a parametric correction. That is the right instrument for the question and it does what it claims. The protein-level linear mixed-effects model is also properly specified for its purpose, with baseline NPX, age, sex and BMI as covariates and per-participant random intercepts. Our objection to the 326-versus-2 comparison is about the coding of its coefficients and the multiplicity partition applied afterwards, not about the model.

10.3 The trial's design and conduct hold up on the points we could check

The endpoints were prespecified: the registry record posted before enrolment completed[14] already contains every endpoint family the abstract names, and the apparent later narrowing occurred in revisions dated after database lock. Blinding is described adequately and maintained through data freeze, with one disclosed exception. The trial used two sources of spirometry equipment, which is not in itself a concern: within-subject device consistency was required, which eliminates the dominant threat, and what remains is a reporting gap rather than a bias. We note the same cancellation argument in the trial's favour here

as we grant to the clock analysis in section 6.4 — it would be inconsistent to accept it in one place and not the other.

10.4 The construct is real in the cohorts where it was developed

Proteomic age acceleration predicts mortality and incident disease, replicates across cohorts and platforms in independent hands, and for the chronological-age models predicts age with high accuracy: ProtAge reaches r = 0.94 in its UK Biobank test set and holds at r = 0.92 and r = 0.94 in two external cohorts[3], and ipfP3GPT reports a mean absolute error of 2.68 years in cross-validation[6]. The mortality-trained models are a different instrument and should not be read as ages, which PAC's own authors make clear[4]. None of this is in question. What the source literature does not support is the specific use made of these models here: reading a within-person change over 12 weeks in a small trial, on a different specimen matrix, without a reliability envelope.

10.5 The data sharing is better than the field's norm

It deserves saying that almost every criticism in this preprint was possible only because the authors deposited their statistics in full. The shipped tables contain the U statistics and group sizes that make the sidedness decidable, the per-term coefficients that expose the parameterisation, the preranked lists that let the ranking metric be identified, and the undisclosed exclusion analysis itself. A paper that reported only its conclusions would have been immune to this reanalysis. That is an argument for the authors' practice, not against it, and we would rather appraise a transparent paper than praise an opaque one.

11. Discussion

11.1 Reanalyses that would resolve most of this

Six analyses, all using data already in hand, would settle the majority of what is disputed above. None requires new samples.

- **Refit the protein-level models with arm-specific slopes** so that placebo and treated arms yield comparable estimands, and repeat the protein counts and the enrichment analysis on that parameterisation. This resolves sections 5.1 and 5.2 directly.
- **Report the clock comparisons two-sided**, with per-arm baseline biological age as a covariate, and state the false-discovery family and threshold as choices.
- **Disclose the Supplementary Table S4 exclusion analysis and its criterion**, and state which reported results depend on it.
- **Recompute ΔBioAge with the fibrosis and matrix features removed.** If the clock shift survives removal of the drug's own pharmacodynamic targets, the aging interpretation gains real support; if it does

not, that is equally informative. This is the single most decisive analysis available and it requires only the existing data.

- **Report the plate layout and the handling of absent clock features**, including how many inputs each clock lost and what was substituted.
- **Report the collection window** and whether repeat visits for a participant were time-matched.

One further study is needed and cannot be done retrospectively: a reliability envelope for these scores. It requires repeat draws from stable individuals, ideally at a fixed time of day, on the same assay and specimen matrix as the intended application, reported as an intraclass correlation and a smallest detectable difference for each clock. Until that exists, no proteomic clock shift in a small trial can be referred to a null distribution, because none has been published.

11.2 What a geroprotective-endpoint trial would require

The ambition behind these papers is sound: if aging-modifying effects exist, current trial designs will miss them, and a molecular endpoint that could be read alongside conventional outcomes would be genuinely valuable. The obstacle is not conceptual but metrological. An endpoint needs a measurement model before it needs a mechanism: a stated precision, a known reference range in the population of interest, demonstrated transfer to the assay and matrix in which it will be used, and a prespecified analysis. Proteomic clocks currently have none of these for serial use, and the four requirements are ordinary rather than onerous.

There is a specific design implication. A clock whose largest weights sit on the pathway the drug targets cannot distinguish geroprotection from pharmacodynamics in that indication, however well it performs as a population predictor. That is not a defect of the clock; it is a mismatch between instrument and application. A geroprotective-endpoint trial would want either a clock with no weight on the drug's target pathway, or a prespecified analysis with those features removed, or an intervention whose targets are not aging-associated matrix proteins. Choosing the instrument after seeing which clocks move is the failure mode to avoid.

11.3 How the results should be framed

Our reading of the evidence is that something real happened to the serum proteome of treated patients; that it is concentrated in fibrosis and extracellular-matrix proteins; and that the claim it constitutes a deceleration of aging is neither established by these data nor excluded by them. The pharmacodynamic explanation has not been ruled out, the absolute biological ages are not on a calibrated scale, and the exploratory framing that a post hoc, completers-only analysis of 42 patients supports is the one the data will bear. Framed that way — as a pharmacodynamic observation with an open aging interpretation

and a named list of analyses that would close it — the substudy is a useful contribution. Framed as detection of geroprotection, it sets a standard that the next twenty papers of this kind will inherit, and that standard is below what the measurements can support.

12. Limitations of this reanalysis

- **We could not refit the mixed models.** Per-protein data are not deposited, so our objections there concern coefficient coding and multiplicity partitioning rather than model fit, which we cannot assess.
- **The FVC contrast is approximate.** It is back-calculated from published intervals rather than computed from per-patient values, which are not deposited. The ANCOVA-with-imputation analysis has no contrast here at all, because we declined to digitize its panel.
- **Two of the six clocks could not be checked.** OrganAge's source publication is closed access and the PAOPAC preprint full text was not retrievable; statements about them rest on their abstracts and we have attributed no unverified figure to either.
- **Our no-effect simulation is a model, not the paper's design.** It reproduces the shared-reference coding and the identified ranking metric with independent normal errors and equal precision across arms. It does not reproduce the correlation structure among proteins, which would affect the width of the null distribution but not its location.
- **We did not assess modelling assumptions** beyond those our specific arguments require, and we did not attempt to obtain unpublished material.

Author contributions, competing interests and acknowledgements

Contributions S.Sz. designed the study, directed the analysis, and reviewed and approved the manuscript; he takes responsibility for its content. The analysis was executed and the manuscript drafted by Claude Science (Anthropic), an AI research assistant, under his direction. Consistent with ICMJE and COPE guidance, an AI tool cannot satisfy authorship criteria and its use is therefore disclosed here rather than credited as authorship. Every statistic in the manuscript is reproducible from the recomputation script, which is available on request, and was checked against the deposited tables. **Competing interests** The author declares no competing interests. The authors of the two papers examined here were not consulted in the preparation of this preprint and have had no opportunity to respond; the author would welcome correction of any point on which the deposited material has been misread.

Data and code availability

This reanalysis uses only material published with the two papers examined here: the Supplementary Tables and Source Data workbooks deposited with refs 1 and 2, and the proteomic data deposited by ref. 1 under accession OMIX008341. No new data were generated. The recomputation script (*recompute_all.py*), its tabular output (*reanalysis_results.csv*, one row per quantity with the method and reproduction status; *clock_comparisons_recomputed.csv*, all 54 comparisons with two-sided q values; *protein_fvc_correlations.csv*) and the figure-generating code are available on request.

References

1. Zhavoronkov, A., Galkin, F., Chen, S., Ren, F. et al. Integration of proteomic aging clocks in a phase 2a clinical trial supports simultaneous geroprotective assessment. *Nat. Biotechnol.* (2026). doi:10.1038/s41587-026-03286-y

2. Xu, Z., Ren, F., Wang, P. et al. A generative AI-discovered TNIK inhibitor for idiopathic pulmonary fibrosis: a randomized phase 2a trial. *Nat. Med.* **31**, 2602–2610 (2025). doi:10.1038/s41591-025-03743-2

3. Argentieri, M. A. et al. Proteomic aging clock predicts mortality and risk of common age-related diseases in diverse populations. *Nat. Med.* **30**, 2450–2460 (2024). doi:10.1038/s41591-024-03164-7

4. Kuo, C.-L. et al. Proteomic aging clock (PAC) predicts age-related outcomes in middle-aged and older adults. *Aging Cell* **23**, e14195 (2024). doi:10.1111/acel.14195

5. Goeminne, L. J. E. et al. Plasma protein-based organ-specific aging and mortality models unveil diseases as accelerated aging of organismal systems. *Cell Metab.* **37**, 205–222 (2025). doi:10.1016/j.cmet.2024.10.005

6. Galkin, F., Chen, S., Aliper, A., Zhavoronkov, A. & Ren, F. *Aging* **17**, 1999–2014 (2025). doi:10.18632/aging.206295

7. Xu, H., Chen, J., Chen, D., Mao, K. & Han, J.-D. J. Proteome-aware organ proxy aging clocks. Preprint. doi:10.64898/2026.04.24.720503

8. Jóhönnuson, E. M. et al. Diurnal rhythm of the human plasma proteome. *Clin. Proteomics* **22**, 29 (2025). doi:10.1186/s12014-025-09551-7

9. Lee-Ødegård, S., Argentieri, M. A., Norheim, F., Drevon, C. A. & Birkeland, K. I. *npj Aging* **12**, 19 (2025). doi:10.1038/s41514-025-00318-w

10. Longitudinal dynamics of organ-specific proteomic aging clocks over a decade of midlife. Preprint. doi:10.64898/2026.02.17.706320

11. Wang, Y. et al. *PLOS Med.* **21**, e1004464 (2024). doi:10.1371/journal.pmed.1004464

12. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).

13. Graham, B. L. et al. Standardization of spirometry, 2019 update. An official American Thoracic Society and European Respiratory Society technical statement. *Am. J. Respir. Crit. Care Med.* **200**, e70–e88 (2019). doi:10.1164/rccm.201908-1590ST

14. ClinicalTrials.gov. A study of ISM001-055 in subjects with idiopathic pulmonary fibrosis. NCT05938920.