

# Omitting tumor proliferation from prognostic gene expression models: a preregistered cross-cancer test, a transcriptome-wide measurement, and a survey of how often the omission is made

SOON WOO NAM

[drswnam@catholic.ac.kr](mailto:drswnam@catholic.ac.kr)

The Catholic University of Korea

---

## Research Article

**Keywords:** prognostic biomarkers, confounding bias, omitted variable, high-dimensional screening, non-collapsibility, preregistration, survival analysis, meta-research

**Posted Date:** October 7th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-11273325/v1>

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

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# Omitting tumor proliferation from prognostic gene expression models: a preregistered cross-cancer test, a transcriptome-wide measurement, and a survey of how often the omission is made

Soon Woo Nam

Hepatobiliary Unit, Division of Gastroenterology, Department of Internal Medicine, Incheon St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea

ORCID 0000-0001-5562-4775

## Corresponding author

Soon Woo Nam, Hepatobiliary Unit, Division of Gastroenterology, Department of Internal Medicine, Incheon St. Mary's Hospital, 56 Dongsu-ro, Bupyeong-gu, Incheon 21431, Republic of Korea. Telephone +82-32-280-7493. Fax +82-32-280-6010. Email drswnam@catholic.ac.kr

Registrations: doi:10.17605/OSF.IO/X5DCF, the cross-cancer test, registered before it was run; doi:10.17605/OSF.IO/N8DW6, the post hoc relationship reported here, registered for a confirmatory test in cohorts I have not examined.

## Abstract

**Background.** A variable associated with both the outcome and a candidate gene moves that candidate's estimate, most for candidates most strongly correlated with it, so a screen of many correlated candidates is exposed wherever one axis of biology moves the candidates and the outcome together. Tumor proliferation is the clearest such case. How often it is omitted, and what omitting it costs, are questions about analytic practice.

**Methods.** Fifty recent prognostic gene expression papers in hepatocellular carcinoma were sampled, and the 41 with a reachable full text were coded for whether their model adjusted for proliferation. For every expressed gene I then fitted two Cox models for overall survival, one adjusted for age, sex and stage, one adding a standardized proliferation score. The shift is the ratio of the two hazard ratios; the landscape is its log regressed on the gene's correlation with the score. That measurement is exploratory. A cross-cancer test of two predictions, with its inclusion rule and refutation thresholds, was registered in advance.

**Results.** None of those 41 adjusted for proliferation (0 percent, 95 percent CI 0.0 to 8.6), 23 reported proliferation activity in their high-risk group, and 28 did not state their covariates in the text. In hepatocellular carcinoma the landscape slope was -0.278 (R-squared 0.785), reproduced at -0.180, and the genes surviving false discovery control fell from 338 to 1. Ten cohorts met the registered rule; two were aggregates, so the eight sharing no patients are primary. The prediction that the slope would be negative in at least 90 percent of them was refuted, at four of eight, as was the prediction of gene-level agreement. Post hoc, a first-order omitted variable expression explained 0.84 of the variance of the per-gene shift, short of the 0.95 fixed in advance.

**Conclusions.** Adjusting for a covariate associated with both the outcome and the candidates moves nearly every estimate in a high-dimensional screen, by an amount the candidate's correlation with it predicts. Whether that removes bias depends on a causal structure these data cannot establish. Its direction tracks that association in the cohort at hand, which my registered prediction assumed away.

Keywords: prognostic biomarkers; confounding bias; omitted variable; high-dimensional screening; non-collapsibility; preregistration; survival analysis; meta-research

## Background

Prognostic gene expression studies follow a common procedure. Expression and survival are drawn from a public cohort, candidate genes or a score built from them are screened against outcome, and the survivors are carried into a multivariable Cox model adjusted for age, sex and stage. The procedure is cheap, it is reproducible in the narrow sense that another investigator with the same matrix will obtain the same numbers, and it supplies a large share of the prognostic biomarker literature[1].

Tumor proliferation is rarely one of those covariates, as the survey reported below shows. It is, however, an unusually strong candidate confounder. Venet and colleagues showed that most randomly chosen gene sets predict breast cancer outcome, and traced this to the proliferation signal that a random set of genes partly captures[2]. Starmans and colleagues showed that a knowledge-based proliferation signature carries prognostic value across many cancer types[3]. What those two papers establish is that proliferation is prognostic in many cancers and that it is captured, to a degree, by almost any set of genes. What they do not establish is how far a given gene's estimated association with outcome moves when proliferation is put into the model, how that movement varies from gene to gene, or whether it behaves the same way in different cancers. A covariate associated with outcome and with the expression of a candidate gene will move that candidate's estimate when it

enters the model. Whether the movement removes bias depends on the causal structure, and two associations alone do not establish that proliferation confounds a gene rather than mediating its effect or acting as a collider, a question I return to in the Discussion[4]. The size and the direction of the movement are set jointly by the covariate's association with outcome and by the sign of its correlation with the gene, so an omitted proliferation term biases a positively correlated gene upward only where proliferation is adverse. The downward case is the more consequential, because a gene whose apparent effect is small is rarely pursued.

The questions this paper asks are therefore about how this research is done rather than about any gene. The first is how often the adjustment is made at all, which is answerable only by reading what published studies report. The second is what the omission costs, which is answerable only by fitting the models both ways and measuring the difference. Establishing the second for one gene says little, because any gene can be an exception; establishing it for every expressed gene at once turns an argument into a measurement. The question is general to any high-dimensional prognostic screen in which one axis of biology moves both the candidate measurements and the outcome, and proliferation in cancer transcriptomics is the case in which that axis is best characterized, most easily measured, and most consistently left out.

This paper reports three things. First, I drew a sample of recent prognostic gene expression papers under a fixed rule and coded each for whether its survival model adjusted for proliferation. Second, in hepatocellular carcinoma, I fitted a survival model for every expressed gene with and without a proliferation score and described how the two estimates differ, using the models as an instrument for measuring the consequence of the omission. Third, I registered the prediction that the pattern found there would hold in every cancer type, and tested it. The registered prediction was refuted. The way it failed is the main result of this paper. No new patient data were generated, and the object of study is the analytic practice rather than the biology of any gene; public cohorts enter as the material on which that practice can be measured, and the only data I produced by hand are the coding of the sampled papers, which is deposited.

## **Methods**

### **Design and reporting**

This is an analysis of public, de-identified data, reported in accordance with the Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK)[5,6] and the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement[7]. No new human or animal data were generated, so ethical

approval and informed consent were not required. The transcriptome-wide landscape in the discovery and replication cohorts is exploratory and was designed after those data had been seen. Preregistration is used here in the sense set out by Nosek and colleagues[8], as a public record of the decision rules written before the data are analyzed. What was registered, before it was run, is the cross-cancer test described below (doi:10.17605/OSF.IO/X5DCF). Analyses designed after their data had been seen are labeled exploratory or post hoc throughout, in the text, in the tables and in the deposited code, and every script writes its own decision rules to a result file on each run, so that a rule edited after the fact is visible in the deposited output rather than only in the code history.

## **Literature survey**

The survey itself is post hoc: it was designed after the transcriptome-wide sweep described below had been seen. Its own design was fixed before any paper was read. PubMed was searched on 2026-09-20 for records with hepatocellular, a prognosis term, a gene expression term and Cox in the title or abstract, published from 2024 onward, in English. The search returned 426 records, 424 unique, which formed the frame, and 50 were drawn without replacement under a fixed seed. Two codings were kept apart. Whether a proliferation measure was a covariate is decidable from a full text, since a term absent from a paper cannot be one of its covariates. Each full text was searched for Ki-67, MKI67, mitotic count or index, and proliferation or cell cycle scores, and every occurrence was read in context. Whether grade was a covariate was coded only from what the text states, and papers whose covariate list appears only in a table were counted as not stating their covariates in the text, rather than having that list imputed from the table. Grade was never counted as an adjustment for proliferation. A sampled paper was eligible if it fitted a Cox model for a survival endpoint with gene expression, or a score built from it, as an explanatory variable. Proportions carry Wilson intervals.

## **Cohorts**

Expression, clinical annotation and overall survival for The Cancer Genome Atlas (TCGA) cohorts were obtained from UCSC Xena[9], using the HiSeqV2 matrices as  $\log_2(\text{normalized count} + 1)$ , the per-cohort clinical matrices, and the pan-cancer Clinical Data Resource survival table[10]. Samples were restricted to tumor samples, one per patient, then to complete data on age, sex, stage and survival. The sample-code filter admits primary, recurrent, additional primary and metastatic tumor codes, so the cohorts are not restricted to primary tissue and they differ in the mix of tumor material they contribute. I did not tabulate that mix by cohort, and a participant flow table is deposited for the discovery cohort only. The discovery cohort is the consortium's

hepatocellular carcinoma cohort[11]. GSE14520, an independent liver cohort[12], was obtained from the Gene Expression Omnibus with GEOquery[13].

### **The proliferation score**

The score is the mean of the gene-wise standardized expression of MKI67, TOP2A, CCNB1, PCNA, BUB1, CCNA2, AURKA and RRM2[14], computed within each cohort and then standardized. As a check on the choice of genes, the score was replaced by one built from HALLMARK\_G2M\_CHECKPOINT, a set in the hallmark collection[15] of the Molecular Signatures Database[16], keeping the 172 of its genes that belong to no Hoshida subclass template[17], so that the substitute measures proliferation without carrying subclass membership. When it was the adjustment variable, its own genes were excluded from the sweep on the same circularity ground as the eight. Because the score was developed in the liver, its behavior in every other cohort was examined post hoc through the mean pairwise correlation among the eight genes, their standardized Cronbach alpha, the share of variance taken by their first principal component, and their agreement with a score built from the 195 genes of that same hallmark set that are not among the eight. The log hazard ratio of the score was also corrected for attenuation by dividing it by the standardized Cronbach alpha of the eight genes.

### **The sweep and the shift**

A gene entered the sweep when its median and interquartile range in the analysis set were both above zero. The eight score genes were excluded, since adjusting a gene for a score containing it is circular. For each remaining gene, two Cox models were fitted for overall survival: one with the standardized expression, age, sex and stage, and one that added the standardized proliferation score. The shift is the exponential of the difference between the gene's coefficient in the second model and in the first, that is, the factor by which its hazard ratio is multiplied when proliferation enters. The landscape is the ordinary least squares regression of the log shift on the gene's Spearman correlation with the proliferation score. Within each cohort, gene-level p values were adjusted by the Benjamini-Hochberg procedure[18], and the adjusted values are reported as q across all swept genes, separately for each model. The discovery analysis in hepatocellular carcinoma also adjusted for histologic grade[19]; grade was not used elsewhere, because it is not recorded in every cohort.

### **What was registered, and what it predicted**

Stage entered every model as a binary variable, stage I or II against stage III or IV, because the finer categories are not comparably recorded across cohorts. The registration fixed the cohort inclusion rule, the covariate set,

the gene filter, the definition of the shift, the multiplicity correction, and the thresholds at which each hypothesis would be treated as confirmed, refuted or equivocal. It also required at least six of the eight score genes to be present in a cohort's matrix; every included cohort carried all eight. No cancer type was named in advance as one to be included. The rule was applied to every cohort with an expression matrix and a stage field, and a cohort entered when at least 200 patients and 60 deaths remained after restriction to complete cases; every exclusion is reported with its reason. Hepatocellular carcinoma, clear cell renal carcinoma and lung adenocarcinoma had been examined before the registration existed and were excluded from the confirmatory set by name. Sex was dropped from the model in any cohort containing only one sex, which applied to ovarian cancer alone.

The registered rule was applied to the cohorts as The Cancer Genome Atlas distributes them, and that distribution contains both single-organ and aggregate cohorts. Ten cohorts met the rule, but they are not ten disjoint groups of patients: COADREAD is COAD together with rectal adenocarcinoma, and LUNG is lung squamous carcinoma together with lung adenocarcinoma. Rectal adenocarcinoma had been excluded on its own for having 84 patients, and lung adenocarcinoma had been excluded by name, so each re-entered the confirmatory set inside an aggregate. My registration did not anticipate the aggregate cohorts. The primary analysis reported here is therefore the eight cohorts in which no patient appears twice, obtained by dropping the two aggregates, and the original ten are reported beside them as a sensitivity analysis. I identified this problem after the analysis had been run, and the choice of the eight is labeled post hoc wherever it appears.

Two hypotheses were registered: that the slope would be negative in at least 90 percent of included cancer types with a median R-squared of at least 0.50, refuted below 75 percent or below 0.30; and that the shift for a gene in one cancer type would agree with its shift in another, confirmed at a median pairwise Spearman correlation of at least 0.30 and refuted below 0.10.

### **Non-collapsibility**

A conditional hazard ratio rises when a prognostic covariate is added even in the absence of confounding[20,21]. To separate that arithmetic from confounding, the genes whose correlation with proliferation fell below 0.05 in absolute value were taken as a null band in which no confounding can operate, and within that band the log shift was regressed on the gene's own log hazard ratio through the origin. The constant of proportionality is the non-collapsibility inflation, and a zero coefficient must give a zero shift, so the intercept is not free. The interval on that coefficient is the one the ordinary least squares fit returns. It rests on

the same independence between genes that I reject for the landscape itself, so it is an optimistic interval and I use it only to indicate an order of magnitude. This estimate was compared with a simulation of 1,000 datasets drawn from the fitted adjusted discovery model with the proliferation score permuted, so that the score keeps its marginal distribution and its coefficient but is independent of the gene, leaving non-collapsibility alone[22]; the interval is the 2.5th and 97.5th percentiles across those simulations. Within the band the genes were also grouped into five bins by quintile of the absolute coefficient within each sign of the coefficient, with the median absolute log shift in each bin kept beside the median shift as a noise floor.

### **How far the omitted variable expression accounts for the shift**

This analysis is post hoc and was designed after the registered test had been seen. In a linear model, omitting a covariate moves a coefficient by the product of that covariate's own coefficient and the coefficient of the omitted covariate regressed on the retained ones. Applied one gene at a time, and with the shift defined as the adjusted coefficient minus the unadjusted one, this predicts a log shift of minus the product of the proliferation coefficient and the coefficient of the standardized gene in an auxiliary regression of the score on that gene and the other covariates. The expression is exact for linear models and only approximate for proportional hazards models, because the partial likelihood is not a linear projection[23,24]. I took the proliferation coefficient from a Cox model containing the score and the covariates but no gene, obtained the auxiliary coefficients for all genes at once by the Frisch-Waugh-Lovell residual form, and defined the residual as the observed log shift minus the predicted one. Two readings of the fits were fixed before the analysis was run, together with what to conclude if neither held: that the expression is an adequate description if it explains at least 0.95 of the variance of the observed shift with a slope within 0.05 of 1; that what it omits is non-collapsibility if the residual is proportional to the gene's own log hazard ratio with a coefficient close to the 0.064 estimated independently in the null band; and that if neither holds, the departure is reported as unexplained by either account. Whether that departure is confined to genes lying close to the score was asked next, in an analysis designed after the largest residuals had been read and labeled post hoc. Because both sides of the auxiliary regression are standardized, the auxiliary coefficient is on the scale of a partial correlation between the gene and the score given the covariates, so a threshold on its absolute value carries the same meaning in either cohort. Genes reaching 0.5, where a gene accounts for a quarter of the residual variance of the score and the two Cox models are close to indistinguishable, were excluded and the same quantities recomputed; thresholds from 0.3 to 0.9 are reported as description rather than as decision. Three readings were fixed before those fits: that the departure is attributable to such genes if the retained genes give a variance explained of at least 0.95 with a slope within 0.05 of 1; that it

is partly attributable to them if the variance explained rises by at least 0.05 without reaching 0.95; and that it is not attributable to them if the variance explained changes by less than 0.02. The landscape was refitted on the retained genes as well, so that the exclusion could be seen not to manufacture the relationship it is applied to.

### **Analyses added after the main results were seen**

Three further analyses are post hoc. The worked example was refitted on disease-specific survival and on the progression-free interval from the pan-cancer Clinical Data Resource[10]. It was refitted with a tissue composition score, defined within each cohort as the first principal component of the stromal and immune markers COL1A1, FN1, PDGFRB, CD3E, CD2, CD68, CSF1R, ITGAM, LYZ and HLA-DRA that are present on the platform and belong to no Hoshida subclass template. It was refitted again with the stromal and immune scores of the ESTIMATE method, which infers the non-tumor content of a bulk expression profile; on RNA-seq data the calibrated conversion of those scores into a purity fraction is not defined, so the combined ESTIMATE score entered the model as a covariate in its own units[25]. The reliability statistics were added one at a time to the regression of the landscape slope on the proliferation log hazard ratio, fitted across all thirteen cohorts. Three further rules were fixed before the analyses they govern were run, and all three are post hoc in design. The proportional hazards assumption was tested for every term of the worked example with scaled Schoenfeld residuals, and the estimate counted as robust only if the hazard ratio stayed at or above 1.25 with an interval excluding 1 under both corrections, the proliferation score entered as a stratification factor on its tertiles and the same score given a coefficient varying as the logarithm of time. The movement of the worked example's hazard ratio when proliferation entered was bootstrapped with 2,000 paired resamples, the same resample fitting both models. In the clinical subsets a variable recorded in fewer than 60 percent of a model set was described rather than modeled.

### **Use of an AI assistant**

An AI assistant (Claude, Anthropic) was used to help write and debug the analysis scripts, to cross-check the reported numbers against the result files, and to draft and edit the text. I reviewed and edited everything it produced and take full responsibility for the content. The study design, the registered hypotheses and decision rules, the interpretation and the conclusions are my own. A large language model does not satisfy the criteria for authorship and is not listed as an author.

### **Software and availability**

Analyses used R 4.6.1 with survival 3.8-6[26] under fixed seeds. Code, the registered rules, every result file and the run logs are deposited at Zenodo, version 2.4.0, doi:10.5281/zenodo.23051386; the concept identifier doi:10.5281/zenodo.22685965 resolves to the most recent version. Version 2.4.0 produced every number reported here.

## Results

### What the literature adjusts for

This survey is post hoc in timing, and its design was fixed before any paper was read. Of 50 sampled papers, all 50 were eligible and 41 had a reachable full text. None of the 41 included a proliferation measure as a covariate in its survival model (0 percent, 95 percent confidence interval, CI, 0.0 to 8.6). Twenty-three of the 41 (56.1 percent, 41.0 to 70.1) reported cell cycle or proliferation pathway activity as a property of their own high-risk group or gene set, through enrichment, gene set enrichment, or hallmark analysis. Twenty-eight of the 41 (68.3 percent, 53.0 to 80.4) did not state the covariates of their multivariable model anywhere in the text. Seven stated an adjustment for grade (Table 1).

### The landscape in hepatocellular carcinoma

What that omission costs is the next question. The discovery sweep adjusts for histologic grade, so it is fitted in the 339 patients who have a recorded grade, among whom there were 114 deaths. In that set 16,902 genes were swept. The shift a gene's hazard ratio took when proliferation entered the model was not random with respect to the gene. The log shift regressed on the gene's Spearman correlation with proliferation gave a slope of -0.278 with an R-squared of 0.785, and the decile medians descended without exception, from a shift of 1.078 in the decile least correlated with proliferation to 0.889 in the decile most correlated with it. In GSE14520, 225 patients and 10,861 genes gave a slope of -0.180 (R-squared 0.796), and the shift correlated at 0.560 between the two cohorts (Figure 1A, B). Replacing the eight-gene score with the 172-gene hallmark score gave -0.296 (R-squared 0.801) with a gene-level agreement of 0.885. These analyses are exploratory.

### What the shift costs

In the same cohort, 2,308 genes had p below 0.05 without a proliferation term and 1,333 had p below 0.05 with one. Under Benjamini-Hochberg control at 0.10 the counts were 338 and 1 (Figure 1C). That collapse should not be read as bias alone. Adding a covariate changes both the coefficient and its standard error, so the number

of surviving genes can move in either direction[27], and across the eight primary cohorts it fell in three, rose in two and was unchanged in three. What a loss of power does not do is move a point estimate, and of those 338 genes, 115 moved by at least 10 percent in the hazard ratio. One gene reached the threshold only with proliferation in the model, and it is not among the 338, so no gene survived the adjustment. The collapse itself is a property of this cohort, in which proliferation is strongly prognostic, and not a general consequence of adjustment: among the eight primary cohorts the median count moved from 326.5 without proliferation to 66.5 with it, and in two cohorts it rose. In the registered ten the same medians are 44 and 66.5, so there the median rises. Which way the median moves depends on which cohorts are in the set, and neither figure should be read as a general consequence of adjustment. The direction reversed for 18.2 percent of all swept genes, falling to 3.7 percent among genes that were significant in either model and whose hazard ratio was at least 1.10 or at most its reciprocal, so most reversals are the noise of genes sitting at the null.

#### **The size of the non-collapsibility contribution**

The following is post hoc. The null band contained 2,963 genes. Within it the log shift was 0.064 times the gene's own log hazard ratio (95 percent CI 0.055 to 0.074), so a gene with a log hazard ratio of 1 has its hazard ratio multiplied by 1.066 when proliferation enters. Across the five bins of negative coefficient the median shift fell steadily below 1, from 0.997 to 0.981 as the coefficient grew. Across the five bins of positive coefficient it rose with the coefficient but crossed 1 only in the upper three, reaching 1.003; in the two bins of smallest positive coefficient the median shift was 0.997, within the noise floor of the band. The pattern is the one non-collapsibility produces and confounding does not, and its magnitude is small. The two sides are not symmetric in size, the negative side moving about six times as far as the positive one, which a purely proportional term would not produce, and I report that rather than smooth it over. Two qualifications belong with it. The band is not centered on a shift of exactly 1, its median being 0.9958, and the rule I fixed before fitting required the regression to pass through the origin, so any constant offset of that kind is absorbed into the slope; whether that inflates or deflates the estimate depends on the sign of the summed coefficients across the band, which I do not report, so I treat 0.064 as an estimate of the proportional term rather than as a bound on it. The simulation is also not an independent method: it draws survival from the same fitted model with the score permuted, and the rule for the null band states that the two must agree, so their agreement is a consistency check rather than corroboration. Applied to the worked example below, it predicted a shift of 1.014, against 1.019 (0.960 to 1.100) from the simulation that made proliferation independent of the gene. For the worked example the observed shift

of 1.105 exceeds the simulated non-collapsibility shift of 1.019, one-sided  $p = 0.023$ , and non-collapsibility accounts for 18.7 percent of the rise. Neither estimate accounts for the landscape.

#### **The registered prediction was refuted**

Of 33 candidate cancer types, 10 met the registered inclusion rule and 23 were excluded, most for having fewer than 200 patients or 60 deaths. The three cohorts excluded by name are additional to that count. Dropping the two aggregate cohorts leaves the eight in which no patient appears twice. Among those eight the slope was negative in four, a proportion of 0.50 against the registered refutation threshold of 0.75, and the median R-squared was 0.579. The median pairwise agreement of the shift across the 28 pairs was -0.268, opposite in sign to the prediction. The two hypotheses are not independent: cohorts whose slopes have opposite signs cannot agree gene by gene, so the second follows from the failure of the first, and the registration has one degree of freedom rather than two. In the original ten the same quantities are five of ten, a median R-squared of 0.586, and a median pairwise agreement of -0.293 over 45 pairs, so neither verdict depends on which set is used. Both hypotheses are refuted (Figure 2A, Table 2).

Two features of that rule limit what its verdict can mean, and I state them beside it. With eight cohorts, a threshold of 90 percent can only be met by eight of eight, so the confirming region is a single point. And the quantity the rule counts is the sign of a slope to which I attach no interval, in cohorts where the confidence interval of the proliferation hazard ratio covers 1 in seven of the eight, so under a state of the world in which proliferation carries little prognostic information in most cancers each sign is close to a coin flip and the rule refutes with high probability. The rule I wrote could therefore fail easily and succeed only on a clean sweep. I would not write it that way again.

The first hypothesis was a conjunction for confirmation and a disjunction for refutation: the plan confirms it only if the slope is negative in at least 90 percent of cohorts and the median R-squared is at least 0.50, and refutes it if the proportion negative falls below 75 percent or the median R-squared falls below 0.30. The proportion negative, 0.50, is below 0.75 in both sets, so the plan refutes the hypothesis as written, even though the median R-squared of 0.579 meets the confirmation threshold. No reading adopted after the result was seen was needed to reach that verdict. The part of the hypothesis that failed is the part about the direction of the slope.

#### **The refutation has a structure**

The following is post hoc. When the eight primary cohorts are ordered by the hazard ratio of the proliferation score in their own model, the slope is negative in every cohort whose estimate exceeds 1 and positive in every cohort whose estimate falls below it. With the three exploratory cohorts added the rule holds in 11 of 11, and in the original ten with their exploratory cohorts it holds in 13 of 13. The interval for that hazard ratio excludes 1 in only two of the ten registered cohorts, so this is a correspondence between two estimates from the same models rather than a division of cancers into those in which proliferation is prognostic and those in which it is not. The slope and the log hazard ratio of the proliferation score are correlated at -0.985 across the eight (Pearson) and at -0.992 across the eleven; regressing the former on the latter gives a coefficient of -0.930 with an intercept of -0.008 across the eight, and -0.956 with an intercept of -0.008 across the original ten (Figure 2).

This correspondence is close to what the definitions require rather than a discovery. Under the standard omitted variable expression, the change in a gene's coefficient is approximately the product of the proliferation coefficient and that gene's partial correlation with the score, and the landscape slope is the regression of that change on the same correlation, so a regression of the slope on the proliferation log hazard ratio should return a coefficient near -1 and an intercept near 0 whatever the data. I therefore present these values as a consistency check on the sweep. How far the expression actually accounts for what happens gene by gene is a separate question, and it is measured below rather than assumed. What is empirical, and what my registered prediction assumed away, is that the proliferation hazard ratio itself varies across cancer types widely enough to change the sign of the slope.

This is what the elementary account of confounding predicts[4,23], since the bias a confounder contributes is the product of its association with the outcome and its association with the exposure. A cohort in which proliferation carries no prognostic information has no confounding to remove, and one in which its association reverses has the shift reverse with it. The second registered hypothesis failed for the same reason, as noted above. My error was to write the prediction in terms of the sign of a slope rather than in terms of the structure that produces it, which amounted to treating a property of the liver cohort as a law. This relationship is a hypothesis generated by these data, not a result established in them, and it has been registered for testing in cohorts that have not been examined (doi:10.17605/OSF.IO/N8DW6).

### **The instrument is not the explanation**

Among the eight primary cohorts the eight genes cohered somewhat less tightly where the slope came out positive, with a median mean inter-item correlation of 0.536 against 0.625, and the same ordering appeared in

the standardized Cronbach alpha, the first principal component share and the agreement with the hallmark score. The two ranges overlap, with 0.497 to 0.639 where the slope was positive against 0.585 to 0.738 where it was negative, so this is a difference in central tendency and not a separation (Table 3). Measurement error pulls a hazard ratio toward 1 and a slope toward 0, so I examined it, and it does not account for the pattern. Agreement with the 195-gene hallmark score ran from 0.775 to 0.946 across all thirteen cohorts, so the eight genes measure proliferation in every one of them. Across the eight primary cohorts the standardized alpha ran from 0.899 to 0.958, so correcting for attenuation changed the log hazard ratio by at most 11.2 percent, reversed its sign nowhere, and moved the coefficient from -0.930 to -0.869; that pair of coefficients is computed from the deposited per-cohort slopes and reliabilities, since the deposited correction itself covers the registered ten and all thirteen. In the original ten the alpha fell as low as 0.897, the largest change was 11.5 percent, and the coefficient moved from -0.956 to -0.894. The reliability statistics and the proliferation hazard ratio are themselves strongly associated across cohorts, and with so few cohorts a nested regression cannot separate their contributions; I ran it across all thirteen cohorts for completeness and do not read the resulting p values, which run from 0.25 to 0.63, as evidence that reliability contributes nothing. Attenuation shrinks an estimate toward the null rather than reversing it, and the positive slopes ran from 0.020 to 0.095. In none of those four cohorts did the confidence interval of the proliferation hazard ratio exclude 1, so what they show is an absence of detectable prognostic effect rather than a protective one. The same caution belongs on the other side: among the four primary cohorts with a negative slope, that interval excluded 1 in one, and among the five with a negative slope in the registered ten it excluded 1 in two.

### **What the first-order expression does not account for**

This analysis is post hoc. The first-order expression accounts for much of the shift and not all of it, and it fails both readings I fixed in advance (Table 4). In the discovery cohort the predicted and observed log shifts correlated at 0.916 and the predicted value explained 0.839 of the variance of the observed one, against the 0.95 I had set; the regression of observed on predicted had a slope of 0.897, against the 0.95 to 1.05 I had set. GSE14520 gave 0.906, 0.820 and 0.867. In both cohorts the observed shift is about a tenth smaller in magnitude than the linear expression predicts. That the expression does not carry over exactly is established[23,24]; what these data add is how far it misses.

What is left over is not small for an individual gene. The median absolute residual was a third of the median absolute shift in both cohorts. The ratio is smallest where the shift is largest, at 0.20 in the lowest decile of the

predicted shift and 0.30 in the highest, and rises above 0.9 in the decile where the predicted shift is closest to zero, so the expression is a good guide to the genes whose estimates move most and a poor one for the rest.

### **Whether the departure is a collinearity artifact**

This analysis is post hoc. The genes whose residuals are largest are not scattered through the sweep. All fifty of the largest residuals sit at auxiliary coefficients between 0.60 and 0.95, which is to say they are close to a linear function of the score once the covariates are removed, and for eight of them the predicted and observed shifts point in opposite directions. The size of the residual rises with that coefficient, at 0.076 per unit (0.074 to 0.079) in the discovery cohort and 0.045 (0.043 to 0.047) in GSE14520, and the 333 genes reaching 0.5, which are 1.97 percent of those swept, carry 40.1 percent of the squared residual, while the highest decile of the coefficient carries 50.9 percent. In GSE14520 the same figures are 188 genes, 1.73 percent, 34.3 percent and 48.5 percent. Read across deciles of that coefficient, the residual is almost the whole of the shift where the gene is nearly independent of the score, at 0.97 in the lowest decile, and about a fifth of it where the gene is closest to the score, at 0.23 in the highest.

Excluding those genes does not restore the expression. The variance explained rises from 0.839 to 0.870 in the discovery cohort and from 0.820 to 0.855 in GSE14520, and the slope moves from 0.897 to 0.868 and from 0.867 to 0.902, so in the discovery cohort it ends further from 1 than it began. Both changes fall between the thresholds I had fixed, above the 0.02 below which I would have concluded that near-collinearity does not matter and below the 0.05 at which I would have concluded that it accounts for part of the departure, so no reading I had written applies and I report that rather than choose the one that suits the argument. The residual per unit of the gene's own log hazard ratio is 0.096 and 0.044 on the retained genes, against the 0.064 of the null band, so the secondary reading fails there as it did before. The landscape is not made by these genes either: refitted on the retained genes, it has a slope of -0.261 with an R-squared of 0.807 in the discovery cohort, against -0.278 and 0.785 on all genes, and -0.186 with 0.824 in GSE14520, against -0.180 and 0.796. One feature of the sign disagreements belongs here: most of them sit where the auxiliary coefficient is smallest, 687 in its lowest decile against 8 in its highest, so a disagreement of sign is mainly a property of genes whose predicted shift is near zero rather than of the genes that are close to the score.

The residual carries the signature of non-collapsibility but not its size. Its median absolute value is symmetric in the sign of the gene's own coefficient and rises at both ends, from 0.013 in the middle decile to 0.022 and 0.023 in the extremes, which is what an inflation proportional to the coefficient produces. But the fitted constant was

0.098 (0.095 to 0.101) in the discovery cohort and 0.053 (0.051 to 0.055) in GSE14520, neither interval contains the 0.064 the null band gave, and the two cohorts do not agree with each other. Adding the gene's own coefficient raised the variance explained only from 0.839 to 0.859 and from 0.820 to 0.838, so non-collapsibility accounts for about a tenth of what the first-order expression leaves. By the readings fixed in advance, the departure is unexplained by either account.

One regularity in it is worth recording. The residual per unit of the gene's own coefficient, divided by the proliferation coefficient in the same cohort, was 0.260 (0.252 to 0.269) and 0.245 (0.235 to 0.256), so the departure scales with the strength of the omitted covariate even though its raw size does not transfer between cohorts. Two cohorts cannot establish that, and I record it as an observation for a later test rather than as a result.

### **A worked example**

CYB5R3 in hepatocellular carcinoma illustrates the mechanism concretely, and it is the observation from which this work started[28]. Its expression correlated with the proliferation score at -0.19, and proliferation was adverse in the same model at 1.588 (1.278 to 1.972), so both conditions for a downward bias were present. That hazard ratio is from the grade-adjusted model in 339 patients, which is the model the sweep used; the 1.47 shown for this cohort in Table 2 is from the model without grade in all 341 patients, fitted so that the liver cohort is comparable with the others. The hazard ratio per standard deviation was 1.241 (1.027 to 1.500) without a proliferation term and 1.371 (1.132 to 1.661) with one, and the direction was the same in GSE14520, where it moved from 1.244 to 1.311. A bootstrap of that movement gives a ratio of hazard ratios of 1.105 (1.035 to 1.213) in the discovery cohort and 1.054 (0.999 to 1.159) in GSE14520, so the movement is larger than resampling noise in the first cohort and is not distinguishable from it in the second. I therefore describe the second as the same direction rather than as a replication. The rule I had fixed in advance for that cohort was met on its own terms, since the adjusted hazard ratio rose to 1.311 with an interval from 1.037 to 1.656, and the deposited verdict file records it as a replication; the bootstrap criterion I added later is the stricter of the two and it is the one I report. The gene does not pass false discovery control in either model, with q values of 0.278 and 0.402, so this is an illustration of how an estimate moves and not a claim that CYB5R3 is a biomarker. It is a fragile illustration. Restricting the cohort to the patients who have a recorded measure of liver function reduces the estimate and widens its interval: 1.229 (0.943 to 1.601) among those with a Child-Pugh class, 1.173 (0.925 to 1.486) with serum albumin, 1.206 (0.955 to 1.522) with a platelet count, and 1.276 (0.973 to 1.674) with an Ishak score. The Ishak subset retained 198 of the 339 patients, below the 60 percent coverage I had fixed for

these subsets, and I report it as falling below that rule. The interval covers 1 in all four, and it does so whether or not the liver function variable is in the model, so the loss is attributable to the smaller sample rather than to the covariate. Grade, which is often used in molecular studies as a surrogate for aggressiveness[19,29], carried no prognostic information in this cohort once age, sex and stage were accounted for (1.199, 0.925 to 1.554, likelihood ratio  $p = 0.462$ ), which is why it cannot stand in for a proliferation term.

The movement is not specific to proliferation, and I report the analysis that shows it. Adding the combined ESTIMATE stromal and immune score, with no proliferation term in the model, moved the same estimate from 1.241 to 1.268, and adding it on top of the proliferation score moved the estimate from 1.371 to 1.412. Any covariate that is prognostic and negatively correlated with a gene will move that gene's estimate in the same direction, which is what the confounding expression says. Proliferation is the case worth measuring because it is correlated with so many genes at once and is so rarely adjusted for, not because it is the only covariate with this property.

## Discussion

Adjusting a prognostic gene expression model for tumor proliferation moves the estimate for nearly every gene, by an amount and in a direction that the gene's correlation with proliferation predicts. In hepatocellular carcinoma the movement is large enough to change what a study of that cohort would report. What I could not show is that this holds everywhere, because it does not.

The registered prediction failed in half the cancer types tested, and the registration is what made that visible. Without it I would have had four cohorts with the sign I expected, an appealing account, and no obligation to count the others. The pattern in the failure is orderly and matches the textbook form of confounding, with the sign of the shift following the sign of the proliferation association in 13 of 13 cohorts. That is a better result than the one I predicted, and it is also the result I am least entitled to assert, since I formed it after seeing the estimates it describes. The honest position is that this paper reports a refutation and a hypothesis, and that the hypothesis is registered elsewhere for a test in data I have not seen.

Why proliferation is prognostic in some cancers and not in others is not answered here, and two kinds of explanation stay open. The cohorts differ in ways that are not biological: ovarian cancer is almost entirely advanced stage, so stage barely varies and its landscape was the weakest of the ten, and colon cancer contributed only 65 deaths. They may also differ biologically, since a measure of proliferation can discriminate little where

nearly every tumor proliferates, or where outcome is governed by another axis. I tested neither, and I would not want the post hoc relationship read as though I had.

Because the score measures proliferation with error, the adjusted models do not remove all of the confounding a perfectly measured proliferative index would remove, so every shift I report is a lower bound on the movement a fully adjusted estimate would show. The reliability of the score is not constant across cohorts, and by the grouping above it is somewhat lower where the slope came out positive. The ordering is not by the distance of the slope from zero: BLCA has the slope closest to zero among the eight and among the highest reliability of any cohort. So under-adjustment cannot be ruled out as a partial explanation for the flatter slopes, but the reliability statistics do not point to it, and with eight cohorts these data cannot separate the two.

Three alternative structures remain open for the underlying claim. Mediation would make proliferation a consequence of the gene rather than a cause of its apparent effect, in which case adjusting is overadjustment. Collider stratification would make conditioning on proliferation induce an association with an unmeasured determinant of both proliferation and death. Which of these holds is a question about causal structure and not about the fit[4]. Non-collapsibility contributes a shift that the null band analysis measures at 1.066 per unit coefficient, which is present but too small to account for the landscape. These data cannot separate the first two from confounding, and the descriptive claim, that the estimate depends systematically on whether proliferation is in the model, holds whichever structure is correct.

The practical consequence does not wait on that resolution. Twenty-three of the 41 papers with a reachable full text reported proliferation pathway activity in their own high-risk group and none adjusted for it, so the confounder was seen and read as biology. Twenty-eight did not state their covariates in the text at all, so a reader cannot tell what a reported hazard ratio is conditional on. I suggest four steps for a prognostic gene expression report, and I note that three of them restate existing reporting guidance rather than add to it. First, estimate the hazard ratio of a proliferation score in one's own cohort and report it, because that value, and not a rule imported from another tumor type, is what these data associate with whether omitting proliferation matters at all. Second, state the adjustment set in the text rather than only in a table. Third, report the gene's estimate with and without a proliferation term, together with the ratio of the two. Fourth, report the gene's correlation with the score, which is what predicts the size of that ratio. Only the first step is specific to this work; the second and third are REMARK and STROBE items and the fourth is one line of descriptive statistics. The first is the one my own registered prediction would have skipped, and it costs one model. None of the four tells a reader which of the two estimates to believe, because that depends on a causal structure these data do not establish;

what they do is make the dependence visible. Grade is not a substitute for any of this, since it carried no prognostic information in the discovery cohort.

Three analyses in the deposit bear on these questions and I report them rather than rest on the general point. For the worked example the movement of the estimate reproduced on disease-specific survival and did not appear on the progression-free interval, where the hazard ratio was 1.012 (0.864 to 1.185) without proliferation. Adjusting that example for a tissue composition score left the movement intact in the discovery cohort and did not reproduce the discovery cohort's pattern in GSE14520. The proportional hazards assumption held in the discovery cohort for the gene, age, sex, stage and grade, but it failed for the proliferation score itself (Schoenfeld  $p = 0.002$ , global  $p = 0.009$ ). Under a rule fixed before the corrected models were fitted, I refitted the worked example with the score entered as a stratification factor on its tertiles and again with a time-varying coefficient. The hazard ratio for the gene was 1.337 (1.099 to 1.627) and 1.357 (1.118 to 1.646), against 1.371 (1.132 to 1.661) with the score as a constant term, so the movement I report does not depend on the violated specification. In the stratified model the assumption then failed for stage instead ( $p = 0.019$ ), while the global test did not reject ( $p = 0.19$ ), and I report that rather than leave it in the deposit. In GSE14520 the assumption held for every term including the score. I did not test it gene by gene across the sweep or in the confirmatory cohorts, so every proliferation hazard ratio in Table 2 and every value plotted in Figure 2 is a time-averaged summary, and the sign rule reported above is a statement about those summaries.

The landscape slope is reported without an interval. Expression is correlated across genes, so the nominal standard error of a regression over roughly 17,000 of them is not interpretable, and I quote no  $p$  value for the landscape fit. The slopes are descriptive quantities, and the comparison of a slope of -0.055 with one of 0.020 should not be read as a difference established by these data.

The limitations are substantial and I state them rather than leave them to be inferred. The discovery cohort and the cohort in which I sought to reproduce it are both liver cohorts of resected patients with preserved organ function and no recorded treatment, so nothing here travels to patients triaged for systemic therapy. The proliferation score is a construct rather than a measured proliferative index such as a mitotic count or Ki-67 staining, and its reliability differed across cohorts in the way reported above. Both discovery cohorts and all confirmatory cohorts are bulk tissue, so a tumor cell effect and a composition effect cannot be fully separated. The confirmatory analysis drew on one consortium with one processing pipeline, which is why the second registration adds cohorts from elsewhere. That consortium distributes aggregate cohorts alongside the single-organ ones, and my registration did not anticipate them, so the registered set had to be reduced after the fact to

the eight in which no patient appears twice. Stage entered every model as a binary variable, which is a coarse adjustment and leaves residual confounding by stage. The literature survey sampled 50 papers from a single tumor type and could determine the proliferation coding in 41 of them, so what it establishes is an upper confidence limit of 8.6 percent rather than zero. I coded every paper myself, with no second coder and therefore no measure of coding agreement, and I searched each full text for Ki-67, MKI67, mitotic count or index, and proliferation or cell cycle scores, so a covariate named only outside what I retrieved would have been missed. Every claim here is associative. The work is by a single author, and the two registrations and the deposited code are what stand in place of a second pair of hands.

One objection to all of this is that the landscape is a re-expression of the model rather than a measurement, since the omitted variable expression already ties the slope to the proliferation coefficient. The objection is right about the slope and it does not carry over to the per-gene shift. Measured gene by gene, the first-order expression leaves a sixth of the variance and is systematically about a tenth too large in magnitude, and what it leaves is not accounted for by non-collapsibility, whose contribution I estimated two ways that do not agree with each other. The strongest version of the objection is that the remainder comes from genes nearly collinear with the score, for which the two models are barely distinguishable and the shift is a product of that near-collinearity rather than an adjustment. That version is testable and I tested it. Those genes do carry a disproportionate share of the residual, 40 percent of it from 2 percent of the genes, and removing them still leaves the expression explaining 0.87 rather than 0.95 of the variance, with a slope no closer to 1 in the discovery cohort than before, while the landscape survives the same exclusion almost unchanged. So the per-gene shift is not recoverable from the two numbers the expression needs, and the size of the departure is an empirical quantity that this design can measure and that I did not know before measuring it. I do not have an explanation for it. The one regularity I can see, that the departure scales with the strength of the omitted covariate, rests on two cohorts and belongs in a later test.

An estimate is a statement about a model, not about a gene. This work measures how much one omitted term changes that statement across an entire transcriptome, shows that the amount is predictable, and shows that the prediction I first wrote for it was wrong in an informative way. Nothing in the argument is particular to proliferation or to cancer. Wherever a screen tests many correlated candidates against an outcome that one shared axis also moves, the direction of the resulting bias is a property of that axis in those data, and it has to be measured there rather than carried over from a setting where someone else measured it.

## Conclusions

Adjusting a prognostic gene expression model for tumor proliferation moves the estimate of nearly every gene, by an amount the gene's correlation with proliferation predicts. The direction of that movement is not a fixed property of the transcriptome: it follows the prognostic strength of proliferation in the cancer at hand, and my registered prediction that it would always point the same way was refuted in half the cancer types tested. What does not depend on the cancer is the reporting. An estimate given without a proliferation term is conditional on a covariate set that most published reports do not state, and in the sample examined here no paper adjusted for proliferation while more than half reported it as a property of their own high-risk group. Reporting the hazard ratio of a proliferation score in one's own cohort, and then the gene estimate with and without that term, costs one model and makes the dependence visible.

## List of abbreviations

BH, Benjamini-Hochberg; BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinoma; CI, confidence interval; COAD, colon adenocarcinoma; COADREAD, colon and rectal adenocarcinoma; GEO, Gene Expression Omnibus; HNSC, head and neck squamous cell carcinoma; HR, hazard ratio; KIRC, kidney renal clear cell carcinoma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; LUNG, lung adenocarcinoma and squamous cell carcinoma combined; LUSC, lung squamous cell carcinoma; OSF, Open Science Framework; OV, ovarian serous cystadenocarcinoma; PC1, first principal component; REMARK, Reporting Recommendations for Tumor Marker Prognostic Studies; SKCM, skin cutaneous melanoma; STAD, stomach adenocarcinoma; STROBE, Strengthening the Reporting of Observational Studies in Epidemiology; TCGA, The Cancer Genome Atlas.

## Declarations

### Ethics approval and consent to participate

This study analyzed only publicly available, de-identified data and generated no new human or animal data, so ethical approval and informed consent were not required.

### Clinical trial number

Not applicable.

### Consent for publication

Not applicable.

## **Availability of data and materials**

The datasets analyzed are publicly available and no new patient data were generated. The Cancer Genome Atlas expression, clinical and survival data are available from the UCSC Xena repository, <https://xenabrowser.net/datapages/>, and the GSE14520 cohort from the Gene Expression Omnibus repository, <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE14520>. The analysis code, the decision rules, every result file and the run logs that support the conclusions of this article are available in the Zenodo repository, <https://doi.org/10.5281/zenodo.23051386>. That deposit also holds the literature survey in full: the search rules, the frame of 424 records, the drawn sample and the coding sheet, which are the only data I generated by hand. Project name: cyb5r3-proliferation-hcc. Archived version: version 2.4.0, <https://doi.org/10.5281/zenodo.23051386>, with the concept identifier <https://doi.org/10.5281/zenodo.22685965> resolving to the most recent version. Operating system: platform independent. Programming language: R. Other requirements: R 4.x with survival, dplyr, data.table, stringr, ggplot2, ggpubr, ggrepel, UCSCXenaTools, GEOquery, Biobase and msigdb, with clinfun and ragg for one test and for compressed TIFF output. License: MIT. No restrictions apply to use by non-academics. The two registered analysis plans are available on the Open Science Framework at <https://doi.org/10.17605/OSF.IO/X5DCF> and <https://doi.org/10.17605/OSF.IO/N8DW6>.

## **Competing interests**

The author declares 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**

SWN is the sole author. SWN designed the study, wrote the preregistrations, performed all analyses, prepared the figures and tables, and wrote the manuscript. SWN read and approved the final manuscript.

## **Acknowledgements**

Not applicable.

## References

1. Riley RD, Hayden JA, Steyerberg EW, Moons KG, Abrams K, Kyzas PA, et al. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. *PLoS Med.* 2013;10:e1001380.
2. Venet D, Dumont JE, Detours V. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput Biol.* 2011;7:e1002240.
3. Starmans MH, Krishnapuram B, Steck H, Horlings H, Nuyten DS, van de Vijver MJ, et al. Robust prognostic value of a knowledge-based proliferation signature across large patient microarray studies spanning different cancer types. *Br J Cancer.* 2008;99:1884-90.
4. Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. *Stat Sci.* 1999;14:29-46.
5. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM. Reporting recommendations for tumor marker prognostic studies (REMARK). *J Natl Cancer Inst.* 2005;97:1180-4.
6. Sauerbrei W, Taube SE, McShane LM, Cavenagh MM, Altman DG. Reporting recommendations for tumor marker prognostic studies (REMARK): an abridged explanation and elaboration. *J Natl Cancer Inst.* 2018;110:803-11.
7. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening of Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet.* 2007;370:1453-7.
8. Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. The preregistration revolution. *Proc Natl Acad Sci U S A.* 2018;115:2600-6.
9. Goldman MJ, Craft B, Hastie M, Repecka K, McDade F, Kamath A, et al. Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol.* 2020;38:675-8.
10. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell.* 2018;173:400-16.e11.
11. Cancer Genome Atlas Research Network. Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell.* 2017;169:1327-41.e23.
12. Roessler S, Jia HL, Budhu A, Forgues M, Ye QH, Lee JS, et al. A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients. *Cancer Res.* 2010;70:10202-12.
13. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics.* 2007;23:1846-7.
14. Whitfield ML, George LK, Grant GD, Perou CM. Common markers of proliferation. *Nat Rev Cancer.* 2006;6:99-106.
15. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database hallmark gene set collection. *Cell Syst.* 2015;1:417-25.
16. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics.* 2011;27:1739-40.
17. Hoshida Y, Nijman SM, Kobayashi M, Chan JA, Brunet JP, Chiang DY, et al. Integrative transcriptome analysis reveals common molecular subclasses of human hepatocellular carcinoma. *Cancer Res.* 2009;69:7385-92.
18. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol.* 1995;57:289-300.
19. Martins-Filho SN, Paiva C, Azevedo RS, Alves VAF. Histological grading of hepatocellular carcinoma: a systematic review of literature. *Front Med (Lausanne).* 2017;4:193.

637 20. Martinussen T, Vansteelandt S. On collapsibility and confounding bias in Cox and Aalen regression models.  
638 Lifetime Data Anal. 2013;19:279-96.

639 21. Hernan MA. The hazards of hazard ratios. Epidemiology. 2010;21:13-5.

640 22. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman & Hall; 1993.

641 23. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with  
642 nonlinear regressions and omitted covariates. Biometrika. 1984;71:431-44.

643 24. Struthers CA, Kalbfleisch JD. Misspecified proportional hazard models. Biometrika. 1986;73:363-9.

644 25. Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour  
645 purity and stromal and immune cell admixture from expression data. Nat Commun. 2013;4:2612.

646 26. Therneau TM, Grambsch PM. Modeling survival data: extending the Cox model. New York: Springer;  
647 2000.

648 27. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating  
649 assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361-87.

650 28. Nam SW. Divergence of the CYB5R3-mARC1 redox axis across the human MASLD-to-hepatocellular-  
651 carcinoma continuum: an exploratory analysis of liver-biopsy and tumor cohorts. Int J Mol Sci. 2026;27:7806.

652 29. Calderaro J, Couchy G, Imbeaud S, Amaddeo G, Letouze E, Blanc JF, et al. Histological subtypes of  
653 hepatocellular carcinoma are related to gene mutations and molecular tumour classification. J Hepatol.  
654 2017;67:727-38.

655 **Table 1**

656 **What the sampled literature adjusts for, and what could be retrieved.**

657 Fifty papers were drawn under a fixed seed from a frame of 424 records; 41 had a reachable full text. Whether a  
658 proliferation measure was a covariate was determined from the full text. Whether grade was a covariate was  
659 coded only from what the paper states in the text. The frame is a PubMed search for hepatocellular carcinoma  
660 prognostic gene expression studies published from 2024 onward, and grade was never counted as an adjustment  
661 for proliferation. n, papers assessed; k, papers meeting the item; CI, confidence interval, computed by the  
662 Wilson method. The first four rows have a denominator of the 41 papers with a reachable full text, and the last  
663 row has a denominator of all 50 sampled papers.

| Coding                                                       | n  | k  | %    | 95% CI       |
|--------------------------------------------------------------|----|----|------|--------------|
| Adjusted for proliferation                                   | 41 | 0  | 0.0  | 0.0 to 8.6   |
| Reported proliferation activity in their own high-risk group | 41 | 23 | 56.1 | 41.0 to 70.1 |
| Did not state the covariate list in the text                 | 41 | 28 | 68.3 | 53.0 to 80.4 |
| Stated an adjustment for grade                               | 41 | 7  | 17.1 | 8.5 to 31.3  |
| No reachable full text                                       | 50 | 9  | 18.0 | 9.8 to 30.8  |

664 **Table 2**

665 **The landscape by cohort.**

666 The shift is the factor by which a gene's hazard ratio is multiplied when the score enters, and every model is for  
 667 overall survival. The primary set is the eight cohorts in which no patient appears twice. The two aggregate  
 668 cohorts are shown separately because each contains a cohort that the registered rule had already removed:  
 669 COADREAD contains rectal adenocarcinoma, excluded for size, and LUNG contains lung adenocarcinoma,  
 670 excluded by name. The three exploratory cohorts were examined before the registration existed. The slope is the  
 671 regression of the log shift on the gene's Spearman correlation with the proliferation score, and R-squared is the  
 672 R-squared of that regression; no interval is given, because expression is correlated across genes and the nominal  
 673 standard error of a regression over roughly 17,000 of them is not interpretable. HR, hazard ratio of the  
 674 standardized proliferation score per 1 standard deviation, in the same model; CI, confidence interval. BH is the  
 675 number of genes at Benjamini-Hochberg q below 0.10, without and with the proliferation term. Every cohort  
 676 here is fitted on a common base of age, sex and stage, except ovarian cancer, which contains one sex and is  
 677 fitted on age and stage. The discovery analysis reported in the text additionally adjusts for histologic grade,  
 678 which is why the liver slope there, -0.278 over 16,902 genes in 339 patients, and its proliferation hazard ratio of  
 679 1.588, differ from the liver values here, which are fitted without grade in 341 patients. For the same reason the  
 680 731 genes surviving false discovery control in the liver row here are not the 338 reported in the abstract and in  
 681 the text: the 731 come from the model without grade in 341 patients and the 338 from the grade-adjusted  
 682 discovery model in 339 patients.

| Cohort                                        | Cancer type                                     | Patients | Deaths | Proliferation HR (95% CI) | Slope  | R-squared | BH without | BH with |
|-----------------------------------------------|-------------------------------------------------|----------|--------|---------------------------|--------|-----------|------------|---------|
| Primary set                                   |                                                 |          |        |                           |        |           |            |         |
| BLCA                                          | Bladder urothelial carcinoma                    | 403      | 176    | 1.08 (0.92 to 1.28)       | -0.055 | 0.405     | 687        | 555     |
| BRCA                                          | Breast invasive carcinoma                       | 755      | 103    | 1.13 (0.93 to 1.36)       | -0.111 | 0.650     | 11         | 0       |
| COAD                                          | Colon adenocarcinoma                            | 270      | 65     | 0.92 (0.71 to 1.18)       | 0.095  | 0.548     | 2          | 2       |
| HNSC                                          | Head and neck squamous cell carcinoma           | 505      | 218    | 1.10 (0.96 to 1.26)       | -0.111 | 0.762     | 1341       | 1684    |
| LUSC                                          | Lung squamous cell carcinoma                    | 485      | 210    | 0.94 (0.82 to 1.08)       | 0.048  | 0.609     | 2          | 2       |
| OV                                            | Ovarian serous cystadenocarcinoma               | 302      | 183    | 0.97 (0.84 to 1.13)       | 0.020  | 0.137     | 0          | 0       |
| SKCM                                          | Skin cutaneous melanoma                         | 407      | 189    | 1.20 (1.03 to 1.40)       | -0.197 | 0.754     | 3954       | 4494    |
| STAD                                          | Stomach adenocarcinoma                          | 373      | 148    | 0.89 (0.76 to 1.05)       | 0.076  | 0.516     | 642        | 131     |
| Aggregate cohorts (sensitivity analysis only) |                                                 |          |        |                           |        |           |            |         |
| COADREAD                                      | Colon and rectal adenocarcinoma                 | 354      | 78     | 0.92 (0.73 to 1.16)       | 0.084  | 0.562     | 0          | 0       |
| LUNG                                          | Lung adenocarcinoma and squamous cell carcinoma | 969      | 387    | 1.15 (1.03 to 1.28)       | -0.150 | 0.892     | 77         | 375     |
| Exploratory                                   |                                                 |          |        |                           |        |           |            |         |
| LIHC                                          | Liver hepatocellular carcinoma                  | 341      | 116    | 1.47 (1.21 to 1.80)       | -0.320 | 0.840     | 731        | 1       |
| KIRC                                          | Kidney renal clear cell carcinoma               | 529      | 174    | 1.41 (1.22 to 1.64)       | -0.292 | 0.752     | 8177       | 5785    |

|      |                     |     |     |                     |        |       |      |    |
|------|---------------------|-----|-----|---------------------|--------|-------|------|----|
| LUAD | Lung adenocarcinoma | 484 | 177 | 1.36 (1.17 to 1.59) | -0.293 | 0.876 | 1665 | 26 |
|------|---------------------|-----|-----|---------------------|--------|-------|------|----|

**Table 3**

**The proliferation score as an instrument, by cohort.**

These checks are post hoc. Inter-item  $r$  is the mean pairwise Pearson correlation among the eight score genes;  $\alpha$  is their standardized Cronbach  $\alpha$ ; PC1 share is the proportion of their variance taken by the first principal component; hallmark agreement is the Pearson correlation of the eight-gene score with a score built from the genes of HALLMARK\_G2M\_CHECKPOINT that are not among the eight. All four are computed in each cohort's own sweep analysis set. Slope sign is the sign of that cohort's landscape slope in Table 2, and the hallmark comparison score is built from the 195 genes of HALLMARK\_G2M\_CHECKPOINT that are not among the eight. Cohort abbreviations are expanded in Table 2.

| Cohort                                           | Slope sign | Inter-item $r$ | Alpha | PC1 share | Hallmark agreement |
|--------------------------------------------------|------------|----------------|-------|-----------|--------------------|
| Primary set                                      |            |                |       |           |                    |
| BLCA                                             | negative   | 0.665          | 0.953 | 0.756     | 0.932              |
| BRCA                                             | negative   | 0.738          | 0.958 | 0.778     | 0.946              |
| HNSC                                             | negative   | 0.585          | 0.923 | 0.657     | 0.914              |
| SKCM                                             | negative   | 0.586          | 0.929 | 0.673     | 0.916              |
| COAD                                             | positive   | 0.497          | 0.899 | 0.598     | 0.884              |
| LUSC                                             | positive   | 0.502          | 0.912 | 0.627     | 0.849              |
| OV                                               | positive   | 0.570          | 0.928 | 0.672     | 0.873              |
| STAD                                             | positive   | 0.639          | 0.949 | 0.740     | 0.927              |
| Aggregate cohorts<br>(sensitivity analysis only) |            |                |       |           |                    |
| COADREAD                                         | positive   | 0.503          | 0.897 | 0.594     | 0.898              |
| LUNG                                             | negative   | 0.729          | 0.963 | 0.797     | 0.924              |
| Exploratory                                      |            |                |       |           |                    |
| LIHC                                             | negative   | 0.758          | 0.959 | 0.783     | 0.919              |
| KIRC                                             | negative   | 0.695          | 0.951 | 0.752     | 0.775              |
| LUAD                                             | negative   | 0.772          | 0.964 | 0.803     | 0.940              |

**Table 4**

**How far the first-order omitted variable expression accounts for the per-gene shift.**

This analysis is post hoc. The predicted log shift for a gene is minus the product of the proliferation coefficient and the coefficient of the standardized gene in an auxiliary regression of the score on that gene and the other covariates; the residual is the observed log shift minus the predicted one. The two readings fixed before the analysis was run were that the expression is adequate if it explains at least 0.95 of the variance with a slope between 0.95 and 1.05, and that what it omits is non-collapsibility if the residual per unit of the gene's own log hazard ratio is close to the 0.064 obtained independently in the null band. Neither reading is met in either cohort.

That constant is then divided by the proliferation coefficient of the same cohort. The last four rows are the post hoc test of whether the departure belongs to genes lying close to the score: the auxiliary coefficient is on the scale of a partial correlation, genes reaching 0.5 were excluded, and the three readings fixed before those fits were that the departure is attributable to such genes if the variance explained reaches 0.95 with a slope within 0.05 of 1, partly attributable if it rises by at least 0.05 without reaching 0.95, and not attributable if it changes by less than 0.02. The observed change falls between those thresholds in both cohorts, so none of the three applies. TCGA-LIHC here is the grade-adjusted discovery sweep in 339 patients and GSE14520 the replication sweep in 225. CI, confidence interval.

| Quantity                                                            | TCGA-LIHC              | GSE14520               |
|---------------------------------------------------------------------|------------------------|------------------------|
| Genes                                                               | 16902                  | 10861                  |
| Proliferation coefficient (gamma)                                   | 0.375                  | 0.217                  |
| Pearson correlation of observed with predicted                      | 0.916                  | 0.906                  |
| Variance of the observed shift explained by the predicted one       | 0.839                  | 0.820                  |
| Slope of observed on predicted                                      | 0.897                  | 0.867                  |
| Median absolute residual on the log scale                           | 0.015                  | 0.009                  |
| Median absolute observed log shift                                  | 0.043                  | 0.028                  |
| Median absolute residual as a share of the median absolute observed | 0.35                   | 0.32                   |
| Variance explained after adding the gene's own coefficient          | 0.859                  | 0.838                  |
| Residual per unit of the gene's own log hazard ratio (95% CI)       | 0.098 (0.095 to 0.101) | 0.053 (0.051 to 0.055) |
| The same divided by gamma (95% CI)                                  | 0.260 (0.252 to 0.269) | 0.245 (0.235 to 0.256) |
| Genes whose auxiliary coefficient reaches 0.5                       | 333                    | 188                    |
| Share of the squared residual those genes carry                     | 0.401                  | 0.343                  |
| Variance explained with those genes excluded                        | 0.870                  | 0.855                  |
| Slope of observed on predicted with those genes excluded            | 0.868                  | 0.902                  |

## Figure legends

### Figure 1. The landscape of the shift in hepatocellular carcinoma.

The shift is the factor by which a gene's hazard ratio is multiplied when the proliferation score enters the model. (A) TCGA-LIHC, 16,902 genes in 339 patients, and (B) GSE14520, 10,861 genes in 225 patients: each gene's Spearman correlation with the proliferation score on the horizontal axis against the log shift on the vertical axis. Shading is gene density on a log scale, the dashed gray line marks a log shift of zero, the red line is the ordinary least squares fit, the white points are the median log shift within deciles of the correlation on the horizontal axis, and the red point labeled CYB5R3 is the worked example described in the Results. Each panel subtitle gives that panel's fit, rounded: the slope is -0.278 in (A) and -0.180 in (B), with variance explained of 0.785 and 0.796. The two panels are drawn on different vertical scales, and each shows the central 99.5 percent of its genes, so a few hundred extreme genes fall outside the drawn range; nothing is removed from any fit or any count. (C) The number of genes in TCGA-LIHC reaching p below 0.05 and Benjamini-Hochberg q below 0.10, in gray without

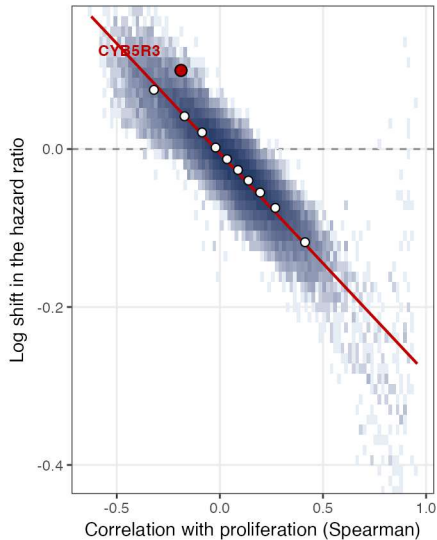
the proliferation score in the model and in blue with it. Every model is a Cox model for overall survival adjusted for age, sex and stage, with histologic grade added in TCGA-LIHC. These analyses are exploratory.

**Figure 2. The registered test and the structure in its failure.**

(A) The slope of each cohort's landscape, ordered from most positive to most negative. The red vertical line marks a slope of zero, and the registration predicted a negative slope in at least 90 percent of the cohorts meeting its rule, that is, to the left of that line; four of the eight primary cohorts fell there, and five of the ten including the aggregates. Dark blue bars are the eight primary cohorts, light blue the two aggregate cohorts that the primary analysis sets aside, and gray the three exploratory cohorts. (B) Each cohort's slope on the vertical axis against the log hazard ratio of the proliferation score in its own model on the horizontal axis. Filled circles are the eight primary cohorts, filled squares the two aggregates and open triangles the three exploratory cohorts. Each point is labeled with its cohort abbreviation, which Table 2 expands. Dashed gray lines mark zero on each axis, and the red line is fitted to the eight primary cohorts only, with the others shown where they fall on its extrapolation; its subtitle carries that fit, a correlation of -0.985 with a coefficient of -0.930 and an intercept of -0.008. The registration also required a median R-squared of at least 0.50 for confirmation, which was met. Panel B is post hoc.

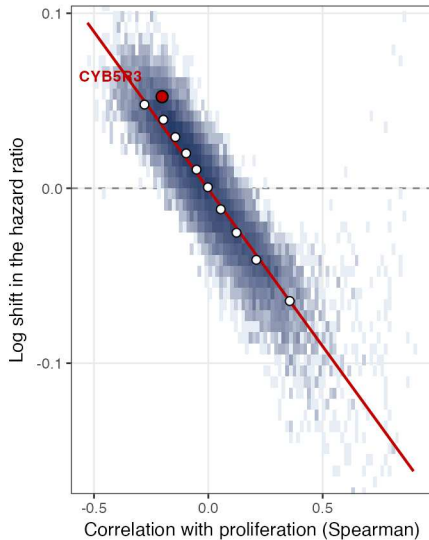
### A TCGA-LIHC

slope -0.278,  $R^2$  0.79, 16,902 genes



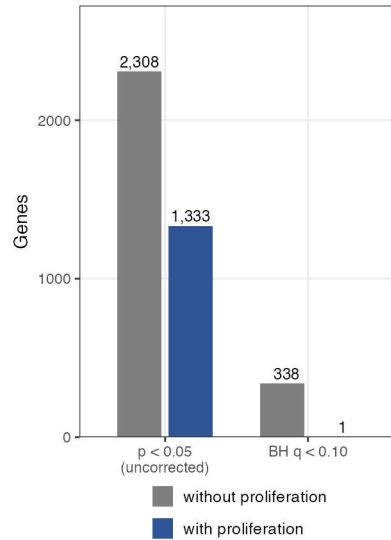
### B GSE14520

slope -0.180,  $R^2$  0.80, 10,861 genes



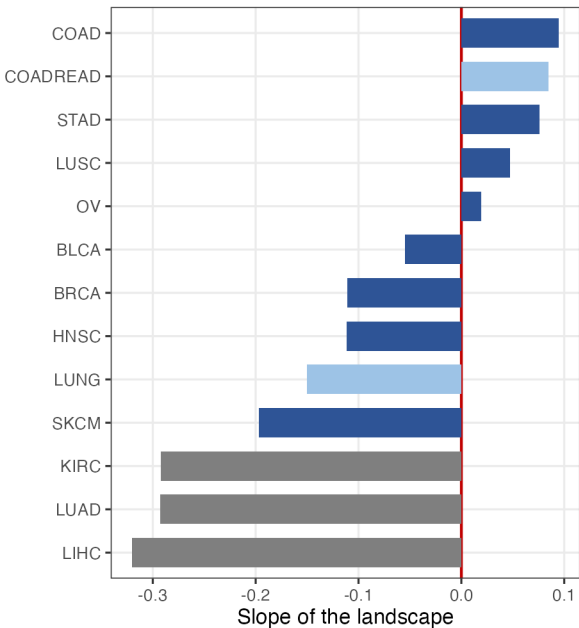
### C Genes passing each threshold

TCGA-LIHC, the same 16,902 genes



## A The registered prediction

negative in at least 90 percent, as registered



## B The structure in the failure (post hoc)

primary cohorts:  $r = -0.985$ , slope  $-0.930$ , intercept  $-0.008$

