Decision Thresholds and Policy
Volume 1 · Number 1 · 2026

What Does a Missing Clinical Field Tell Us? A Record-Completeness Study in a Public HCC Cohort

Adrian Reynolds[1] · Wesley Hayes[2] · Matthew Fisher[3]

[1] Finland · [2] Denmark · [3] Norway

Original publication: November 5, 2026 | Article 688

Abstract

Only eight of the 165 patients in a public hepatocellular-carcinoma dataset have complete values for all forty-nine predictors. We examined whether the number of missing fields itself relates to the released one-year survival label. After age and sex adjustment, ten additional missing fields corresponded to a 0.077 higher survival proportion, with an interval from -0.115 to 0.269. Adding performance-status categories reduced the estimate to 0.045, again with a wide interval spanning zero. The file contains 826 missing predictor cells despite a repository summary flag indicating no missingness. This small record audit supports explicit completeness reporting, while offering little precise evidence about the survival association of a simple missing-field count.

Exploratory reanalysis of historical deidentified records. The analysis estimates descriptive associations and makes no treatment recommendation.

What Does a Missing Clinical Field Tell Us? A Record-Completeness Study in a Public HCC Cohort

Exploratory reanalysis of historical deidentified records. The analysis estimates descriptive associations and makes no treatment recommendation.

Abstract

Only eight of the 165 patients in a public hepatocellular-carcinoma dataset have complete values for all forty-nine predictors. We examined whether the number of missing fields itself relates to the released one-year survival label. After age and sex adjustment, ten additional missing fields corresponded to a 0.077 higher survival proportion, with an interval from -0.115 to 0.269. Adding performance-status categories reduced the estimate to 0.045, again with a wide interval spanning zero. The file contains 826 missing predictor cells despite a repository summary flag indicating no missingness. This small record audit supports explicit completeness reporting, while offering little precise evidence about the survival association of a simple missing-field count.

1. Completeness is observable even when its mechanism is unknown

Clinical datasets are often described through the variables they contain. Equally relevant is which variables are absent for each patient. A missing value may reflect a decision not to order a test, unavailable documentation, a transfer between systems, or another feature of care. Counting missing fields does not identify those mechanisms, but it creates an observable description of how much information a record supplies. That description can be studied without inventing values for the absent measurements.

Our research question is whether record-level missingness burden is associated with the supplied one-year survival label in a small public HCC cohort. We examine the association first with age and sex adjustment and then with additional performance-status categories. The aim is an audit of the information structure available to a secondary analyst. It is not an attempt to develop a deployable survival score or estimate how obtaining another test would change a patient's outcome.

The cohort is available through Santos et al., 2015a and was introduced in the study by Santos et al., 2015b. The original work addressed a prediction problem with methods tailored to a small, imbalanced sample. We use its released records for a different question about completeness. We neither reproduce its oversampling strategy nor compare our descriptive regression's fit to its reported predictive performance. Such a comparison would mix distinct objectives and validation designs.

2. Why the clinical context does not supply a treatment estimand

Contemporary HCC evidence can distinguish treatment settings and viral-infection subgroups, as illustrated by Li et al., 2026. The present public cohort does not provide the treatment-assignment, eligibility, and temporal information needed to recreate that comparison. Its one-year outcome cannot identify the efficacy or safety of a first-line immunotherapy or targeted therapy. The citation locates the disease context while the article's actual estimand remains a record-completeness association.

The distinction also separates individual records from evidence synthesis. Guidance on systematic reporting in Page et al., 2021, network comparisons in Higgins et al., 2012, and heterogeneity in Higgins and Thompson, 2002 concerns combining studies under their respective assumptions. Our patients are not separate trials or treatment contrasts. We perform no network meta-analysis and infer no comparative

treatment ranking from the dataset. A patient-level missingness count is not a substitute for the covariates needed to assess treatment-effect modification.

The certainty perspective of Guyatt et al., 2008 is useful when communicating the limitations of this evidence. It does not transform a public-data association into clinical guidance. In particular, the observed completeness pattern cannot tell a clinician which laboratory measurement to order or whether a treatment is appropriate. Those decisions require current clinical evidence and patient-specific assessment beyond the scope of a historical anonymous matrix.

3. Reconcile the resource description with the actual file

The downloaded archive contains a comma-separated patient matrix and an original descriptive file. The matrix has 165 rows, forty-nine predictor columns, and one final outcome column. Question marks mark absent predictor values. The original description identifies the outcome as survival at one year, encoded one for living and zero for dying. We find 102 living and 63 dying labels, matching that description. No event dates or censoring times accompany those binary labels.

The current repository API says that missing values are absent. The file and its archived description contradict that summary: we count 826 absent predictor cells, or 10.216 percent of the 8,085 possible predictor entries. The analysis follows the actual file and explicitly records the mismatch. This is a source-description issue, not evidence that an individual patient lacks clinical care or that an unobserved laboratory result had a particular value.

The original description also reports very low full-record completeness. Our direct calculation identifies eight complete records, corresponding to 4.85 percent of the cohort. The remaining patients have at least one absent predictor, with the largest count reaching twenty-three. Restricting every analysis to all-variable complete cases would therefore retain a tiny and potentially selective subset. We do not estimate a complete-case survival regression on those eight patients or call their observed outcomes representative of the whole cohort.

Reporting principles in Benchimol et al., 2015 and von Elm et al., 2007 support distinguishing the acquired data structure from summary claims about it. The original file, field order, missing-value marker, and binary label definition are preserved. We do not rely on a blank contemporary variable table to assign new meanings to columns. The archival description supplies the field definitions, while direct parsing supplies their observed counts and completeness.

4. Construct a deliberately simple missingness measure

For each patient, we count question-mark values across all forty-nine predictor fields. The outcome column is excluded from this calculation. The count treats every field equally: an absent historical characteristic and an absent laboratory value each contribute one. That weighting is transparent but clinically crude. It does not imply that all missing variables are equally consequential, equally costly to collect, or missing for the same reason.

The explanatory variable is the count divided by ten. Thus its coefficient represents an association for ten additional absent fields, a scale chosen for interpretability rather than a clinical threshold. The observed count ranges from zero to twenty-three. Age is centered at sixty years and divided by ten, so its coefficient refers to a decade. The released gender field is used as a male indicator under the source coding. Age, gender, and performance status have no missing entries in this matrix.

This construction retains all 165 patients without imputing any clinical value. Missingness theory in Little and Rubin, 2002 explains why that choice does not establish that the missing-data mechanism is

ignorable. The missing-indicator issues discussed by Groenwold et al., 2012 are also relevant: a count used as the object of a descriptive audit is not a general solution to confounding from absent clinical covariates. We do not claim that adding the count repairs the information lost from those variables.

Imputation methods such as Buuren and Groothuis-Oudshoorn, 2011 and Stekhoven and Bühlmann, 2012 could create completed predictor matrices under additional modeling choices. They are not run here. Substituting imputed values before defining the count would erase the observed absence pattern that the question concerns. A later prediction study might compare imputation approaches, but would need to perform fitting within an appropriate validation design and preserve the distinction between observed and imputed evidence.

5. Two descriptions of the same patients

The first specification is a linear-probability regression of the one-year living label on missingness burden, age in decades, and the male indicator. The second adds indicators for performance-status categories one through four, with active status zero as the reference. We use categories rather than assuming that every status step has an equal outcome association. All covariates in both formulas are observed or derived from the original absence pattern, and both models retain the same sample.

The performance distribution is eighty patients at status zero, thirty at one, thirty-two at two, eighteen at three, and five at four. Small upper-status groups limit the precision and stability of category contrasts. The role of the second model is adjustment sensitivity within the available records. It does not establish that performance status blocks every confounding pathway or that the remaining missingness coefficient is causal.

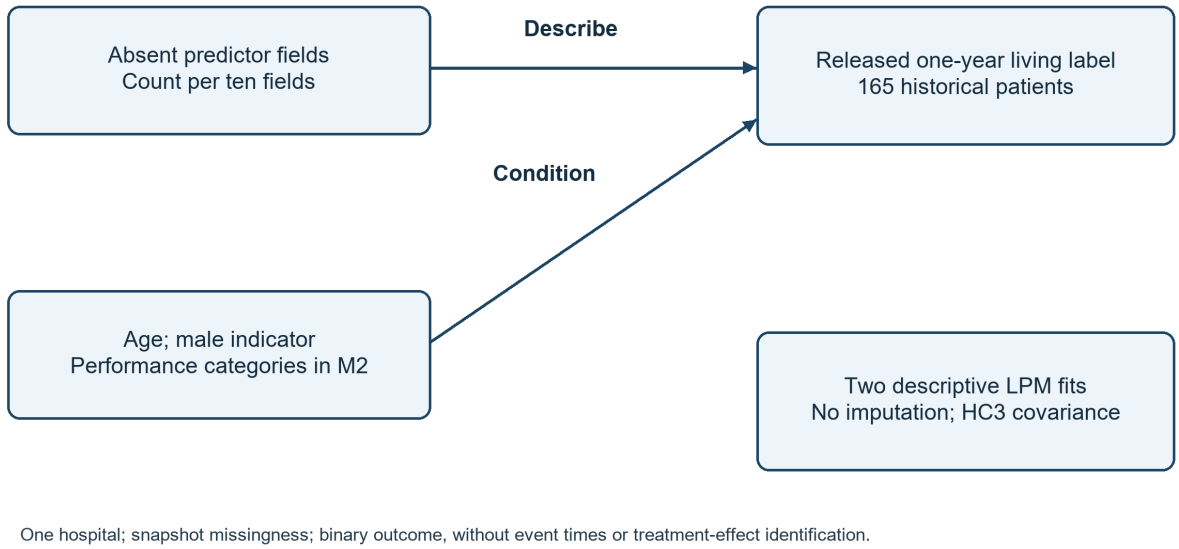

Figure 1. A missing-field count and fully observed demographic controls enter two descriptive survival-label models, with performance categories added in the second.

Heteroskedasticity-consistent HC3 standard errors follow the variance concerns introduced by White, 1980. The regression uses normal-reference intervals under the recorded implementation. It does not cluster by hospital because the resource contains one hospital, and a cluster label with a single group cannot establish a population of hospitals. Dependence considerations in Cameron and Miller, 2015 make clear that adding a covariance label cannot compensate for absent independent clinical settings.

Table 1. Missingness-burden and demographic coefficients in descriptive linear-probability models of the released living label.

Model	Term	Estimate (SE)	95% CI	N
M1	Ten additional absent fields	0.0769 (0.0981)	[-0.1154, 0.2693]	165
M1	Age per decade	-0.0522 (0.0297)	[-0.1104, 0.0060]	165
M1	Male indicator	0.0026 (0.1037)	[-0.2006, 0.2058]	165
M1: fit	R-squared = 0.0248	-	-	165
M2	Ten additional absent fields	0.0452 (0.0872)	[-0.1257, 0.2161]	165
M2	Age per decade	-0.0344 (0.0290)	[-0.0912, 0.0224]	165
M2	Male indicator	0.0275 (0.1006)	[-0.1697, 0.2247]	165
M2: fit	R-squared = 0.1668	-	-	165

Model labels: M1 = Age/sex-adjusted missingness association; M2 = Additional performance-status adjustment.

The linear model provides a directly interpretable difference in the fitted outcome proportion. It can produce out-of-range predictions and is not presented as a calibrated patient-risk calculator. Model-development guidance in Steyerberg et al., 2010 and risk-of-bias assessment in Wolff et al., 2019 concerns much more than estimating a coefficient table. No external calibration sample, clinical utility threshold, or prospective evaluation is available here. We consequently use the fit summaries only to describe the executed regressions.

6. A weakly resolved association, not evidence of equivalence

Patients labeled as dying have a mean of 4.683 missing fields and a median of four. Patients labeled as living have a mean of 5.206 and a median of five. These unadjusted descriptions are close relative to the observed count range. They do not establish comparable clinical information across outcome groups, since different absent fields can yield the same count. A history field and a laboratory field remain indistinguishable in this deliberately coarse burden measure.

In the age- and sex-adjusted model, the missingness coefficient is 0.07695 per ten absent fields, with standard error 0.09814 and interval -0.11540 to 0.26930. Its wide interval includes potentially negative and positive associations. After performance adjustment, the coefficient is 0.04520 with standard error 0.08720 and interval -0.12571 to 0.21611. The point estimate decreases, but the data do not resolve a precise independent missingness association in either specification.

The first model's R-squared is 0.0248 and the second's is 0.1668. Additional performance categories explain more variation in this sample, but that improvement is not an out-of-sample survival-prediction result. Age has coefficient -0.05222 per decade before performance adjustment and -0.03440 after it; both corresponding intervals include zero. We retain those uncertain estimates rather than interpreting their signs as established prognostic effects.

Interpretation under Wasserstein and Lazar, 2016 distinguishes an imprecise interval from evidence that no association exists. The concerns raised by Ioannidis, 2005 are relevant when a small dataset encourages broad conclusions from whichever specification yields a favorable significance threshold.

Here both prespecified descriptions and their uncertainty remain visible. We did not search for an absence-count cutoff that maximizes association with the outcome, or remove patients because their records weaken a preferred direction.

7. The absent mechanism is the main scientific limitation

The source does not identify when each field became available, why it was absent, or whether documentation completeness changed after an outcome-relevant event. A retrospective snapshot therefore cannot establish that the count would have been available at a specific prediction time. This temporal ambiguity is one reason we describe an association audit rather than a prospective prognostic model. The count may summarize the data-extraction process as well as patient assessment.

Causal reasoning in Greenland et al., 1999, observational-comparison principles in Stuart, 2010, and sensitivity perspectives in ROSENBAUM and RUBIN, 1983 emphasize that recorded adjustment does not create random treatment or random information collection. A patient's health could affect both the completeness of measurement and survival, while clinical workflows could affect documentation independently. The available matrix cannot separate these pathways. In particular, increasing the recorded count by deleting fields is not the same intervention as a clinical process that fails to collect those fields.

Measurement-error discussion in Carroll et al., 2006 also applies to the nominally observed values and the outcome label. The archive provides no independent adjudication of every entry. Influence ideas in Cook, 1977 could identify records with unusual covariate combinations, especially among five high-status patients. Those records are retained here. A sensitivity analysis deleting them would change the population represented by the fitted association and would need its own transparent rationale.

The endpoint is binary at one year. Competing-risk methods such as Fine and Gray, 1999 require event and follow-up information that this release does not supply. Applying a time-to-event label to our regression would therefore be incorrect. Likewise, discrimination concepts in Fawcett, 2006 and scoring principles in Gneiting and Raftery, 2007 do not turn the table into a completed prediction validation. We calculate neither a held-out AUC nor a proper-score comparison between competing clinical models.

8. A modest audit with a concrete reporting consequence

The most definite finding is the information pattern itself: nearly every patient has at least one absent field, and the actual file contradicts a high-level no-missingness flag. The association with the released outcome is much less definite. This contrast is useful for secondary research because a precise completeness count and an uncertain clinical association can coexist. Neither needs to be suppressed to tell a simpler story about the dataset.

Multiplicity control in Benjamini and Hochberg, 1995 would matter if a later analysis screened many missingness patterns and patient subgroups. Our two models do not enumerate those possibilities. The documentation aims of Gebru et al., 2021 support retaining actual field absence, source discrepancies, coding, and the same-patient comparison. Wilkinson et al., 2016 motivates preserving the raw archive and derived counts so another analyst can reproduce the audit without relying on changing repository summaries. Further clinical interpretation requires richer timing, measurement, treatment, and validation evidence than this small release contains.

References

1. Fine, Jason P., and Robert J. Gray. "A Proportional Hazards Model for the Subdistribution of a Competing Risk." *Journal of the American Statistical Association*, vol. 94, no. 446, 1999, pp. 496-509. https://doi.org/10.1080/01621459.1999.10474144.

2. Higgins, Julian P. T.; Thompson, Simon G.. (2002). Quantifying heterogeneity in a meta-analysis. Statistics in Medicine. https://doi.org/10.1002/sim.1186.

3. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.

4. Benchimol, Eric I.; Smeeth, Liam; Guttmann, Astrid; Harron, Katie; Moher, David; Petersen, Irene; Sørensen, Henrik T.; von Elm, Erik; Langan, Sinéad M.; , . (2015). The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLOS Medicine. https://doi.org/10.1371/journal.pmed.1001885.

5. Page, Matthew J, et al. "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews." *BMJ*, 2021, pp. n71. https://doi.org/10.1136/bmj.n71.

6. Guyatt, Gordon H, et al. "GRADE: an emerging consensus on rating quality of evidence and strength of recommendations." *BMJ*, vol. 336, no. 7650, 2008, pp. 924-926. https://doi.org/10.1136/bmj.39489.470347.ad.

7. Benjamini, Yoav and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 57, no. 1, 1995, pp. 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

8. Buuren, Stef van; Groothuis-Oudshoorn, Karin. (2011). mice: Multivariate Imputation by Chained Equations inR. Journal of Statistical Software. https://doi.org/10.18637/jss.v045.i03.

9. Santos, Miriam Seoane; Abreu, Pedro Henriques; García-Laencina, Pedro J.; Simão, Adélia; Carvalho, Armando. (2015a). HCC Survival [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5TS4S.

10. Cameron, A. Colin, and Douglas L. Miller. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources*, vol. 50, no. 2, 2015, pp. 317-372. https://doi.org/10.3368/jhr.50.2.317.

11. ROSENBAUM, PAUL R., and DONALD B. RUBIN. "The central role of the propensity score in observational studies for causal effects." *Biometrika*, vol. 70, no. 1, 1983, pp. 41-55. https://doi.org/10.1093/biomet/70.1.41.

12. Steyerberg, Ewout W., et al. "Assessing the Performance of Prediction Models." *Epidemiology*, vol. 21, no. 1, 2010, pp. 128-138. https://doi.org/10.1097/ede.0b013e3181c30fb2.

13. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; ’t Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18.

14. Higgins, J. P. T.; Jackson, D; Barrett, J. K.; Lu, G; Ades, A. E.; White, I. R.. (2012). Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. Research Synthesis Methods. https://doi.org/10.1002/jrsm.1044.

15. Wasserstein, Ronald L.; Lazar, Nicole A.. (2016). The ASA Statement on p -Values: Context, Process, and Purpose. The American Statistician. https://doi.org/10.1080/00031305.2016.1154108.

16. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.

17. Ioannidis, John P. A.. (2005). Why Most Published Research Findings Are False. PLoS Medicine. https://doi.org/10.1371/journal.pmed.0020124.

18. Li, Qinfei; Li, Hong; Ma, Haowei; Chen, Wei. (2026). Efficacy and safety of first-line immunotherapy and targeted therapy in advanced HCC: a network meta-analysis with subgroup analysis based on HBV and HCV infection. Frontiers in Immunology, 17, 1706446. https://doi.org/10.3389/fimmu.2026.1706446.

19. Greenland, Sander, Judea Pearl, and James M. Robins. “Causal Diagrams for Epidemiologic Research.” *Epidemiology*, vol. 10, no. 1, 1999, pp. 37-48. https://doi.org/10.1097/00001648-199901000-00008.

20. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.

21. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.

22. Fawcett, Tom. (2006). An introduction to ROC analysis. Pattern Recognition Letters. https://doi.org/10.1016/j.patrec.2005.10.010.

23. Gneiting, Tilmann; Raftery, Adrian E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. https://doi.org/10.1198/016214506000001437.

24. Stekhoven, Daniel J.; Bühlmann, Peter. (2012). MissForest-non-parametric missing value imputation for mixed-type data. Bioinformatics. https://doi.org/10.1093/bioinformatics/btr597.

25. Groenwold, Rolf H.H.; White, Ian R.; Donders, A. Rogier T.; Carpenter, James R.; Altman, Douglas G.; Moons, Karel G.M.. (2012). Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. Canadian Medical Association Journal. https://doi.org/10.1503/cmaj.110977.

26. Stuart, Elizabeth A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science*, vol. 25, no. 1, 2010, pp. 1-21. https://doi.org/10.1214/09-sts313.

27. White, Halbert. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. https://doi.org/10.2307/1912934.

28. Santos, Miriam Seoane; Abreu, Pedro Henriques; García-Laencina, Pedro J.; Simão, Adélia; Carvalho, Armando. (2015b). A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. Journal of Biomedical Informatics. https://doi.org/10.1016/j.jbi.2015.09.012.

29. von Elm, Erik, et al. "The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies." *PLoS Medicine*, vol. 4, no. 10, 2007, pp. e296. https://doi.org/10.1371/journal.pmed.0040296.

30. Wolff, Robert F., et al. "PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies." *Annals of Internal Medicine*, vol. 170, no. 1, 2019, pp. 51-58. https://doi.org/10.7326/m18-1376.