





Causal Considerations Can Inform the Interpretation of Surprising Associations in Medical Registries

Alberto Carmona-Bayonas, Paula Jiménez-Fonseca, Javier Gallego & Pavlos Msaouel


To cite this article: Alberto Carmona-Bayonas, Paula Jiménez-Fonseca, Javier Gallego & Pavlos Msaouel (2021): Causal Considerations Can Inform the Interpretation of Surprising Associations in Medical Registries, *Cancer Investigation*, DOI: [10.1080/07357907.2021.1999971](https://doi.org/10.1080/07357907.2021.1999971)



To link to this article: <https://doi.org/10.1080/07357907.2021.1999971>

 View supplementary material 

 Published online: 25 Nov 2021.

 Submit your article to this journal 




 Article views: 117

 View related articles 

 View Crossmark data 



Causal Considerations Can Inform the Interpretation of Surprising Associations in Medical Registries

Alberto Carmona-Bayonas^a , Paula Jiménez-Fonseca^b , Javier Gallego^c and Pavlos Msaouel^d 

^aHematology and Medical Oncology Department, Hospital Universitario Morales Meseguer, UMU, IMIB, Murcia, Spain; ^bMedical Oncology Department, Hospital Universitario Central de Asturias, ISPA, Oviedo, Spain; ^cMedical Oncology Department, Hospital General de Elche, Elche, Spain; ^dDepartment of Genitourinary Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

ABSTRACT

An exploratory analysis of registry data from 2437 patients with advanced gastric cancer revealed a surprising association between astrological birth signs and overall survival (OS) with $p = 0.01$. After dichotomizing or changing the reference sign, p -values < 0.05 were observed for several birth signs following adjustments for multiple comparisons. Bayesian models with moderately skeptical priors still pointed to these associations. A more plausible causal model, justified by contextual knowledge, revealed that these associations arose from the astrological sign association with seasonality. This case study illustrates how causal considerations can guide analyses through what would otherwise be a hopeless maze of statistical possibilities.

ARTICLE HISTORY

Received 25 August 2021
Revised 24 October 2021
Accepted 26 October 2021

KEYWORDS

Zodiac sign; horoscope;
causal inference; Bayesian;
frequentist; seasonality

Introduction

Much has been written about inductive, hypothesis-free methods to study high-dimensional biological datasets obtained by next-generation sequencing and multi-omics integration without making strong assumptions about the underlying data-generating processes (1). In the hope that the observed data associations can reveal biological processes worth exploring further, analyses may be performed without precise, pre-specified, or contextually directed causal hypotheses (2). Although this approach can sometimes be of value (1), multiple authors have raised concerns, because such analyses can result in spurious correlations and false inferences, resulting in a lack of reproducibility, as can be demonstrated by genome-wide association studies (3).


The impact of hypotheses-free interrogations of clinical datasets, such as observational registries and *post-hoc* subgroup analyses from randomized clinical trials (RCTs), has received less attention. The analytical complexity of large-

scale clinical registries, in particular, is far greater (4,5). Real-world data (RWD) studies of this nature have increased by more than 600% in the last decade (6), with growing interest in using these resources to support clinical and regulatory decision-making (6–9).

To capture the underlying reality that leads to observing the generated data, causal inference calls for the introduction of multiple additional assumptions, such as confounding relationships between variables (10). If a refutational statistical test detects a violation of any assumption, the result is not attributable to a single, isolated hypothesis and all alternative mechanisms must be accounted for as possible competing explanations (11). Accordingly, improper inferences derived from large RWD studies are often due to incorrect parameterizations of the models, misattributions of causal links, incorrect selection of confounders, or use of statistical models based on implausible causal assumptions (12–15).

Categorical variables, those that can take a fixed, limited number of possible values, are

CONTACT Pavlos Msaouel  pmsaouel@mdanderson.org Department of Genitourinary Medical Oncology, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX 77030, USA.

 Supplemental data for this article can be accessed [here](#).

© 2021 Taylor & Francis Group, LLC

particularly prone to poor specification since the analysis depends on arbitrary assumptions that hinder the search for patterns between different levels (e.g., categorical variables can be coded or combined in multiple ways to reflect the desired comparison, among other issues) (16–18). All this amplifies model-possibility counting to an infinitely large number of potential specifications even from simple data (10).

Various strategies have been proposed to reduce false signals in this scenario, due to the enormous multiplicity of models (19). While there is no practical consensus on this matter (20–23) and the importance of context and costs has been underestimated (24), the statistical literature often invokes the need to make adjustments for multiple tests. Bayesian shrinkage is another way to reduce the probability of random fluctuations. Thus, the Bayesian approach assigns prior probabilities that can meaningfully quantify *a priori* knowledge and expectations (25,26). In estimating causal effects, assigning skeptical prior probabilities can be particularly useful to discourage results deemed implausible *a priori* (27). This skeptical perspective can be formalized through meticulously chosen priors and is a key aspect of Bayesian modeling (24,26,28). A limitation of the Bayesian approach is that in certain settings it can be very hard to specify a prior probability that is acceptable by all subject matter experts and data analysts (29–33). Although frequentist approaches do not quantify prior probabilities, they too encode prior knowledge and skepticism, for example, by using shrinkage-estimation methods to calibrate the repeated-sampling accuracy of their estimates (34). Admittedly, skepticism can enhance the reliability of our inferences, but it can also impede learning, particularly in cases of extreme skepticism. The statistician Dennis Lindley, a strong supporter of Bayesian methods, coined the term “Cromwell’s rule” to describe why such extreme skepticism can be counterproductive (35). In clinical research, extreme skepticism can unjustifiably favor the null hypotheses of no difference or no effect; a prejudice known as “nullism” (36). The fallacy of “absence of evidence is not evidence of absence”, which misleads researchers into failing to capture subtle but real effects stand out among the consequences of

nullism (37). The complementarity of Bayesian and frequentist procedures allows data analysts to examine data and models in alternative ways, detecting more sources of error and uncertainty, thereby lessening the risk of nullism and other biases (38,39). Bayesian and frequentist ideas can be incorporated in multilevel (hierarchical) modeling, which encompasses frequentist, Bayes, semi-Bayes, as well as shrinkage (empirical-Bayes) methods (24,34,40,41). However, as with all statistical methods, multilevel modeling is sensitive to the causal assumptions regarding the underlying data-generating process (41,42). Furthermore, while there is no single, universally valid approach to analyzing all medical RWD registry datasets, contextual clinical and biological information can facilitate the interpretation of RWD analyses.

Motivated by the above considerations, we set out to interrogate these concepts in a real-world setting using the gastric cancer registry AGAMENON. Unexpectedly, we found an association between survival outcomes and patient zodiac signs. By considering the causal network that may have generated this association, we were able to arrive at a plausible explanation for this association. Our results illustrate how rare and surprising findings can be satisfactorily explained through the correct specification of statistical models structured according to causal considerations informed by contextual knowledge that lies outside of the dataset itself.

Methods

Patients

The data comes from the AGAMENON-SEOM (SEOM is the Spanish acronym of the Sociedad Española de Oncología Médica) hospital-based, gastric cancer registry in which researchers from 37 Spanish institutions participate. Basic eligibility criteria include individuals >18 years old, with advanced tumors of the stomach, esophagus, or gastroesophageal junction. The database is managed through a website (<http://www.agamenonstudy.com/>) designed to guarantee data reliability and control for missing and inconsistent data, with telephone and online monitoring

(PJF). The overall characteristics of this database, including eligibility criteria, clinical aspects, quality criteria, data monitoring, baseline patient characteristics, and outcomes, have been previously reported (5,43–46).

Variables and study design

During an analysis of prognostic factors (5), we decided to use negative controls, such as the sign of the zodiac, having specified *a priori* that this variable does not affect survival. Such negative controls are used in epidemiology when the scientific community generally accepts beforehand that these variables lack any causal effect. Therefore, an unexpected result would suggest some problem in the data, the analytical approach, or the underlying assumptions. The analysis of the prognostic effect of the horoscope has also been used to illustrate various statistical pitfalls, such as the binning of categorical variables, or the need for multiple hypotheses to be tested, or as a source of implausible associations (47,48).

Overall survival (OS) was used as the analysis outcome because it is considered the gold standard clinical endpoint in oncology (49). OS was defined as the period from the beginning of the first line of chemotherapy until death, censoring subjects without an event at the time of analysis. The 12 signs of the zodiac and their associated elements (fire, earth, air, and water) were considered. Seasonality was incorporated by considering the fraction of the year that has passed, which results from dividing the full calendar time of birth by 366 or 365, depending on whether the year is a leap year or not. The secular year of birth was used to capture long-term trends, taking as a reference the year 1921 (birth of the oldest subject in the registry), adding the fraction of the elapsed year in progress at the date of birth.

Statistical analyses

The evaluation of the effect of the astrological sign on OS was carried out using the log-rank test and the likelihood-ratio test (where H_0 is a null hypothesis; i.e., the absence of effect). We then fitted several Cox proportional hazards (PH) models to compare the effect of individual zodiac signs.

Several reference categories, or clustering criteria, were used to exemplify the errors associated with dichotomization based on observed results (47). Holm-Bonferroni's method was applied to account for these multiple comparisons (50).

p-Values are a purely refutational metric that represents the probability that the chosen test statistic would be as or more extreme than observed given all the assumptions used to compute it (11,19,51). *S*-values are interpreted as bits of information against the test model and are intended to facilitate the translation of abstract statistical results as simple physical experiments, such as coin tossing (19,52). *S*-values are expressed as negative logarithms of the *p*-value and, when the base 2 is used for the logarithm, then the *S*-values are measured in bits of information against the tested hypothesis and background assumptions (19). For example, a *p*-value = 0.05 yields $-\log_2(0.05) \sim 4.32$ bits of information in the data against the test model (hypothesis and background assumptions), equivalent to the surprise we should feel when tossing a presumed fair coin 4 times and getting heads for all tosses.

Furthermore, Bayesian Cox PH models were fitted to evaluate the influence on the likelihood of progressively more skeptical priors $\sim N(0, 0.1)$ and $\sim N(0, 0.05)$ (25,53). Finally, a rigorous analysis of seasonality was conducted, taking into account orthogonal trigonometric functions in a chronobiological model (54). These models are formulated by including the following terms, where *t* and *T* represent the calendar day and total days of the year, while β_1 and β_2 are the coefficients from which the estimated phase and amplitude are derived:

$$\beta_1 \times \sin\left(\frac{2\pi t}{T}\right) + \beta_2 \times \cos\left(\frac{2\pi t}{T}\right)$$

The annual trends, added to the fraction of the corresponding year to avoid abrupt leaps, were modeled non-linearly using restricted cubic splines. The variation inflation factor (VIF) was used to evaluate the multicollinearity of the variables. Diagnostics of the Cox regression models were further performed using conventional information criteria, such as the Widely Applicable Information Criterion (WAIC), Pseudo-BMA, or the Leave-One-Out Cross-Validation (LOO-CV)

applied to attain the out-of-sample predictive performance (55). For frequentist models, this comparison was carried out using the Akaike information criterion (AIC) and Bayesian information criterion (BIC).

All analyses were performed with the R v4.0.3 software package, with the DescTools, survminer, survival, and brms libraries (56,57). Directed acyclic graphs (DAGs) were built using the Dagitty software (<http://www.dagitty.net/>) (58). The complete analysis R code used is available as [Supplementary File 1](#).

Results

Prognostic effect of the zodiac signs

At the time of the analysis, the AGAMENON-SEOM database had 2473 registered patients and

2057 death events. The log-rank test revealed a significant association between zodiac signs and OS with $\chi^2 = 23.0$ on 11 degrees of freedom (df), p -value = 0.01. We subsequently fitted Cox PH models with the horoscope as a categorical variable (Figure 1). In model 1, the category “Capricorn” was used as a reference, because it had the highest number of events ($N=203$) (Figure 1). Once again, the zodiac sign was associated with OS (Likelihood ratio test = 24.0 on 11 df, $p=0.01$). This is equivalent to ~ 7 bits of information against the null hypothesis and background assumptions. Note here that these background assumptions presuppose that the zodiac sign (expressed as a categorical variable) is reflective of the underlying data-generating process. The p -value and corresponding S-value cannot on their own discern which aspects of the tested model (null hypothesis or background

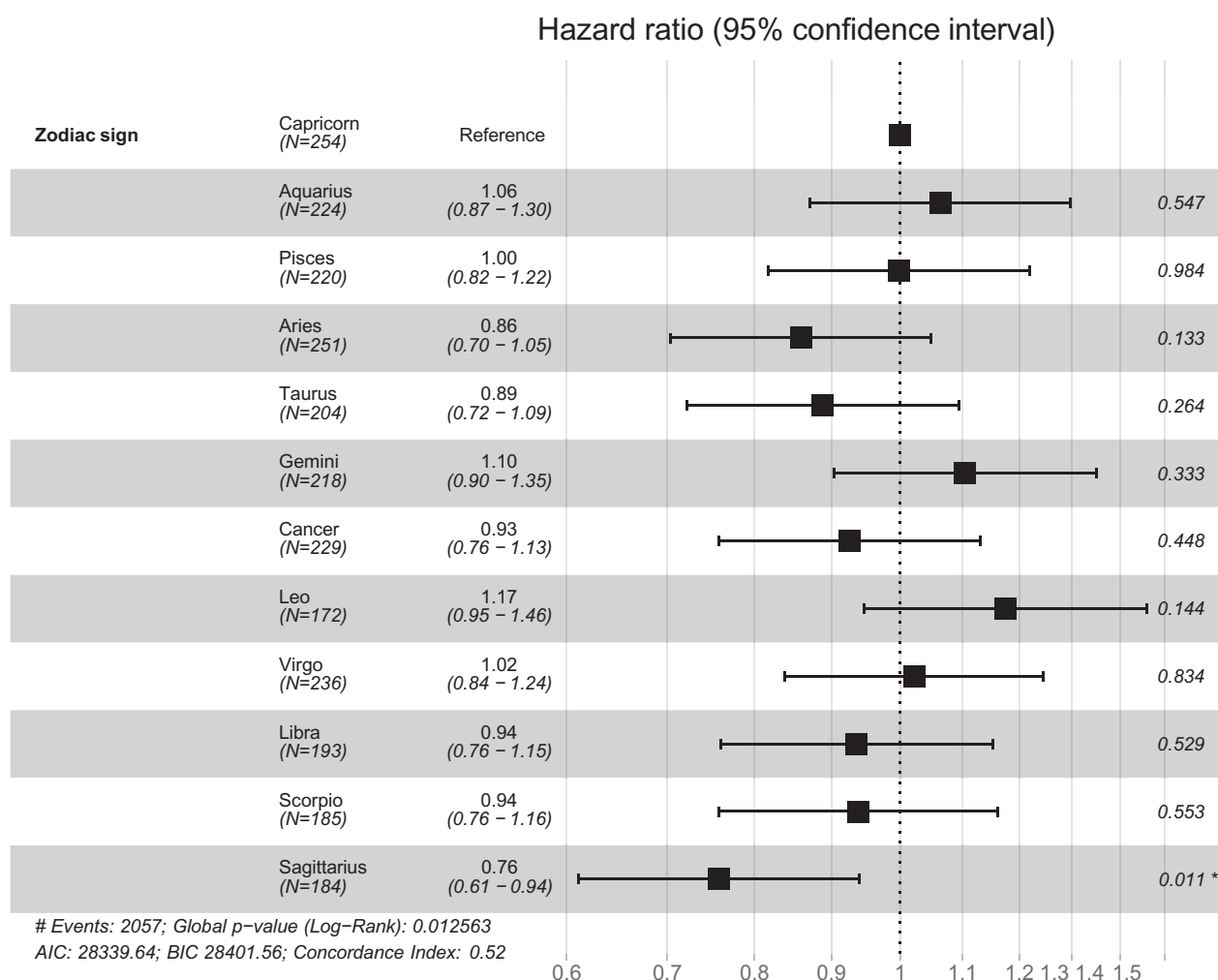


Figure 1. Cox proportional hazards model for overall survival. The model has been parameterized here considering the zodiac as a categorical variable. Capricorn subjects have been considered as the reference since they constitute the group containing the most individuals and events. AIC: Akaike information criterion; BIC: Bayesian information criterion.

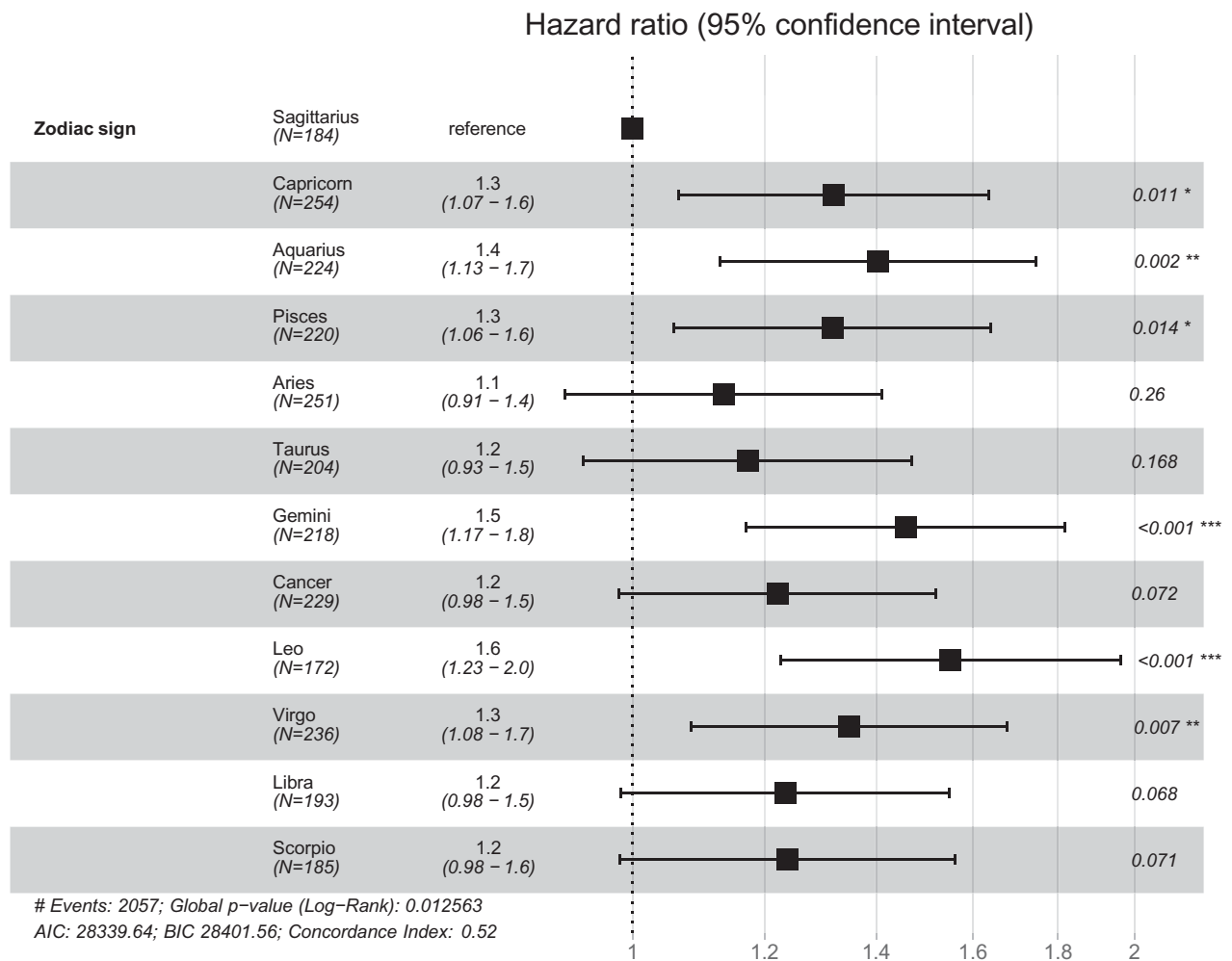


Figure 2. Cox proportional hazards model for overall survival. On this occasion, the Sagittarius subjects are the reference. The decision was made based on the observed results of model 1. The reader can appreciate that the consequence of this decision is that multiple categories are significantly associated with survival.

assumptions) are refuted by the data. According to the specification of this model 1, participants born under the sign of Sagittarius had a better prognosis, with an OS hazard ratio (HR) of 0.76 (95% confidence interval [CI], 0.61–0.94), p -value = 0.011.

Model-possibility counting and multiple comparisons

Model 1 is one of the thousands of possible specifications. As they are combinations without replacement, the number of pairwise comparisons (Wald tests) between the 12 elements of the zodiac (e.g., Capricorn vs. Sagittarius) is:

$$\binom{12}{2} = 66$$

Given that the choice between them is arbitrary, the validity of multiplicity adjustment is

equally valid for any of these specifications. Applying the binomial formula, the probability of reaching a false conclusion in any of them is 96.7%:

$$1 - \binom{66}{0} (0.05^0) (0.95^{66-0})$$

For example, if Sagittarius is considered the reference category (model 2), then six of the zodiac signs have a worse prognosis ($p < 0.05$) (Figure 2). After adjusting for multiple comparisons, both Gemini and Leo were still associated with worse outcomes ($p < 0.05$). Model-possibility counting scales up quickly when considering more complex models. The number of ways a set with N elements can be partitioned into disjoint, non-empty subsets is described by the n th Bell number minus one; therefore the zodiac sign can be partitioned in 4,213,596 ways, each one

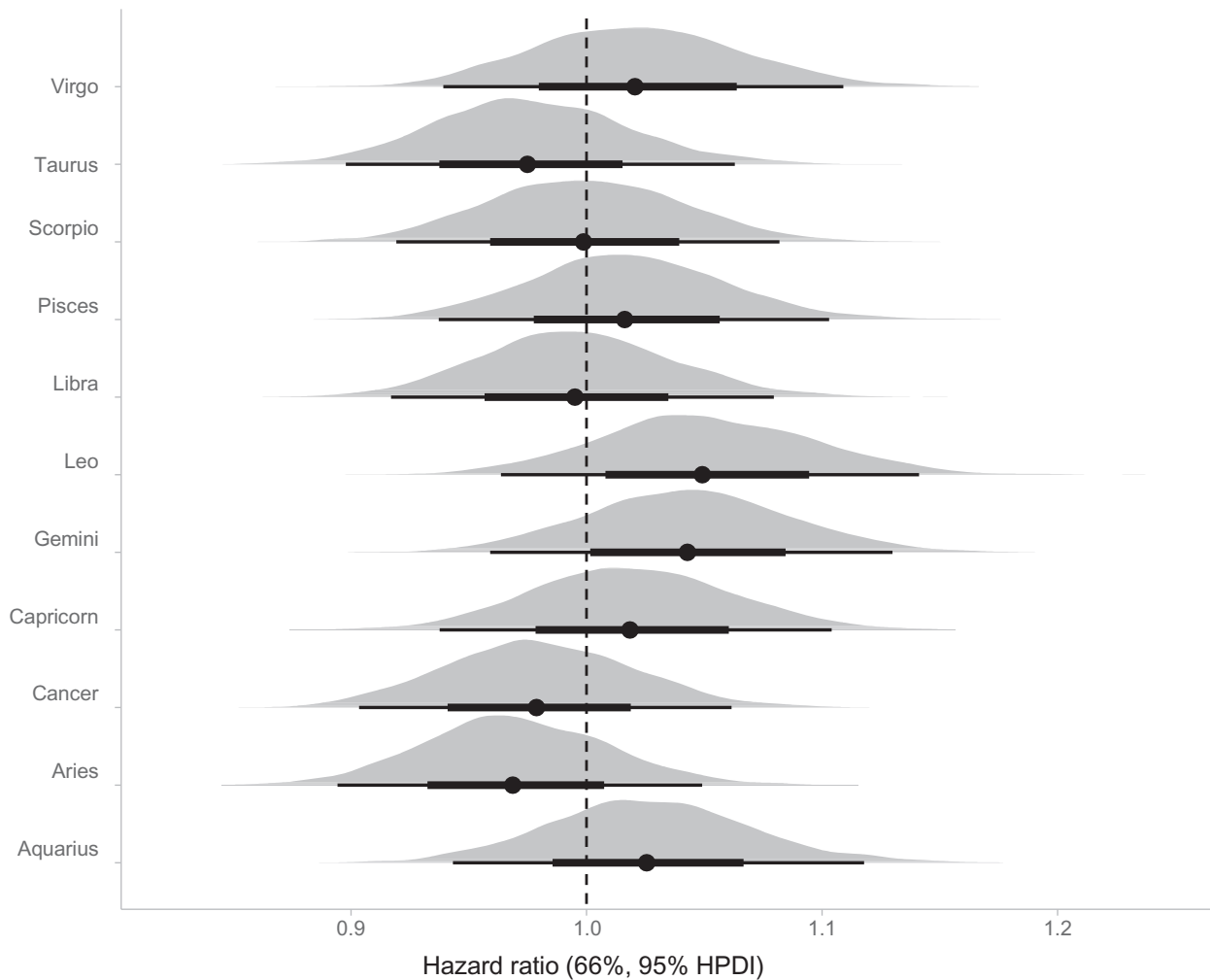


Figure 3. Half-eye plot with the (Bayesian) Cox proportional hazards model with a skeptical prior. The Sagittarians constitute the reference group. The skeptical prior, normal (0, 0.05), conditions the posterior distribution of effect, discouraging implausible results, but ultimately prevents learning.

representing a different model (59). In one of these models, Sagittarians have a higher median OS than all others combined (13.4 vs. 10.5 months), with HR 0.79 (95% CI, 0.61–0.93), p -value = 0.011, corresponding ~ 7 bits of refutational information (Figure 1). All these considerations are similar when the parameterization is performed with the Gregorian calendar (data not shown).

Skeptical Bayesian approach

To further interrogate the unexpected association of OS with a patient zodiac sign, we fitted Bayesian versions of these Cox PH models. In model 3, a moderately skeptical prior $\sim N(0, 0.1)$ with Capricorn as reference category was used. Even under this skeptical model, Sagittarians

demonstrated a better prognosis with HR 0.87 (95% credible interval [CrI], 0.78–0.97), which represents a posterior probability of favorable effect of 98%. Model 4 uses a more skeptical prior $\sim N(0, 0.05)$, but this time with Sagittarian patients as the reference category. With this highly skeptical model, none of the other zodiac elements demonstrated a substantially worse OS compared with Sagittarians (Figure 3). For example, Leos demonstrated an HR of 1.05 (95% CrI, 0.96–1.13), with a posterior probability of effect size >0 of $\sim 60\%$ compared with Sagittarians.

Chronobiological model

We decided to use context-derived causal considerations to narrow our statistical considerations

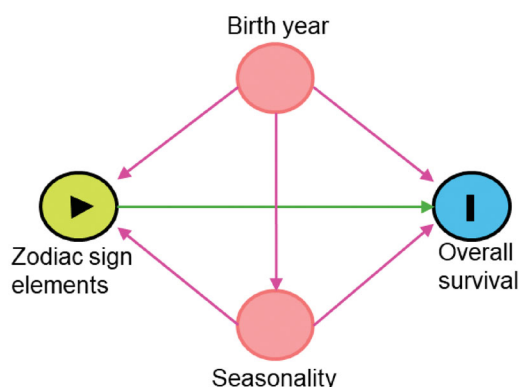


Figure 4. Directed acyclic graph of the chronobiological causal model used to investigate the relationship between zodiac sign elements (exposure) and overall survival (outcome) in our dataset. Seasonality is associated with the zodiac sign elements at birth and can affect overall survival through multiple plausible mechanisms, such as gestational nutrition, seasonal infections, and sunlight exposure. A birth year can influence seasonality mechanisms (e.g., seasonal effects can be attenuated by modern infection controls and food distribution) as well as overall survival. Seasonality and birth year are confounders that must be accounted for to determine the true causal effect of the zodiac sign on overall survival.

based on plausible models of the underlying reality. We accordingly generated a DAG to guide further analyses of our dataset (Figure 4). The DAG is based on the hypothesis that seasonal mechanisms, such as sunlight exposure, gestational nutrition, and infections, can affect OS and are reflected in the astrological month of birth (60). Furthermore, we assumed that birth year can also affect seasonality (for example, modern infection controls and food distribution can flatten seasonal effects), in addition to its association with zodiac signs and overall survival. The DAG suggested that seasonality and birth year are confounders of the association between zodiac signs and OS. A frequentist Cox PH model was accordingly built that adjusted for both the seasonality of the date of birth and the year of birth (model 5). Under this simple model, none of the zodiac signs was significantly associated with OS. However, model diagnosis revealed high multicollinearity for the coefficients associated with seasonality (correlation with the horoscope). To mitigate this effect, the astrological signs were gathered into their four basic elements (model 6) as described in the Methods section. This procedure drastically reduced the multiplicity of models. The final result does not refute the null

hypothesis that the zodiac sign does not independently influence prognosis (Likelihood test = $\chi^2 = 1.028$, on 3 df, p -value = 0.794), corresponding to ~ 0 bits of refutational information (see Figure 5). The model diagnostics are shown in Supplementary Figure 1. No multicollinearity, influential observations (data points whose deletion would noticeably change the result of the estimate), outliers, or violation of the PH assumption were observed for any of the covariates. When comparing out-of-sample performance, the chronobiological model 6 is superior to the naïve model 1, according to the BIC (28391.11 vs. 28401.56, respectively), but not according to the AIC (28351.71 vs. 28339.64). We next fitted the Bayesian counterpart of the chronobiological model using a normal (0,1) prior. The estimates are comparable to the Cox frequentist model (data not shown). Data-based diagnostics, such as LOO-CV (0.003 vs. 0.996), WAIC (0.003 vs. 0.996), and Pseudo-BMA weighting (0.23 vs. 0.77) suggested that the chronobiological model did not represent the data-generating process more plausibly.

Discussion

In the present study, we used multiple different approaches to investigate a surprising result that arose during the survival analysis of a real-world cancer registry. Our initial statistical model suggested an association between patient zodiac signs and cancer prognosis as measured by OS. Causal considerations, based on contextual knowledge outside the dataset itself, provided a plausible chronobiological explanation of this association. The resultant chronobiological model suggested that both seasonality and birth year were confounders for the association between zodiac signs and OS. Adjustment for both these confounders negated the effect of zodiac signs on OS in our patient cohort.

The initially surprising association was due to the erroneous translation of an implausible causal construct into our statistical models, which ignored that the horoscope (or the month of the year) represents the effect of continuous calendar time on OS (61,62). Discretization of continuous data by the horoscope yielded arbitrary statistical

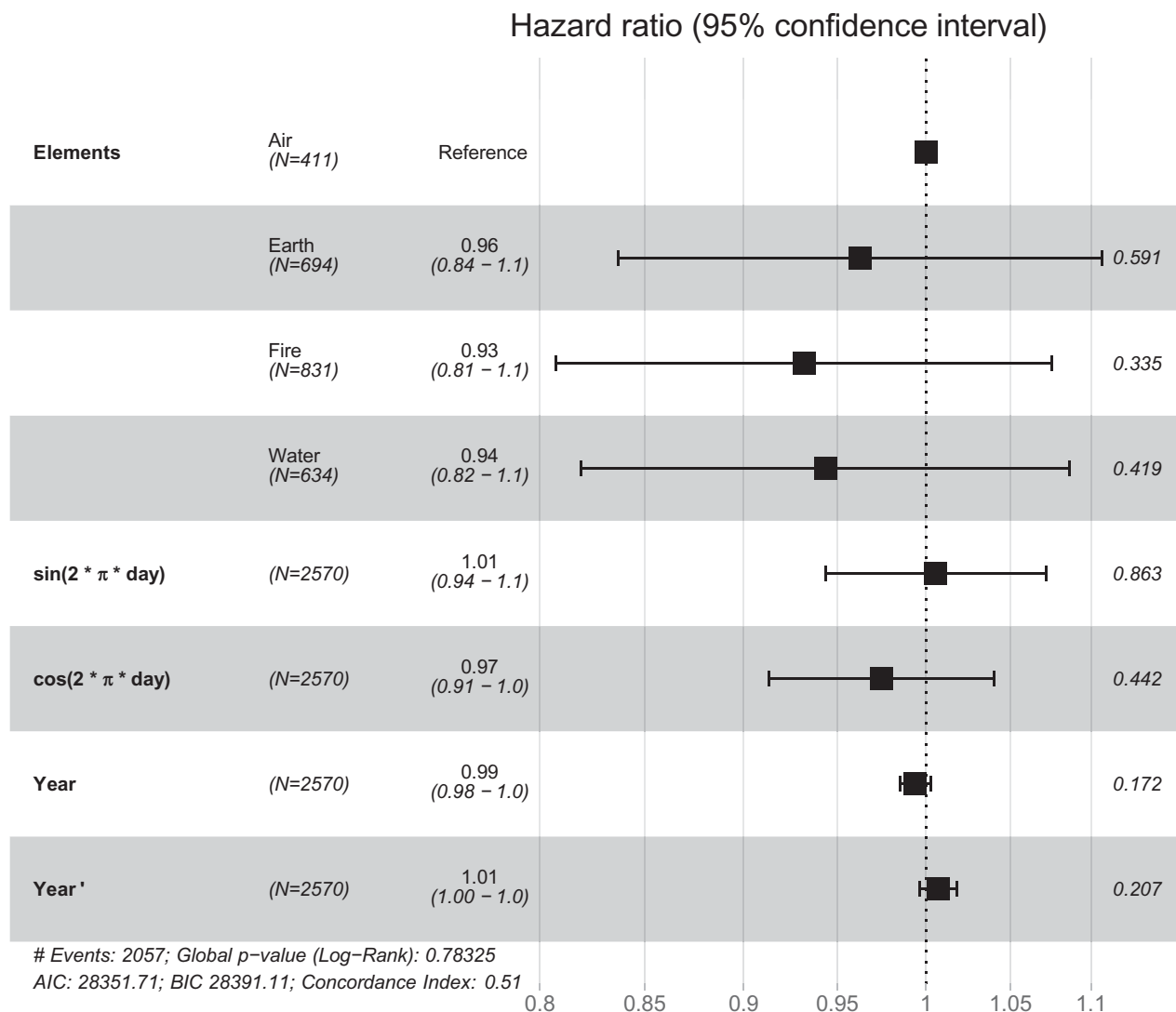


Figure 5. Chronobiological model. Calendar time is captured by orthogonal trigonometric functions (sine and cosine). To avoid multicollinearity, the zodiac is expressed through its elements (see Methods).

models that ignored the fundamental mechanism of seasonality. Even when focused solely on the zodiac sign, the potential possible models are up to 4,213,596. However, the true estimate of potential analyses is unquantifiable. This problem is accentuated even further with continuously modeled data. For instance, if birth days were used instead of birth months, the number of imaginable models would make 4,213,596 seem miniscule. This hopeless maze of equally valid alternatives cannot be navigated with hypothesis-free approaches that are unable to select plausible models from the innumerable possible alternatives (10). The first important step is to accept that the overwhelming majority of logically possible partitions and parameterizations would be

immediately rejected as complex beyond any practical or empirical justification. Thus, parsimony heuristics allow us to narrow our models down to a smaller, finite collection, although this still fails to fully resolve the issue (63). For example, approaches attempting to naively simplify models, e.g., by dichotomizing k levels into a 2×2 table, do not fully resolve the exuberant proliferation of models and further compromise the analysis by introducing arbitrary discontinuities, especially if dichotomization is performed based on “conveniently” observed results (17,61).

Controlling family-wise error rates introduces additional dilemmas, such as the need to define what is considered a family of tests, and increases the probability of type II errors (64). In an

extreme case, adjustments would be needed for comparisons that the analyst has not even considered carrying out (24). Had we casually rejected the association between OS and the zodiac sign as a random “fluke”, we would have missed a potentially interesting hypothesis-generating signal connecting OS with seasonality. Multiplicity considerations should factor in the cost of missed discoveries relative to false discoveries (65).

The Bayesian approach shares the same need for appropriate specification of analyses models (24,38). Bayesian analyses can compute the probability of hypotheses as a function of the data and allow the incorporation of prior knowledge through plausible prior probability distributions (25,26,28). However, the use of strongly skeptical priors, such as spiked priors focusing on the null effect, presupposes that we have very strong prior evidence that seasonality in no way affects OS in patients with advanced gastric cancer. Taken to the extreme, such “nullism” hinders the acquisition of evidence from the data (38). Some of the most successful scientific strategies have been inspired by investigating anomalies (e.g., outliers or unexpected observations for theory), as proposed by philosophers of science, such as Popper, Kuhn, and Lakatos (66–69). Therefore, strongly skeptical priors, while they may be useful in certain contexts (3), do not substitute for the need to properly parametrize our statistical models based on plausible causal frameworks.

Researchers may be tempted to carelessly perform statistical analyses based on limited rationale without taking advantage of the contextually rich background information available within the scientific community (38). Moreover, research objectives may be arbitrarily changed, hypotheses modified to adapt to observed results, subgroups modeled without having contextually-rich or reliable data, and *post-hoc* analyses may be misinterpreted (70) leading to the proliferation of implausible models and conclusions (71). In this case study, the 7 bits of refutational information we initially found alerted us to a problem in our background assumptions and warned us of the lack of causal context in our analysis (11,72). Some of what we today treat as pure superstition had its origin in observations accrued over many generations. Vague astrological notions were not

outlandish at a time when seasonal effects were likely far more profound than in modern generations. The contemporary null spike on astrological effects is a product of precise astrometry that emerged from our modern era. On the other hand, an association between seasonality and survival outcomes has recently been suggested for pediatric tumors (73) and other pathologies (74–76), and adjusting for seasonality is therefore a sound strategy. More generally, the initial unexpected association between the zodiac sign and survival is an example of a problem enunciated by the theoretical physicist and philosopher of science Pierre Duhem more than 100 years ago: the impossibility of evaluating a hypothesis in isolation without causal models that require assumptions or auxiliary hypotheses (42,77). These challenges are usually more prominent in explanatory analyses of large datasets (78).

There can be causally incorrect models that make better predictions than causally correct ones (79). Hence, measures of model fit may not on their own be helpful in selecting the best causal model. Thus, the conventional information criteria, such as AIC, BIC, WAIC, or the LOO-CV procedure, do not supplant the theoretical justification of our statistical models. Such complex causal relationships can be represented by DAGs, which help visualize causal relationships between variables based on plausible and preferably *a priori* specified causal models that are external to the data (12,15). In our case, neither AIC nor WAIC was capable of identifying the most causally correct model. In contrast, the BIC suggested the best explanatory model that accounted for seasonality and birth year. This is consistent with the notion that, whereas the AIC is suitable for prediction because it is asymptotically equivalent to cross-validation, the BIC is more appropriate for selecting causal models because it attempts to estimate the underlying data-generating process (80). This illustrates how prediction and explanation are fundamentally different processes, although they are often crudely interchanged (78,81).

In conclusion, this case study illustrates the importance of using carefully developed statistical analysis models, supported by plausible background assumptions derived from contextual

knowledge outside of the dataset itself. Our analysis warns against the indiscriminate use of priors focused on the null effect and of measures of goodness-of-fit, which cannot substitute for sound causal reasoning. When faced with unexpected results, all statistical analysis steps should be carefully and transparently audited in light of the available knowledge.

Acknowledgments

Various statisticians and clinicians have commented on these results in <https://discourse.datamethods.org>, providing informal insights into the results. The insights of Sander Greenland greatly contributed to this manuscript, although we alone are responsible for any possible errors. We thank the anonymous reviewer for providing many insightful comments and suggestions. We also thank the IRICOM S.L. team for the support of the website registry.

Ethical approval

All procedures followed were in accordance with the ethical standards of the committee responsible for human experimentation (institutional and national) and with the Helsinki Declaration of 1964 and later versions. Informed consent was obtained from all patients before being included in the study. This work is original and has not been previously presented elsewhere. All authors have approved the manuscript and publication.

Author contributions

P.J.F., A.C.B., J.G., and P.M. developed the project, analyzed the data, and drafted the manuscript. The other authors recruited patients and provided clinical information, comments, and improvements to the manuscript. All authors participated in the interpretation and discussion of data, and the critical review of the manuscript.

Declaration of interest

P.M. has received honoraria for service on a Scientific Advisory Board for Mirati Therapeutics, Bristol Myers Squibb, and Exelixis; consulting for Axiom Healthcare Strategies; non-branded educational programs supported by Exelixis and Pfizer; and research funding for clinical trials from Takeda, Bristol Myers Squibb, Mirati Therapeutics, Gateway for Cancer Research, and UT MD Anderson Cancer Center. All other authors state no conflict of interest related to this study.

Funding

This is an academic study. The study was supported by the authors themselves. Pavlos Msaouel is supported by a Career Development Award by the American Society of Clinical Oncology, a Research Award by KCCure, the MD Anderson Khalifa Scholar Award, the Andrew Sabin Family Foundation Fellowship, and the MD Anderson Physician-Scientist Award.

ORCID

Alberto Carmona-Bayonas  <http://orcid.org/0000-0002-1930-9660>

Paula Jiménez-Fonseca  <http://orcid.org/0000-0003-4592-3813>

Pavlos Msaouel  <http://orcid.org/0000-0001-6505-8308>

Data availability statement

All data will be provided upon request to the corresponding author.

Code availability

The R code is available as [Supplementary File 1](#).

References

1. Knight J. Hypothesis-free research. *Trends Genet.* 2000;16:25. doi:10.1016/S0168-9525(00)00104-9.
2. Efron B, Hastie T. Computer age statistical inference, student edition: algorithms, evidence, and data Science. Cambridge: Cambridge University Press; 2021. (Institute of Mathematical Statistics Monographs).
3. Eicher T, Kinnebrew G, Patt A, Spencer K, Ying K, Ma Q, et al. Metabolomics and multi-omics integration: a survey of computational methods and resources. *Metabolites.* 2020;10(5):202. doi:10.3390/metabo10050202.
4. Carmona-Bayonas A, Jimenez-Fonseca P, Garrido M, Custodio A, Hernandez R, Lacalle A, et al. Multistate models: accurate and dynamic methods to improve predictions of thrombotic risk in patients with cancer. *Thromb Haemost.* 2019;119(11):1849–1859. doi:10.1055/s-0039-1694012.
5. Jimenez-Fonseca P, Carmona-Bayonas A, Martínez de Castro E, Custodio A, Pericay Pijaume C, Hernandez R, et al. External validity of docetaxel triplet trials in advanced gastric cancer: are there patients who still benefit? *Gastric Cancer.* 2021;24(2):445–456. doi:10.1007/s10120-020-01116-x.
6. Karim S, Booth CM. Effectiveness in the absence of efficacy: cautionary tales from real-world evidence. *J*

- Clin Oncol. 2019;37(13):1047–1050. doi:10.1200/JCO.18.02105.
7. Cave A, Kurz X, Arlett P. Real-world data for regulatory decision making: challenges and possible solutions for Europe. *Clin Pharmacol Ther.* 2019;106(1):36–39. doi:10.1002/cpt.1426.
 8. Msaouel P, Lee J, Thall PF. Making patient-specific treatment decisions using prognostic variables and utilities of clinical outcomes. *Cancers.* 2021;13(11):2741. doi:10.3390/cancers13112741.
 9. Msaouel P. Impervious to randomness: confounding and selection biases in randomized clinical trials. *Cancer Invest.* 2021;39(10):783–788. doi:10.1080/07357907.2021.1974030.
 10. Patel CJ, Burford B, Ioannidis JP. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *J Clin Epidemiol.* 2015;68(9):1046–1058. doi:10.1016/j.jclinepi.2015.05.029.
 11. Greenland S, Rafi Z. To aid scientific inference, emphasize unconditional descriptions of statistics. *arXiv Preprint 2019;arXiv:190908583.*
 12. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology.* 1999;10(1):37–48.
 13. Castañón E, Sanchez-Arreaez A, Alvarez-Manceñido F, Jimenez-Fonseca P, Carmona-Bayonas A. Critical reappraisal of phase III trials with immune checkpoint inhibitors in non-proportional hazards settings. *Eur J Cancer.* 2020;136:159–168. doi:10.1016/j.ejca.2020.06.003.
 14. van Rein N, Cannegieter SC, Rosendaal FR, Reitsma PH, Lijfering WM. Suspected survivor bias in case-control studies: stratify on survival time and use a negative control. *J Clin Epidemiol.* 2014;67(2):232–235. doi:10.1016/j.jclinepi.2013.05.011.
 15. Shapiro DD, Msaouel P. Causal diagram techniques for urologic oncology research. *Clin Genitourin Cancer.* 2021;19(3):271.e1–271.e7. doi:10.1016/j.clgc.2020.08.003.
 16. Rutkowski L, Svetina D, Liaw Y-L. Collapsing categorical variables and measurement invariance. *Struct Equat Model.* 2019;26(5):790–802. doi:10.1080/10705511.2018.1547640.
 17. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med.* 2006;25(1):127–141. doi:10.1002/sim.2331.
 18. Altman DG. Categorising continuous variables. *Br J Cancer.* 1991;64(5):975. doi:10.1038/bjc.1991.441.
 19. Rafi Z, Greenland S. Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. *BMC Med Res Methodol.* 2020;20(1):244. doi:10.1186/s12874-020-01105-9.
 20. Stacey AW, Pouly S, Czyz CN. An analysis of the use of multiple comparison corrections in ophthalmology research. *Invest Ophthalmol Vis Sci.* 2012;53(4):1830–1834. doi:10.1167/iov.11-8730.
 21. Dmitrienko A, D’Agostino RB. Multiplicity considerations in clinical trials. *N Engl J Med.* 2018;378(22):2115–2122. doi:10.1056/NEJMr1709701.
 22. Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology.* 1990;1(1):43–46.
 23. Hollis S, Fletcher C, Lynn F, Urban H-J, Branson J, Burger H-U, et al. Best practice for analysis of shared clinical trial data. *BMC Med Res Methodol.* 2016;16(Suppl 1):76. doi:10.1186/s12874-016-0170-y.
 24. Greenland S, Hofman A. Multiple comparisons controversies are about context and costs, not frequentism versus Bayesianism. *Eur J Epidemiol.* 2019;34(9):801–808. doi:10.1007/s10654-019-00552-z.
 25. van de Schoot R, Kaplan D, Denissen J, Asendorpf JB, Neyer FJ, van Aken MAG, et al. A gentle introduction to Bayesian analysis: applications to developmental research. *Child Dev.* 2014;85(3):842–860. doi:10.1111/cdev.12169.
 26. Thall PF. Statistical remedies for medical researchers. *Springer Series in Pharmaceutical Statistics.* New York (NY): Springer Publishing; 2020.
 27. Pedroza C, Han W, Truong VTT, Green C, Tyson JE. Performance of informative priors skeptical of large treatment effects in clinical trials: a simulation study. *Stat Methods Med Res.* 2018;27(1):79–96. doi:10.1177/0962280215620828.
 28. McElreath R. Statistical rethinking: a Bayesian course with examples in R and Stan. 2nd ed. Boca Raton (FL): Taylor and Francis, CRC Press; 2020. (CRC texts in statistical science).
 29. Sander G. Relaxation penalties and priors for plausible modeling of nonidentified bias sources. *Statist Sci.* 2009;24(2):195–210. doi:10.1214/09-STS291.
 30. Greenland S. Comment: the need for syncretism in applied statistics. *Statist Sci.* 2010;25(2):158–161. doi:10.1214/10-STS308A.
 31. Gelman A, Simpson D, Betancourt M. The prior can often only be understood in the context of the likelihood. *Entropy.* 2017;19(10):555. doi:10.3390/e19100555.
 32. Greenland S. Sensitivity analysis, Monte Carlo risk analysis, and Bayesian uncertainty assessment. *Risk Anal.* 2001;21(4):579–583. doi:10.1111/0272-4332.214136.
 33. Senn S. You may believe you are a Bayesian but you are probably wrong. *Ration Markets Morals.* 2011;2(42):48–66.
 34. Greenland S. Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *Int J Epidemiol.* 2006;35(3):765–775. doi:10.1093/ije/dyi312.
 35. Lindley DV. Making decisions. 2nd ed. London; New York (NY): Wiley; 1985.

36. Greenland S. Invited commentary: the need for cognitive science in methodology. *Am J Epidemiol.* 2017; 186(6):639–645. doi:[10.1093/aje/kwx259](https://doi.org/10.1093/aje/kwx259).
37. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ.* 1995;311(7003):485. doi:[10.1136/bmj.311.7003.485](https://doi.org/10.1136/bmj.311.7003.485).
38. Greenland S, Poole C. Rejoinder: living with statistics in observational research. *Epidemiology.* 2013;24(1): 73–78. doi:[10.1097/EDE.0b013e3182785a49](https://doi.org/10.1097/EDE.0b013e3182785a49).
39. James B. The case for objective Bayesian analysis. *Bayesian Anal.* 2006;1(3):385–402. doi:[10.1214/06-BA115](https://doi.org/10.1214/06-BA115).
40. Greenland S. Methods for epidemiologic analyses of multiple exposures: a review and comparative study of maximum-likelihood, preliminary-testing, and empirical-Bayes regression. *Stat Med.* 1993;12(8): 717–736. doi:[10.1002/sim.4780120802](https://doi.org/10.1002/sim.4780120802).
41. Greenland S. Principles of multilevel modelling. *Int J Epidemiol.* 2000;29(1):158–167. doi:[10.1093/ije/29.1.158](https://doi.org/10.1093/ije/29.1.158).
42. Greenland S. Chapter 31: the causal foundations of applied probability and statistics. In: Dechter R, Halpern J, Geffner H, editors. *Probabilistic and causal inference: the works of Judea Pearl*. New York (NY): ACM Books; 2021.
43. Jiménez Fonseca P, Carmona-Bayonas A, Hernández R, Custodio A, Cano JM, Lacalle A, et al. Lauren subtypes of advanced gastric cancer influence survival and response to chemotherapy: real-world data from the AGAMENON National Cancer Registry. *Br J Cancer.* 2017;117(6):775–782. doi:[10.1038/bjc.2017.245](https://doi.org/10.1038/bjc.2017.245).
44. Cotes Sanchís A, Gallego J, Hernandez R, Arrazubi V, Custodio A, Cano JM, et al. Second-line treatment in advanced gastric cancer: data from the Spanish AGAMENON registry. *PLOS One.* 2020;15(7): e0235848. doi:[10.1371/journal.pone.0235848](https://doi.org/10.1371/journal.pone.0235848).
45. Custodio A, Carmona-Bayonas A, Jiménez-Fonseca P, Sánchez ML, Viudez A, Hernández R, et al. Nomogram-based prediction of survival in patients with advanced oesophagogastric adenocarcinoma receiving first-line chemotherapy: a multicenter prospective study in the era of trastuzumab. *Br J Cancer.* 2017;116(12):1526–1535. doi:[10.1038/bjc.2017.122](https://doi.org/10.1038/bjc.2017.122).
46. Carmona-Bayonas A, Jiménez-Fonseca P, Echavarría I, Sánchez Cánovas M, Aguado G, Gallego J, et al. Surgery for metastases for esophageal-gastric cancer in the real world: data from the AGAMENON national registry. *Eur J Surg Oncol.* 2018;44(8): 1191–1198. doi:[10.1016/j.ejso.2018.03.019](https://doi.org/10.1016/j.ejso.2018.03.019).
47. Austin PC, Mamdani MM, Juurlink DN, Hux JE. Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. *J Clin Epidemiol.* 2006;59(9):964–969. doi:[10.1016/j.jclinepi.2006.01.012](https://doi.org/10.1016/j.jclinepi.2006.01.012).
48. Szydło RM, Gabriel I, Olavarria E, Apperley J. Sign of the Zodiac as a predictor of survival for recipients of an allogeneic stem cell transplant for chronic myeloid leukaemia (CML): an artificial association. *Transplant Proc.* 2010;42(8):3312–3315. doi:[10.1016/j.transproceed.2010.07.036](https://doi.org/10.1016/j.transproceed.2010.07.036).
49. Korn EL, Freidlin B, Abrams JS. Overall survival as the outcome for randomized clinical trials with effective subsequent therapies. *J Clin Oncol.* 2011;29(17): 2439–2442. doi:[10.1200/JCO.2011.34.6056](https://doi.org/10.1200/JCO.2011.34.6056).
50. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat.* 1979;6(2):65–70.
51. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol.* 2016;31(4):337–350. doi:[10.1007/s10654-016-0149-3](https://doi.org/10.1007/s10654-016-0149-3).
52. Cole SR, Edwards JK, Greenland S. Surprise! *Am J Epidemiol.* 2021;190(2):191–193. doi:[10.1093/aje/kwaa136](https://doi.org/10.1093/aje/kwaa136).
53. Bendtsen M. A gentle introduction to the comparison between null hypothesis testing and Bayesian analysis: reanalysis of two randomized controlled trials. *J Med Internet Res.* 2018;20(10):e10873. doi:[10.2196/10873](https://doi.org/10.2196/10873).
54. Stolwijk AM, Straatman H, Zielhuis GA. Studying seasonality by using sine and cosine functions in regression analysis. *J Epidemiol Community Health.* 1999;53(4):235–238. doi:[10.1136/jech.53.4.235](https://doi.org/10.1136/jech.53.4.235).
55. Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat Comput.* 2017;27(5):1413–1432. doi:[10.1007/s11222-016-9696-4](https://doi.org/10.1007/s11222-016-9696-4).
56. Therneau TM, Grambsch PM. *Modeling survival data: extending the Cox model*. New York (NY): Springer; 2000. (Statistics for biology and health).
57. Bürkner P-C. brms: An R package for bayesian multi-level models using Stan. *J Stat Soft.* 2017;80(1):1–28. doi:[10.18637/jss.v080.i01](https://doi.org/10.18637/jss.v080.i01).
58. Textor J, van der Zander B, Gilthorpe MS, Liskiewicz M, Ellison GT. Robust causal inference using directed acyclic graphs: the R package ‘dagitty’. *Int J Epidemiol.* 2016;45(6):1887–1894. doi:[10.1093/ije/dyw341](https://doi.org/10.1093/ije/dyw341).
59. Becker HW, Riordan J. The arithmetic of Bell and Stirling numbers. *Am J Math.* 1948;70(2):385–394. doi:[10.2307/2372336](https://doi.org/10.2307/2372336).
60. Barnett AG, Dobson AJ. *Analysing seasonal health data*. Berlin; London: Springer; 2010. (Statistics for biology and health).
61. Austin PC, Goldwasser MA. Pisces did not have increased heart failure: data-driven comparisons of binary proportions between levels of a categorical variable can result in incorrect statistical significance levels. *J Clin Epidemiol.* 2008;61(3):295–300. doi:[10.1016/j.jclinepi.2007.05.007](https://doi.org/10.1016/j.jclinepi.2007.05.007).
62. Thoresen M. Spurious interaction as a result of categorization. *BMC Med Res Methodol.* 2019;19(1):28. doi:[10.1186/s12874-019-0667-2](https://doi.org/10.1186/s12874-019-0667-2).

63. Kelly TK. Simplicity, truth, and probability. In: Prasanta B, Malcolm F, editors. *Handbook on the philosophy of statistics*. Dordrecht: Elsevier; 2010. p. 983–1026.
64. Glickman ME, Rao SR, Schultz MR. False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. *J Clin Epidemiol*. 2014;67(8):850–857. doi:10.1016/j.jclinepi.2014.03.012.
65. Greenland S. Analysis goals, error-cost sensitivity, and analysis hacking: Essential considerations in hypothesis testing and multiple comparisons. *Paediatr Perinat Epidemiol*. 2021;35(1):8–23. doi:10.1111/ppe.12711.
66. Kuhn TS, Science IS. Religion. The structure of scientific revolutions. Chicago (IL): University of Chicago Press; 1996.
67. Greenland S. Induction versus Popper: substance versus semantics. *Int J Epidemiol*. 1998;27(4):543–548. doi:10.1093/ije/27.4.543.
68. Popper KR. *The logic of scientific discovery*. London; New York (NY): Routledge; 1992.
69. Lakatos I, Feyerabend P, Motterlini M. *For and against method: including Lakatos's lectures on scientific method and the Lakatos-Feyerabend correspondence*. Chicago (IL): University of Chicago Press; 1999.
70. Spears MR, James ND, Sydes MR. 'Thursday's child has far to go'-interpreting subgroups and the STAMPEDE trial. *Ann Oncol*. 2017;28(10):2327–2330. doi:10.1093/annonc/mdx410.
71. Carmona-Bayonas A, Jimenez-Fonseca P, Fernández-Somoano A, Álvarez-Manceño F, Castañón E, Custodio A, et al. Top ten errors of statistical analysis in observational studies for cancer research. *Clin Transl Oncol*. 2018;20(8):954–965. doi:10.1007/s12094-017-1817-9.
72. Fisher RA. Note on Dr. Berkson's criticism of tests of significance. *J Am Stat Assoc*. 1943;38(221):103–104. doi:10.1080/01621459.1943.10501783.
73. Basta NO, James PW, Craft AW, McNally RJQ. Season of birth and diagnosis for childhood cancer in Northern England, 1968–2005. *Paediatr Perinat Epidemiol*. 2010;24(3):309–318. doi:10.1111/j.1365-3016.2010.01112.x.
74. Woodhouse PR, Khaw KTee, Plummer M, Meade TW, Foley A. Seasonal variations of plasma fibrinogen and factor VII activity in the elderly: winter infections and death from cardiovascular disease. *Lancet*. 1994;343(8895):435–439. doi:10.1016/S0140-6736(94)92689-1.
75. Lee BK, Gross R, Francis RW, Karlsson H, Schendel DE, Sourander A, et al. Birth seasonality and risk of autism spectrum disorder. *Eur J Epidemiol*. 2019;34(8):785–792. doi:10.1007/s10654-019-00506-5.
76. Murray G, Allen NB, Rawlings D, Trinder J. Seasonality and personality: a prospective investigation of Five Factor Model correlates of mood seasonality. *Eur J Pers*. 2002;16(6):457–468. doi:10.1002/per.462.
77. Duhem PMM. *The aim and structure of physical theory*. Princeton (NJ): Princeton University Press; 1954.
78. Efron B. Prediction, estimation, and attribution. *J Am Stat Assoc*. 2020;115(530):636–655. doi:10.1080/01621459.2020.1762613.
79. Pearl J. The limitations of opaque learning machines. In: *Possible minds: twenty-five ways of looking at AI*. City of Westminster (London, England): Penguin Press; 2019. p. 13–19.
80. Heinze G, Wallisch C, Dunkler D. Variable selection – a review and recommendations for the practicing statistician. *Biom J*. 2018;60(3):431–449. doi:10.1002/bimj.201700067.
81. Shmueli G. To explain or to predict? *Statist Sci*. 2010;25(3):289–310. doi:10.1214/10-STS330.