ELSEVIER

Journal of Clinical Epidemiology 59 (2006) 964–969

Journal of Clinical Epidemiology

Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health

Peter C. Austin[a,b,c,*], Muhammad M. Mamdani[a,d], David N. Juurlink[a,e,f], Janet E. Hux[a,c,e,f]

[a]*Institute for Clinical Evaluative Sciences, G1 06, 2075 Bayview Avenue, Toronto, Ontario, M4N 3M5 Canada*
[b]*Department of Public Health Sciences, University of Toronto, Toronto, Ontario, Canada*
[c]*Department of Health Policy, Management and Evaluation, University of Toronto, Canada*
[d]*Faculty of Pharmacy, University of Toronto, Canada*
[e]*Clinical Epidemiology and Health Care Research Program (Sunnybrook & Women's College Site), Canada*
[f]*Division of General Internal Medicine, Sunnybrook & Women's College Health Sciences Centre and the University of Toronto, Canada*

Accepted 19 January 2006

Abstract

Objectives: To illustrate how multiple hypotheses testing can produce associations with no clinical plausibility.

Study Design and Setting: We conducted a study of all 10,674,945 residents of Ontario aged between 18 and 100 years in 2000. Residents were randomly assigned to equally sized derivation and validation cohorts and classified according to their astrological sign. Using the derivation cohort, we searched through 223 of the most common diagnoses for hospitalization until we identified two for which subjects born under one astrological sign had a significantly higher probability of hospitalization compared to subjects born under the remaining signs combined ($P < 0.05$).

Results: We tested these 24 associations in the independent validation cohort. Residents born under Leo had a higher probability of gastrointestinal hemorrhage ($P = 0.0447$), while Sagittarians had a higher probability of humerus fracture ($P = 0.0123$) compared to all other signs combined. After adjusting the significance level to account for multiple comparisons, none of the identified associations remained significant in either the derivation or validation cohort.

Conclusions: Our analyses illustrate how the testing of multiple, non-prespecified hypotheses increases the likelihood of detecting implausible associations. Our findings have important implications for the analysis and interpretation of clinical studies.

Keywords: Subgroup analyses; Multiple comparisons; Hypothesis testing; Astrology; Data mining; Statistical methods

1. Introduction

The second International Study of Infarct Survival (ISIS-2) demonstrated that the use of aspirin during the acute phase of acute myocardial infarction reduced mortality in a group of more than 17,000 patients [1]. A subgroup analysis demonstrated that aspirin increased mortality of patients born under the astrological sign of Gemini or Libra. This biologically implausible finding reinforced the authors' contention that frivolous subgroup analyses should be avoided.

Although the subgroup analysis in the ISIS-2 trial was intended as an amusing illustration of a fundamental statistical construct, other investigators have examined the effect of astrologic signs more rigorously. For example, Gurm and Lauer [2] conducted a study to examine the belief that those born under the sign of Leo are "big-hearted" and at increased risk for heart disease. They examined 32,386 patients who underwent exercise stress testing at the Cleveland Clinic between 1990 and 1999 and found a slight excess of deaths among Leos (9.6% vs. 8.7%). This effect disappeared in a matched propensity score analysis ($P = 0.3$). Furthermore, they found no correlation between astrological signs and abnormality on stress testing.

While an undue reliance on astrologic phenomena as a guide to health and healthcare may put subjects at risk for adverse outcomes [3], we examined the relationship between birth sign and health outcomes with a different intent. The purpose of the current study was to demonstrate the pitfalls of multiple hypothesis testing and of conducting analyses without prespecified hypotheses. We hypothesized

* Corresponding author. Tel.: +1-416-480-6131; fax: +1-416-480-6048.
E-mail address: peter.austin@ices.on.ca (P.C. Austin).

0895-4356/06/$ – see front matter
doi: 10.1016/j.jclinepi.2006.01.012

that we could generate numerous statistically significant associations, but that these would be neither reproducible nor biologically plausible. For illustrative purposes, we studied the association between astrological signs and health.

2. Methods

We conducted a population-based retrospective cohort study using administrative databases covering 10,674,945 residents of Ontario aged 18–100 years. The Registered Person's Database (RPDB) contains basic demographic data on all residents of Ontario, Canada. We extracted information on all residents of Ontario between the ages of 18 and 100 in 2000 and who were alive on their birthday in 2000. We then randomly assigned these individuals to equally sized derivation and validation cohorts. From the birth date, we determined the astrological sign under which each person was born.

The Canadian Institute of Health Information (CIHI) hospital discharge abstract database contains data on all hospital separations in the province of Ontario. We examined all admissions to Ontario hospitals among subjects aged 18 to 100 years during a 2-year period (January 1, 2000 to December 31, 2001), who were classified as either urgent or emergent admissions (i.e., elective or planned admissions were excluded). Each admission was classified according to the most responsible diagnosis, using the first three digits of the ICD-9 coding scheme. Diagnoses were then ranked from most frequent to least frequent. Both the CIHI discharge abstract database and the RPDB database contain encrypted versions of residents' health card numbers, permitting the two databases to be deterministically linked in an anonymous fashion.

Beginning with the most frequently occurring urgent or emergent diagnosis for hospitalization, we determined whether persons in the derivation cohort were hospitalized with that diagnosis in the 365 days following their birthday in 2000. We then determined the proportion of subjects born under each astrological sign who were hospitalized with that same diagnosis in the year subsequent to their birthday in 2000. We then identified the astrological sign with the highest hospitalization rate for that diagnosis. We then determined whether the probability of admission for that diagnosis was statistically significantly different for residents born under this astrological sign than for residents born under all other astrological signs combined (i.e., we compared the probability of admission between residents born under one astrological sign and residents born under all other signs—a two-sample comparison of binomial proportions). Statistical significance was assessed using Fisher's exact test, and a two-tailed significance level of 0.05 was used to denote statistical significance. This process was repeated for all diagnoses, beginning with the most frequent, until two diagnoses were identified for each astrological sign. This phase of the study served as the hypothesis-generating phase.

In the validation cohort, we explicitly tested the 24 hypotheses associating astrological sign and illness that were generated in the derivation cohort.

3. Results

The number of Ontario residents who were aged between 18 and 100 years and who were alive on their birthday in 2000 was 10,674,945. The derivation cohort included 5,337,472 residents and the validation cohort included 5,337,473 residents. There were 895 diagnoses for which patients had emergent and urgent hospitalizations between January 1, 2000 and December 31, 2001.

In the derivation cohort, it was necessary to search sequentially through admissions for the 223 most common causes for hospitalization to identify two diagnoses for which the probability of hospitalization was statistically significantly greater for residents born under each astrological sign compared to residents born under the remaining 11 astrological signs. These 223 diagnoses accounted for 91.8% of all urgent and emergent hospitalizations in Ontario in 2000 and 2001. Of these 223 diagnoses, there were 72 (32.3%) for which residents born under one astrological sign had a significantly higher probability of hospitalization compared to residents born under the other astrological signs combined ($P < 0.05$). The number of diagnoses for which residents born under a given astrological sign had a significantly higher probability of hospitalization compared to residents born under the 11 other astrological signs combined ranged from a low of 2 (Scorpio) to a high of 10 (Taurus), with a mean of 6 diagnoses for each astrological sign. The P-values for the 72 significant associations ranged from 0.0003 to 0.0488. The two most frequently occurring diagnoses for which each astrological sign had a higher probability of hospitalization compared to the other astrological signs combined are described in Table 1. The P-values for testing the significance of the association between a particular astrological sign and the probability of the diagnosis-specific admission ranged from 0.0006 to 0.0475 among these 24 potential associations. In Table 1, we also report the relative risk comparing the probability of hospital admission for residents born under the given astrological sign with the probability of hospital admission for residents born under all other astrological signs combined. The relative risks ranged from a low of 1.10 to a high of 1.80. For example, the probability of hospitalization for lymphoid leukemia was 80% greater for Scorpios than it was for residents born under the 11 other astrological signs combined.

We tested the associations identified in Table 1 in the validation cohort. Of the 24 associations identified in the derivation cohort, only 2 remained statistically significant in the validation cohort. In the validation cohort, residents born under the sign of Leo had a significantly higher probability of hospitalization due to gastrointestinal hemorrhage

Table 1
Diagnoses for which residents with given astrological sign had a higher probability of hospitalization compared to residents born under the remaining astrological signs combined: results from derivation cohort

Astrological sign	ICD-9 code	Diagnosis	*P*-value	Relative risk
Aries	733	Other disorders of bone and cartilage	0.0402	1.27
	008	Intestinal infections due to other organisms	0.0058	1.41
Taurus	820	Fracture of neck of femur	0.0368	1.11
	562	Diverticula of intestine	0.0006	1.27
Gemini	998	Other complications of procedures, NEC	0.0330	1.15
	303	Alcohol dependence syndrome	0.0154	1.30
Cancer	560	Intestinal obstruction without mention of hernia	0.0475	1.12
	285	Other and unspecified anemias	0.0388	1.27
Leo	578	Gastrointestinal hemorrhage	0.0041	1.23
	V58	Encounter for other and unspecified procedure and aftercare	0.0397	1.17
Virgo	823	Fracture of tibia and fibula	0.0355	1.26
	643	Excessive vomiting in pregnancy	0.0344	1.40
Libra	808	Fracture of pelvis	0.0108	1.37
	430	Subarachnoid hemorrhage	0.0377	1.44
Scorpio	566	Abscess of anal and rectal region	0.0123	1.57
	204	Lymphoid leukemia	0.0395	1.80
Sagittarius	784	Symptoms involving head and neck	0.0376	1.30
	812	Fracture of humerus	0.0458	1.28
Capricorn	799	Other ill-defined and unknown causes or morbidity and mortality	0.0105	1.29
	634	Abortion	0.0242	1.28
Aquarius	413	Angina pectoris	0.0071	1.23
	481	Other bacterial pneumonia	0.0375	1.33
Pisces	428	Heart failure	0.0013	1.13
	411	Other acute and subacute forms of ischemic heart disease	0.0182	1.10

Abbreviation: NEC = not elsewhere classified.

compared to other residents of Ontario, with a relative risk of 1.15 ($P = 0.0483$). Similarly, residents born under the sign of Sagittarius had a significantly higher probability of hospitalization for fractures of the humerus compared to residents born under the remaining 11 astrological signs, with a relative risk of 1.38 ($P = 0.0125$). The remaining 22 associations were no longer significant in the validation cohort ($0.0743 \leqslant P \leqslant 0.9574$).

4. Discussion

We identified at least two diagnoses for which Ontario residents born under each astrological sign had a significantly higher probability of hospitalization compared to residents born under the remaining astrological signs combined. Two of these 24 associations remained statistically significant when tested in an independent validation cohort. These observations yield several important lessons about hypothesis testing, study design, and the interpretation of the results of clinical studies.

4.1. The pitfalls of multiple significance tests

First, it was relatively simple to generate numerous statistically significant associations when we examined a large number of potential associations. We began the study with no prespecified hypotheses. Rather, we searched sequentially through a list of diagnosis codes until at least two diagnoses had been found for each astrological sign, for which residents born under that sign were significantly more likely to be hospitalized compared to residents born under the remaining astrological signs combined. This exercise implicitly involved multiple comparisons for each diagnosis. For each astrological sign, we computed the proportion of persons born under that sign who were hospitalized for that diagnosis in the year subsequent to their birthday in 2000. We then selected the astrological sign for which persons born under that sign had the highest probability of hospitalization. This implicitly involved 66 pairwise comparisons, because there are $\binom{12}{2} = \frac{12!}{10!2!}$ ways of selecting distinct pairs from a set of 12 objects.

The finding that 22 of 24 statistically significant findings generated in the derivation cohort were not confirmed in the validation cohort illustrates the dangers inherent in studies involving multiple, non-prespecified hypotheses.

4.2. Adjusting P-values for multiple comparisons

Second, our observation that two of the associations identified in the derivation set were confirmed in the

validation set does not necessarily provide evidence that those born under the sign of Leo have a significantly higher probability of hospitalization for gastrointestinal hemorrhage, or that those born under the sign of Sagittarius have a higher probability of hospitalization for fractures of the humerus. Under the null hypothesis, *P*-values are uniformly distributed between 0 and 1. The likelihood of a type I error—identifying a statistically significant association where none exists—is 5%, when using a 0.05 significance level. When testing 24 hypotheses in which the null hypothesis is true, the likelihood that at least one will be found to be significant simply by chance is 70.8%. Thus, by not making appropriate adjustments for the testing of multiple hypotheses, we greatly increased our risk of falsely ''uncovering'' an association between astrological sign and illness. Had we instead endeavored to preserve an overall type I error rate of 0.05, we would have had to use a significance level of 0.00213 for each of the 24 individual hypothesis tests (this is marginally less conservative than a Bonferroni correction, which would have used a significance level of 0.05/24 = 0.00208; both methods require that the multiple comparisons be independent of one another). Using this significance level, none of the 24 hypothesized associations would have been significant in the validation cohort. Sankoh et al. [4] discuss the relative merits of different methods in adjusting for the testing of multiple endpoints in clinical trials. In particular, they note that the Bonferroni adjustment (which is an approximation to our exact method) ignores most of the information from the data and is too conservative when there are many outcomes [4]. Bender and Lange [5] provide an overview of methods to adjust for multiple testing in medical and epidemiological studies.

Similarly, in the derivation cohort, there were implicitly 14,718 comparisons (223 diagnoses × 66 pairwise comparisons per diagnosis). To retain an overall 5% type I error rate, one would need to use a significance level of 0.000003485 for an individual hypothesis test. Using this significance level, none of the 72 associations identified in the derivation cohort would have been identified as statistically significant. We should note that there were 72 diagnoses for which the astrological sign with the *highest* probability of hospitalization had a significantly higher probability of hospitalization compared to that for the remaining astrological signs combined. It is highly likely that there were other astrological signs (but not the one with the highest probability of hospitalization) that had a significantly higher probability of hospitalization compared to residents born under the remaining 11 astrological signs combined. While these comparisons were implicitly considered in our design, they were not reported on in the current study. Our study illustrates that in a trial with multiple hypothesis tests (either secondary outcomes or subgroup analyses), the significance level used should be adjusted to preserve an overall type I error of a desired level. It is common in randomized clinical trials to examine one primary outcome or endpoint and multiple secondary endpoints. However, as the number of secondary endpoints or subgroup analyses increases, the risk of erroneously identifying a significant association also increases. To quantify the prevalence of subgroup analyses and the number of endpoints in clinical trials, we examined all 131 randomized clinical trials published in the *Journal of the American Medical Association*, the *New England Journal of Medicine*, the *Lancet*, and the *British Medical Journal* between January 1 and June 30 of 2004. The mean and median number of subgroups in which endpoints were compared between treatment arms were 5.1 and 2, respectively (IQR = 0–6), while the mean and median number of significance tests of efficacy and safety endpoints were 26.5 and 19, respectively (IQR = 9–32). The maximum number of distinct subgroups in which endpoints were compared between treatment arms was 68, while the maximum number of observed endpoints was 185.

4.3. The importance of biologic plausibility

Third, none of the hypotheses generated using the derivation cohort had any apparent biologic plausibility. Despite confirming 2 of the 24 prespecified hypotheses in the validation cohort, there is no currently apparent mechanism by which Leos might be predisposed to gastrointestinal hemorrhage or Sagittarians to humeral fractures. In interpreting the subgroup analyses from the ISIS-2 trial, the authors argued that the results were not biologically plausible, and should be ignored. Caution is required in interpreting results that do not have apparent biological plausibility. In particular, it is important that biologically plausible associations be specified during the design of the study, because it is tempting to construct biologically plausible reasons for observed subgroup effects after having observed them [6]. Our study demonstrates that data-driven statistical methods may result in conclusions that are neither reproducible nor biologically plausible.

4.4. Subgroup analyses in clinical trials

Subgroup analyses are common in randomized controlled trials. Indeed, the subgroup analysis reported by the ISIS-2 investigators [1] motivated the current study. Many investigators have cautioned against subgroup analyses in randomized controlled trials. It has been argued that such analyses should be prespecified, and that there should be a pre-specified biologically plausible explanation for the proposed subgroup analysis [6]. Furthermore, it has been suggested that one should not be guided by statistical significance, but rather by trends and consistency, because such analyses are frequently underpowered [6]. Similarly, Sleight [7] cautions against subgroup analyses in randomized clinical trials, suggesting that plausible explanations for specific findings can often be found for conclusions that were, in reality, spurious. If our categorization of residents

had been based upon clinical criteria or demographic characteristics rather than astrological sign, it is likely that post hoc plausible explanations could have been constructed for many of the associations identified. Both Yusuf et al. [8] and Oxman and Guyatt [9] provide guidelines for interpreting the results of subgroup analyses. Freemantle [10] suggested that a purist approach would be to examine subgroup analyses and secondary endpoints only if the primary endpoint is statistically significant. Recently, Rothwell [11] discussed arguments for and against subgroup analyses and provided guidelines for designing and interpreting subgroup analyses. There are increasing calls for the registration of trial protocols prior to the start of randomized clinical trials [12], an initiative that could reduce the number of frivolous subgroup analyses. The current study adds a cautionary note concerning the practice of conducting numerous significance tests, such as those often performed in the setting of a randomized trial.

4.5. Validation studies

The current study used both derivation and validation datasets. Only 2 of the 24 significant associations that were identified using the derivation cohort remained statistically significant in the validation cohort. The use of derivation and validation datasets has been frequently advocated in the statistical literature [13]. The use of a validation dataset allows one to assess the reproducibility of findings obtained in the derivation cohort, and serves to protect oneself from identifying spurious findings in a single dataset. We suggest that when surprising associations are obtained, either as a result of subgroup analyses or analysis of secondary outcomes in clinical trials, researchers seek to reproduce these findings in separate studies.

This concept is nicely illustrated by two major clinical trials. The Prospective Randomized Amlodipine Survival Evaluation (PRAISE) study examined the effect of amlodipine in patients with congestive heart failure and found no benefit in the primary analysis. In a prespecified subgroup analysis, amlodipine reduced the risk of fatal and nonfatal events in patients with severe nonischemic heart failure ($P = 0.04$) [14]. Furthermore, amlodipine seemed to prevent a secondary outcome (mortality) in the same patients ($P < 0.001$). The PRAISE-2 trial, which was explicitly designed to examine the effect of amlodipine in nonischemic heart failure patients, found no effect on mortality or cardiac events [15]. This trial was never reported in detail. Similarly, the Evaluation of Losartan in the Elderly (ELITE) trial suggested a survival benefit in elderly heart failure patients treated with the angiotensin II antagonist losartan compared to the ACE inhibitor captopril [16]. This finding was not replicated in the ELITE II trial [17]. The results of the PRAISE/PRAISE-2 and ELITE/ELITE II trials illustrate that subgroup analyses, even when specified, can result in findings that are not subsequently reproducible.

4.6. Data mining

Finally, there is an increasing interest in "data mining" as a means of hypothesis generation, particularly in commercial endeavors. Data mining has been variously described as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data" [18] and as a "semi-automatic extraction of patterns, changes, associations, anomalies, and other statistically significant structures from large data sets" [19]. Data mining is often conducted in large datasets and often does not involve prespecified hypotheses. In the current study, we began with no prespecified hypotheses, and used automated methods to detect apparently significant associations. Despite the addition of a validation cohort, two unanticipated associations remained significant. Our study therefore serves as a cautionary note regarding the interpretation of findings generated by data mining, and suggests that conclusions obtained from data mining should be viewed with a healthy degree of skepticism.

In conclusion, we were able to identify multiple significant associations, all of them clinically implausible, between astrological sign and the probability of hospitalization for specific diagnoses. Two of these associations remained statistically significant when tested in an independent validation cohort. Our study emphasizes the hazards of testing multiple, non-prespecified hypotheses.

Acknowledgments

The Institute for Clinical Evaluative Sciences (ICES) is supported in part by a grant from the Ontario Ministry of Health and Long-Term Care. The opinions, results, and conclusions are those of the authors and no endorsement by the Ministry of Health and Long-Term Care or the Institute for Clinical Evaluative Sciences is intended or should be inferred. Drs. Austin, Mamdani, and Juurlink are supported by New Investigator awards from the Canadian Institutes of Health Research (CIHR).

References

[1] ISIS-2 Collaborative Group. Randomized trial of intravenous streptokinase, oral aspirin, both, or neither among 17187 cases of suspected acute myocardial infarction: ISIS-2. Lancet 1988;2(8607):349–60.
[2] Gurm HS, Lauer MS. Predicting incidence of some critical events by sun signs—the PISCES Study. ACC Curr J Rev 2003;Jan/Feb:22–4.
[3] Philips DP, Ruth TE, Wagner LM. Psychology and survival. Lancet 1993;342(8880):1142–5.
[4] Sankoh AJ, Huque MF, Dubey SD. Some comments on frequently used multiple endpoint adjustment methods in clinical trials. Stat Med 1997;16:2529–42.
[5] Bender R, Lange S. Adjusting for multiple testing—when and how? J Clin Epidemiol 2001;54:343–9.
[6] Topol EJ, Califf RM, Van de Werf F, Simoons M, Hampton J, Lee KL, et al. Perspectives on large-scale cardiovascular clinical trials for the new millennium. Circulation 1997;95:1072–82.

[7] Sleight P. Debate: subgroup analyses in clinical trials – fun to look at, but don't believe them? Curr Control Trials Cardiovasc Med 2000;1:25–7.
[8] Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. J Am Med Assoc 1991;266:93–8.
[9] Oxman AD, Guyatt GH. A consumer's guide to subgroup analysis. Ann Intern Med 1992;116:78–84.
[10] Freemantle N. Interpreting the results of secondary end points and subgroup analyses in clinical trials: should we lock the crazy aunt in the attic? BMJ 2001;322:989–91.
[11] Rothwell PM. Subgroup analysis in randomized controlled trials: importance, indications, and interpretation. Lancet 2005;365:176–86.
[12] DeAngelis CD, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, et al. International Committee of Medical Journal Editors. Clinical trial registration: a statement from the International Committee of Medical Journal editors. J Am Med Assoc 2004;292:1363–4.
[13] Picard RR, Berk KN. Data splitting. Am Stat 1990;44:140–7.
[14] Packer M, O'Connor CM, Ghali JK, Pressler ML, Carson PE, Belkin RN, et al. Effect of amlodipine on morbidity and mortality in severe chronic heart failure. N Engl J Med 1996;335:1107–14.
[15] Thackray S, Witte K, Clark AL, Cleland JGF. Clinical trials update: OPTIME-CHF, PRAISE-2, ALL-HAT. Eur J Heart Fail 2000;2:209–12.
[16] Pitt B, Segal R, Martinez FA, Meurers G, Cowley AJ, Thomas I, et al. Randomized trial of losartan versus captopril in patients with heart failure (Evaluation of Losartan in the Elderly Study, ELITE). Lancet 1997;349:747–52.
[17] Pitt B, Poole-Wilson PA, Segal R, Martinez FA, Dickstein K, Camm AJ, et al. Effect of losartan compared with captopril on mortality in patients with symptomatic heart failure: randomized trial—the Losartan Heart Failure Survival Study ELITE II. Lancet 2000;355:1582–7.
[18] Everitt BS. The Cambridge dictionary of statistics, 2nd edition. Cambridge: Cambridge University Press; 1998.
[19] Accessed December 14. Available at http://www.rgrossman.com/dm.htm 2005.