Refuted Causal Claims from Observational Studies
https://bit.ly/experimentGuideRefutedObservationalStudies

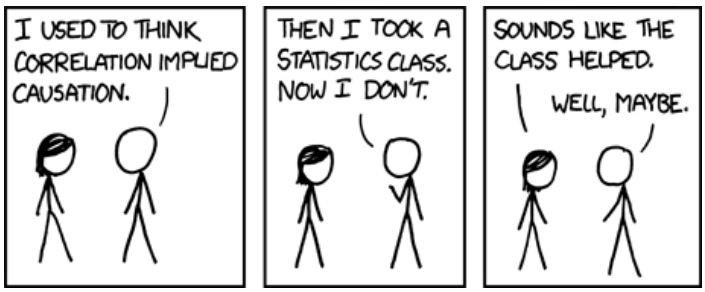
REFUTED CAUSAL CLAIMS FROM OBSERVATIONAL STUDIES

RON KOHAVI

UPDATED 21 DEC 2019

Commented [RK1]: I'd love to hear about more examples. Please comment here or by e-mail and I'll acknowledge contributions.

Commented [RK2]: Use allowed in https://xkcd.com/license.html
You can use them freely (with some kind of link) in not-for-profit publications, and I'm also okay with people reprinting occasional comics (with clear attribution) in publications like books, blogs, newsletters, and presentations.



Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there.'
-- Randall Munroe in https://xkcd.com/552/

Any claim coming from an observational study is most likely to be wrong
-- S. Stanley Young and Alan Karr (2011)

"activity bias" [is] a tendency to overestimate the causal effects of advertising using online behavioral data
-- Randall Lewis, Justin Rao, David Reiley (2011)

Why you care: It is common to find a correlation between A and B and incorrectly conclude that A causes B or vice-versa. While correlations are a great source of hypotheses, causal claims based on observations alone should have a lower level of trust than properly run randomized controlled experiments; this is true even when attempts are made to control for several factors. We review famous examples where causality was claimed as likely in observational studies, but later refuted in studies higher in the hierarchy of evidence, such as randomized controlled experiments.

Refuted Causal Claims from Observational Studies
https://bit.ly/experimentGuideRefutedObservationalStudies

Introduction

While many observational studies are later confirmed by randomized controlled experiments (Concato, Shah, & Horwitz, 2000), others are refuted. Ioannidis (2005) evaluated claims coming from highly cited studies; of six observational studies included in his study, five failed to replicate. Stanley Young and Alan Karr (2011) compared published results from medical hypotheses shown to be significant using observational studies (i.e., uncontrolled) with randomized clinical trials, considered more reliable. Of 52 claims in 12 papers, none replicated in the randomized controlled trials, and in five of the 52 cases, the direction was statistically significant in the opposite direction of the observational study. Their (albeit overstated in our opinion) conclusion: "Any claim coming from an observational study is most likely to be wrong."

In the famous observational studies below, causality was claimed as likely, yet later refuted. The examples should help the reader see the risks and pitfalls, and the reason observational studies are ranked low in the Hierarchy of Evidence (Greenhalgh, 2014). Additional examples are available in the Design and Interpretation of Clinical Trials Lecture 8B: High-Profile Cases (Holbrook & Drye). One key lesson is that if you *can* run a randomized controlled experiment, then strongly prefer it to an observational study. There are, of course, cases where only observational studies can be run, but claims of causality should be trusted less, and one should attempt to control for confounders and the pitfalls shown below.

Bloodletting

In Bad Medicine, Wootton (2007) claimed that "for 2,400 years patients have believed that doctors were doing them good; for 2,300 years they were wrong." He notes that for "two thousand years, from the first century BC until the mid-nineteenth century, the main therapy used by doctors was bloodletting—opening a vein in the arm with a special knife called a lancet—which weakened and even killed patients."

Figure 0.1: Lancet

Doctors and researchers were fooled by correlation: bloodletting had a calming effect, and thus doctors believed it was helpful.

So strong is this belief that doctors thought that many diseases, including hepatitis, pneumonitis, and ophtalmia were manifestations of the inflammation of organs, and bloodletting was deemed

an efficient treatment. After years of using lancets, leeches were deemed a better way to suck the blood. In 1833 alone, France imported 42 million leeches for medical use (Morabia, 2006).

In 1799, President George Washington died after three different doctors each performed bloodletting and extracted more than half his blood volume in a short period when he was sick. It is now believed that this procedure led to "preterminal anemia, hypovolemia, and hypotension" (Vadakan, 2004) and premature death of the first US president.

In 1828, Pierre-Charles-Alexandre Louis published an article and then a book on the effects of bloodletting (Louis, 1836; Morabia, 2006), which is one of the earliest (non-randomized) controlled experiments. Louis took 77 patients from a very homogeneous group with the same, well-characterized form of pneumonia. He analyzed the duration of the disease and the frequency of death by the timing of the first bloodletting (early in days 1-4, or later in days 5-9). The result: 44% of the patients who had been bled early died compared to 25% of those bled late (Morabia, 2006).

Dramatic Overestimates of Effects of Advertising

Lewis, Rao, and Reiley (2011) compared the effectiveness of online advertising as estimated by observational studies and controlled experiments, which they called the "gold standard." They ran three experiments:

1. Advertisements (display-ads) were shown to users, and the question was: what is the increase (lift) in the number of users who search using keywords related to the brand shown in the ad. Using several observational studies of 50 million users, including three regression analyses with control variables, the estimated lift ranged from 871% to 1198%. These widely diverging from the more trustworthy controlled experiment lift of 5.4%.
   The reason is that users who actively visit Yahoo! on a given day are much more likely both to see the display ad and to do a Yahoo! search. The ad exposure and the search behavior are highly positively correlated, but the display ads have very little causal impact on the searches.
   One might conjecture that the exaggerated results depend on increased activity on a single site (Yahoo!), but the next experiments show this not to be the case.
2. Videos were shown to users, and the question was whether these would lead to increased activity. Users were recruited through Amazon Mechanical Turk and exposed half to a 30-second video advertisement promoting Yahoo.com services (the Treatment), and half to a political video advertisement (the Control). An observational study of the treatment group before and after the exposure overstated the effects of the ad by 350%. Being active on Amazon Mechanical Turk on a given day increased the chance of participating in the experiment and in being active on Yahoo!

3. An ad campaign was shown to users on Yahoo! The question was: what is the impact to rival firms. It turns out that exposed users were much more likely to sign up at the competitor's website on the day they saw the ad, compared to the week prior to the exposure. Is this spillover? No, the experiment showed that the control group exhibited a nearly identical lift on the day they came to the Yahoo! site but did not see the campaign ads. The causal effect was very close to zero.

Uncontrolled Factor: Genetics.

Does Night Light Causes Myopia?

An article in Nature (Quinn, Shin, Maguire, & Stone, 1999) used an observational study to establish a correlation between young children ages 0-2, who slept at night with room lighting, and myopia later in life. They showed the following graph and claimed that the prevalence of myopia increased "markedly with increased levels of night-time ambient lighting during sleep before the age of 2 years." Their analysis showed that the percentage increase in myopia and high myopia was statistically significant with p-value < 0.00001.
The authors recognized the fact that this was an observational study and wrote: "Although it does not establish a causal link, the statistical strength of the association of night-time light exposure and childhood myopia does suggest that the absence of a daily period of darkness during early childhood is a potential precipitating factor in the development of myopia." That said, they then claimed that causality is likely: "Despite these qualifications, it seems prudent that infants and young children sleep at night without artificial lighting in the bedroom."

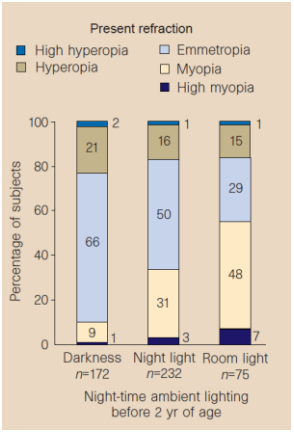


Figure 0.2: Exposure to light correlated with increase of Myopia: from 10% for those who slept the dark to 34% for those with night light to 55% for those with room light

The story generated significant interest, including CNN article and video (Etheridge, 1999),which wrote "Young children who sleep with a light on may have a substantially higher risk of developing nearsightedness as they get older, says a new study in the journal Nature."

The article then claims that "…the study offers a novel explanation for the increasing prevalence of myopia over the past two centuries, as populations shifted from agricultural to urban environments. The current findings suggest that the greater ambient nighttime light levels associated with industrialization may be a factor in the high incidence of myopia in developed nations."

A year later, two papers were published in Nature that were unable to replicate the result, and which highlighted a confounding factor: parental myopia. The first paper (Zadnik, et al., 2000) show that the proportion of myopic children in those subjected to a range of nursery-lighting conditions is remarkably uniform. Furthermore, they made an important observation: there was a strong statistically significant association between myopic parents and nursery lightings (p-value<0.001), which was not controlled for in the original study. Their conclusion is that "myopia is unlikely to develop in children as a result of exposure to night-time lighting as infants." The second paper (Gwiazda, Ong, & Thorn, 2000) was also unable to replicate the results, and found that myopic parents are more likely to employ night-time aids for their children and that there is an association between myopia in parents and their children. They likewise wrote that "we question whether parents need to be concerned about causing myopia in their children by lighting their nurseries at night."

Three years later, a third paper (Guggenheim, Hill, & Yam, 2003) showed that myopia occurred with approximately equal frequency in those who slept with and without light exposure at night. The study's conclusion is that "night-time light exposure during infancy is not a major risk factor for myopia development in most population groups."

Confounders

Does Vitamin C Reduce Coronary Heart Disease?

Observational studies showed that vitamin C reduces Coronary Heart Disease (CHD). Randomized controlled trials (RCTs) showed the complete opposite: vitamin C increases CHD. The controlled experiments are considered more trustworthy, and a study was done to assess the reasons for the disparity (Lawlor, Smith, Bruckdorfer, Kundu, & Ebrahim, 2004). The study shows that ten socioeconomic position indicators are linearly associated with vitamin C concentrations, including number of bathrooms in house, shared bedroom, car access, etc. Behavior factors also differed, including current smoker, exercise, low fat diet, BMI > 30 (obesity), and alcohol consumption. The conflicting results are therefore likely the result of inadequate adjustment for the complexity of social and environmental exposures. The key problem is that it is impossible to know if enough confounders have been accounted for.

Causal Insufficiency

Twin Studies

In 2007, an observational study concluded that adolescents who lose their virginity earlier than their peers are more likely to become juvenile delinquents (Armour & Haynie, 2007). The study used the National Longitudinal Study of Adolescent Health with over 7,000 adolescents, but it was an observational study. The study included a number of covariates as statistical controls, including gender, race, receipt of public assistance, parental education, family structure, previous substance use and depression, importance of religion, school GPA, relative pubertal status, and virginity pledge status, yet there is always the possibility of another confounding covariate; causal sufficiency is impossible to prove.

The causal claims in the study, such as "experiencing early or late sexual debut continues to have consequences for delinquent behavior occurring in young adulthood'' are not justified. The fact that federal "abstinence" programs were already part of the curriculum likely made it easier for people to accept the causal claim (Weiss, 2007).

The study did not control for genetic factors, yet several twin studies have demonstrated that siblings who are more genetically similar exhibit more similar ages at first sex. A later paper (Harden, Mendle, Hill, Turkheimer, & Emery, 2008) compared 534 monozygotic twins from the same database as the original study. Twin studies are one of the best Natural Experiments, with research showing that results can usually be generalized (representativeness assumption) (McGue, 2014). Differences in delinquency between the twins cannot be attributed to genetic confounds nor to familial environment, such as sociodemography, family structure, and family relationships. Comparing monozygotic twins, therefore provides a more rigorous test (although still not as powerful as a fully controlled experiment). The conclusion from the later paper was the opposite of the initial study: earlier age at first sex predicted lower levels of delinquency in early adulthood!

Time-Sensitive Confounder (Death)

In several very large observational medical studies, hormone replacement therapy (HRT) was suggested as a way to reduce Coronary Heart Disease (CHD) for postmenopausal women (Grodstein, et al., 1996; Grodstein, et al., 2000). Over 50,000 women from the Nurses' Healthy Study were followed-up for up to 16 years and the results indicated a "marked decrease in the risk of major coronary heart disease." observational studies were so convincing that many doctors prescribed Premarin for HRT over years. In 2001, it was the third most-prescribed drug in the United States (Patterson, 2002).

A later study based on randomized clinical trials (controlled experiments), part of the Women's Health Initiative (WHI) concluded that HRT "does not confer cardiac protection and may increase the risk of CHD among generally healthy postmenopausal women" (Manson, et al., 2003). While the planned duration of the study was 8.5 years, it was terminated after 5.2 years

because “the overall risks exceeded the benefits” (although that may have been a rushed decision in hindsight (Robinson, 2018)).

Why were the results inconsistent? One interesting reason is that the control and treatment populations in the observational studies were not comparable (a common scenario is observational studies) but the difference is unusual: some women who used HRT were excluded from the observational study because they…died (Holbrook & Drye; Hernan, et al., 2008; Manson, et al., 2003). Most of the women in the observational study’s treatment group were already taking HRT for two to more than five years. The WHI (randomized controlled) study noted that there was increased risk in the first two years and a pattern of decreased risk over time. As stated in the Design and Interpretation of Clinical Trials Lecture 8B: High-Profile Cases (Holbrook & Drye) “The women were different … in terms of the extent of their use of a hormone replacement therapy. So some women in the observational study actually had this immortal time where they survived long enough to get into the observational study, and the women who died early on were less likely to get into the study… The women who died didn't enroll in the study.”

External Validity Problems

In some attempts to reproduce results, there are questions about whether the population being experimented is similar. Lithgow, Driscoll, and Phillips (2017) share such an amazing effort to explain reproducibility problems in labs involving 100,000 worms and four years of effort, yet they write “Despite more experiments and additional publications, we couldn't work out why the labs were getting different lifespan results. To this day, we still don't know.” In contrast, in the three examples shared above, it was exactly the same population, so the overestimates highlight the risks of assuming observational data can be used to estimate treatment effects (causality).

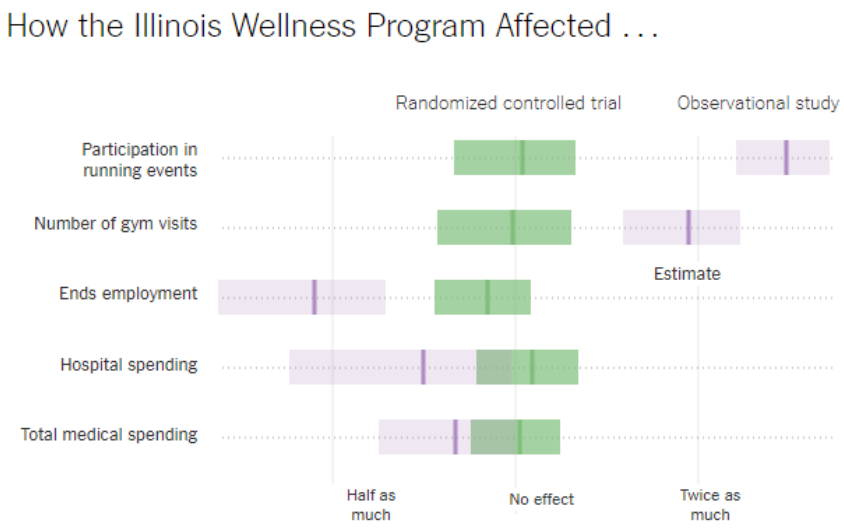
Refuted Causal Claims from Observational Studies
https://bit.ly/experimentGuideRefutedObservationalStudies

Other Examples

Here are some more examples

1. Clofibrate was used to treat elevated serum cholesterol levels, and growing in popularity, before a randomized clinical trial showed that it increases mortality, resulting in the drug being banned in several countries (Department of Clinical Epidemioology and Biostatistics, 1981).

2. Workplace Wellness programs cover over 50 million workers and are intended to reduce medical spending, improve productivity, and improve well-being. Of claims made by 115 prior observational studies, 83% were ruled out based on a large randomized controlled experiment with over 12,000 employees (Carroll, 2018; Jones, Molitor, & Reif, 2018). Here are several factors with their 95% confidence interval shown in green for the randomized controlled experiment, and in purple if the same data is analyzed as an observational study (the x-axis is the multiplicative treatment effect with the center "no effect" indicating 1.0).



Source: What Do Workplace Wellness Programs Do? Evidence from the Illinois Workplace Wellness Study

For example, participation in running events, such as a five- or ten-kilometer run, showed a large increase from 3.3% to 9.2% in the observational analysis, but only 6.0 to 6.5% in the randomized controlled experiment.

Acknowledgement

Thanks to Tommy Guy for feedback on an earlier presentation of this material. Thanks to Stuart Buck for sharing the Workplace Wellness article.

REFERENCES

Armour, S., & Haynie, D. L. (2007, February). Adolescent Sexual Debut and Later Delinquency. *Journal of Youth and Adolescence, 36*(2), 141-152. doi:https://doi.org/10.1007/s10964-006-9128-4

Carroll, A. E. (2018, Aug 6). *Workplace Wellness Programs Don't Work Well. Why Some Studies Show Otherwise.* Retrieved from The New York Times: https://www.nytimes.com/2018/08/06/upshot/employer-wellness-programs-randomized-trials.html

Concato, J., Shah, N., & Horwitz, R. I. (2000). Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs. *The New England Journal of Medicine, 342*(25), 1887-1892. doi:https://www.nejm.org/doi/10.1056/NEJM200006223422507

Department of Clinical Epidemioology and Biostatistics. (1981, May 1). How to read clinical journals to distinguish useful from useless or even harmful therapy. *CMA Journal, 124.* Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1705333/pdf/canmedaj01481-0042.pdf

Etheridge, P. (1999, May 13). *Night-light may lead to nearsightedness.* Retrieved from CNN.com: www.cnn.com/HEALTH/9905/12/children.lights/index.html

Greenhalgh, T. (2014). *How to Read a Paper: The Basics of Evidence-Based Medicine.* BMJ Books. Retrieved from https://www.amazon.com/gp/product/B00IPG7GLC

Grodstein, F., Manson, J. E., Colditz, G. A., Willett, W. C., Speizer, F. E., & Stampfer, M. J. (2000). A prospective, observational study of postmenopausal hormone therapy and primary prevention of cardiovascular disease. *Annals of Internal Medicine, 133*(12), 933-941. doi:http://dx.doi.org/10.7326/0003-4819-133-12-200012190-00008

Grodstein, F., Stampfer, M. J., Manson, J. E., Colditz, G. A., Willett, W. C., Rosner, B., . . . Hennekens, C. H. (1996, August 15). Postmenopausal Estrogen and Progestin Use and the Risk of Cardiovascular Disease. *The New England Journal of Medicine, 335*, 453-461. doi:https://www.nejm.org/doi/full/10.1056/NEJM199608153350701

Guggenheim, J. A., Hill, C., & Yam, T.-F. (2003, May). Myopia, genetics, and ambient lighting at night in a UK sample. *British Journal of Ophthalmology, 87*(5), 580-582. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1771677/

Gwiazda, J., Ong, E., & Thorn, H. F. (2000, March 9). Myopia and ambient night-time lighting. *Nature, 404*, 144.

Harden, K. P., Mendle, J., Hill, J. E., Turkheimer, E., & Emery, R. E. (2008, April). Rethinking Timing of First Sex and Delinquency. *Journal of Youth and Adolescence, 37*(4), 373-385. doi:https://doi.org/10.1007/s10964-007-9228-9

Hernan, M. A., Alonso, A., Logan, R., Grodstein, F., Michels, K. B., Stampfer, M. J., . . . Robins, J. A. (2008, November). Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart

disease. *Epidemiology, 19*(6), 766-779. doi:https://dx.doi.org/10.1097%2FEDE.0b013e3181875e61

Holbrook, J., & Drye, L. T. (n.d.). Design and Interpretation of Clinical Trials, Lecture 8B: High-Profile Cases. John Hopkins University. Retrieved from https://www.coursera.org/lecture/clinical-trials/lecture-8b-high-profile-cases-sT0iu

Ioannidis, J. P. (2005). Contradicted and Initially Stronger Effects in Highly Cited Clinical Research. *294*(2).

Jones, D., Molitor, D., & Reif, J. (2018, June). *What Do Workplace Wellness Programs Do? Evidence from the Illinois Workplace Wellness Study.* Retrieved from The National Bureau of Economic Research - Working Papers: https://www.nber.org/papers/w24229

Lawlor, D. A., Smith, G. D., Bruckdorfer, K. R., Kundu, D., & Ebrahim, S. (2004, May 22). Those confounded vitamins: what can we learn from the differences between observational versus randomised trial evidence? *Lancet, 363*, 1724-1727. Retrieved from www.faculty.umb.edu/pjt/epi/lawlor04.pdf

Lewis, R. A., Rao, J. M., & Reiley, D. (2011). Here, There, and Everywhere: Correlated Online Behaviors Can Lead to Overestimates of the Effects of Advertising. *Proceedings of the 20th ACM International World Wide Web Conference (WWW20)*, (pp. 157-166). Retrieved from https://ssrn.com/abstract=2080235

Lithgow, G. J., Driscoll, M., & Phillips, P. (2017, August 24). A long journey to reproducible results. *Nature, 48*, 387-388. Retrieved from https://www.nature.com/news/a-long-journey-to-reproducible-results-1.22478

Louis, P.-C.-A. (1836). *Researches On The Effects Of Bloodletting In Some Inflammatory Diseases.* Boston: Hilliard, Gray, & Company. Retrieved from https://www.amazon.com/gp/product/133106824X

Louis, P.-C.-A. (1836). *Researches on the Effects of Bloodletting in Some Inflammatory Diseases, and on the Influence of Tartarized Antimony and Vesication in Pneumonitis* (Forgotten Books reprint 2018 ed.). (C. G. Putnam, Trans.) Boston, MA: Hilliard, Gray, and Company.

Manson, J. E., Hsia, J., Johnson, K. C., Rossouw, J. E., Assaf, A. R., Lasser, N. L., . . . Cushman, M. (2003). Estrogen plus Progestin and the Risk of Coronary Heart Disease. *The New England Journal of Medicine, 349*, 523-534. doi:https://www.nejm.org/doi/10.1056/NEJMoa030808

McGue, M. (2014). Introduction to Human Behavioral Genetics, Unit 2: Twins: A Natural Experiment . Retrieved from https://www.coursera.org/learn/behavioralgenetics/lecture/u8Zgt/2a-twins-a-natural-experiment

Morabia, A. (2006, Mar). Pierre-Charles-Alexandre Louis and the evaluation of bloodletting. *Journal of the Royal Society of Medicine, 99*(3), 158-160. doi:10.1258/jrsm.99.3.158

Morabia, A. (2006, March). Pierre-Charles-Alexandre Louis and the evaluation of bloodletting. *Journal of the Royal Society of Medicine, 99*, 158-160.

Patterson, K. (2002, Sept 17). *What Doctors Don't Know (Almost Everything).* Retrieved from The New York Times Magazine: https://www.nytimes.com/2002/05/05/magazine/what-doctors-don-t-know-almost-everything.html?pagewanted=print

Quinn, G. E., Shin, C. H., Maguire, M. G., & Stone, R. A. (1999, May 13). Myopia and Ambient Lighting at Night. *Nature, 399*, 113-114.

Robinson, G. K. (2018). What Properties Might Statistical Inferences Reasonably be Expected to Have?—Crisis and Resolution in Statistical Inference. *The American Statistician*, 1-10. doi:https://doi.org/10.1080/00031305.2017.1415971

Weiss, R. (2007, November 11). Study Debunks Theory on Teen Sex, Deliquency. *Washington Post*, A03. Retrieved from www.washingtonpost.com/wp-dyn/content/story/2007/11/11/ST2007111100542.html

Wikipedia: Pierre Charles Alexandre Louis. (2018). *Pierre Charles Alexandre Louis.* Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Pierre_Charles_Alexandre_Louis

Young , S. S., & Karr, A. (2011). Deming, data and observational studies: A process out of control and needing fixing. *Significance, 8*(3). doi:https://doi.org/10.1111/j.1740-9713.2011.00506.x

Zadnik, K., Jones, L. A., Irvin, B. C., Kleinstein, R. N., Manny, R. E., Shin, J. A., & Mutti, D. O. (2000, March 9). Myopia and ambient night-time lighting. *Nature, 404*, 143-144.