Research Methodology and Biostatistics Series VIII – Statistical Relationship Versus Cause–Effect Relationship Between Medical Factors

Abhaya Indrayan[1*]

[1]Department of Clinical Research, Max Healthcare, New Delhi

Correspondence:

Abhaya Indrayan

E-mail: abhaya.indrayan@maxhealthcare.com

DOI: https://doi.org/10.62830/mmj2-04-32c

Abstract:

A statistical relationship refers merely to a change in one factor accompanied by the change in one or more of the others. A relationship between two or more factors can be considered cause-effect when it is statistically significant to begin with and satisfies most of the following criteria: (i) biological plausibility, (ii) experimental evidence, (iii) dose-response relationship, (iv) consistency across time and space, (v) specificity, and, (vi) statistical significance, as mentioned earlier. These criteria are hard to achieve; hence, establishing a cause-effect relationship is challenging. This article provides a brief overview of the distinction between a cause-effect relationship and a statistical relationship.

Key words: Statistical Relation, Cause-Effect, Regression, Spurious Correlation.

Introduction

A statistical relationship refers to the change in one factor accompanied by an increase or decrease in one or more of the others. Although modern medical science divides the human body into tissues, organs, and systems, these systems function in tandem with one another with intricate dependence. The body also interacts with the environmental factors such as microbes, dietary intake, and physical activity. Psychological factors such as tension, stress, family support, and appreciation or criticism, also make a substantial contribution to medical parameters. With such a large number of factors at work, it is natural to investigate how and to what extent these factors affect health parameters, and how various health parameters interact with one another. Statistically, these relationships are studied through their structure and strength, and more exactly by ascertaining the cause-effect mechanisms involved.

Statistical Relationships

A statistical relationship between two or more factors is studied in two dimensions: the structure of the relationship and the strength of the relationship.

The statistical study of relationships requires identifying one or more variables as dependent and others as independents. The dependents may represent the outcome, response, or consequence, whereas the independents may be the risk factors, antecedents, or predictors that are anticipated to determine the outcome. In a statistical relationship, the independents are known as regressors and the structure of the relationship is described by regression. The dependent may be qualitative, such as survival or death; however, for ease of understanding, we will consider only quantitative dependents for the moment. For example, Skovlund[1] discussed the following regression between C-reactive protein (CRP) concentration and quality of life scores (QoL) in patients with metastatic bowel cancer:

$$QoL = 75.6 - 0.177 \times CRP$$

This is an example of a simple regression with one dependent variable (QoL) and one independent variable (CRP). The QoL on the left side in a regression is not the value of an individual but the average QoL among subjects with a specific CRP level — a fact often overlooked in research. This example is linear, as it can be plotted as a straight line, but relationships can also be curvilinear or even multidimensional surfaces when several independents are involved. A regression can include not only multiple independents but, in rare cases, multiple dependent variables as well.

Regression is obtained as the 'best fit' to the data, although this best fit may still be poor. The adequacy of fit for quantitative variables is measured by the Pearson correlation coefficient. This ranges form –1 to +1, with 0 indicating no correlation. When one dependent variable is related to several independents, the multiple correlation coefficient is used. A correlation exceeding 0.8 in either direction is considered strong. However, a big limitation of the Pearson correlation is that it measures only a linear relationship. Two variables may have a strong non-linear (curved) relationship, yet still produce a low correlation coefficient. For example, lung function shows a curvilinear relationship with age — it increases from childhood to adulthood and decreases in old age. It is not linear. To assess non-linear relationships, the coefficient of determination is used instead of the multiple correlation coefficient. Nevertheless, even these measures may be spurious, as discussed later in this communication.

When a qualitative variable, such as disease severity (none, mild, moderate, serious, critical), is investigated for its dependence on various laboratory or radiological parameters, the regression is termed logistic regression. The strength of the relationship in this case is measured mostly by Nagelkerke R^2, which is similar to the square of the multiple correlation coefficient.

When the variable of interest is a duration, such as survival time in cancer patients or hospitalisation duration in liver transplant recipients, a Cox regression is used. This approach accounts for the generally highly skewed distribution of durations and the presence of censored observations due to limited follow-up.

Cause-Effect Relationships

The relationships presented in the previous section are statistical in nature and they can sometimes be spurious. For example, the incidence of cardiovascular disease (CVD) in India has been increasing, while the birth rate has been declining over the past 50 years. These two variables show a strong negative relationship, with a correlation coefficient of nearly –0.7. The relationship exists statistically, but does it have any medical significance? Clearly not, it is spurious, fostered by factors such as increasing life expectancy and sedentary lifestyle in the case of CVD, and greater awareness of the disadvantages of large families in the case of birth rates. Both trends are part of the same development process.

Correlation do not imply causation — increasing the birth rate will not reduce the incidence of CVD. Louizi *et al.*[2] discussed the role of carotid stenosis as a possible explanation for a spurious correlation between white matter hyperintensities and coronary artery disease.

Cause is a strong term. Although its usage varies, it generally refers to a factor directly responsible for an outcome, which precedes the effect. For instance, smoking can be considered a cause of lung cancer, although the mechanism is mediated through cotinine, which has the potential to alter cell structure. A similar situation exists between obesity and diabetes — the association appears direct but works through metabolic mediators.

It is sometimes argued that a factor qualifies as a cause if it is both necessary and sufficient for the effect to appear. However, this is an overly strict condition in medical contexts. Obesity is neither necessary nor sufficient to cause diabetes, yet it is still considered as a causal factor. Thus, most medical relationships fall at least partly within the statistical domain. If the presence of a factor increases the chances of an outcome and its absence reduces the likelihood, the factor can reasonably be considered a cause, although this could be one of several possible causes. This aspect is often overlooked, despite its relevance to interpreting medical outcomes.

'Just how much evidence is enough to conclude a cause-effect relationship?' is a crucial question that does not have a clear answer. Nonetheless, the following set of criteria is often used to infer a cause-effect relationship:

(i) **Biological plausibility –** The strongest evidence for a cause-effect relationship arises from a biological explanation of how the presence of a factor triggers a biological response that produces the effect. This mechanism may be obscure in some

cases until it is fully explored, and the epistemic uncertainties due to incomplete knowledge may also be a limiting factor. Bellavite and Imbriano[3] described the biological mechanism underlying the protective effects of hesperidin, reinforcing the causal relationship between skin aging and hesperidin.

(ii) **Experimental evidence –** A clinical trial with pre-defined inclusion and exclusion criteria, and features such as randomisation and blinding, provides controlled conditions that attribute the observed effect almost exclusively to the intervention. Since clinical trials cannot ethically be conducted for factors with unknown or harmful effects, the other criteria in this list are often relied upon to infer cause-effect relationship in such cases.

(iii) **Dose-response relationship –** A cause-effect relationship is supported when a factor of greater magnitude produces a greater effect. For example, blood pressure (BP) levels and coronary events exhibit a dose-response relationship. Liu and Qiao[4] studied a similar relationship between sleep duration and the mediation of phenotypic age acceleration.

(iv) **Consistency –** This criterion has several facets: (a) the effect must occur across different populations where the cause exists; (b) the relationship should persist across different time periods; and most importantly; (c) under identical conditions, a specific magnitude of the cause should consistently produce a similar effect. For instance, the relationship between homocysteine levels and coronary disease was found to lack consistency and was therefore discarded as a cause (Unadkat *et al*).[5]

(v) **Specificity –** For a true cause-effect relationship, the absence of the factor should be associated with a reduced magnitude of the outcome, if not its complete absence. This can be viewed as extension of the dose-response concept, though the emphasis here is on absence. For example, if anaemia is absent in pregnant women, can that reasonably ensure normal birth weight?

(vi) **Statistical significance –** Empiricism is the backbone of medical research; conclusions are drawn from data-based evidence. However, all empirical studies rely on samples, as future cases cannot be included. Sampling fluctuations mean we can never be completely certain, and Type-I and Type-II errors are unavoidable. Statistical significance and adequate study power are therefore the best safeguards for keeping these errors within acceptable limits and ensuring confidence that the observed effect is real. Thus, a relationship must be statistically significant to support the conclusion that it exists.

None of these criteria alone is sufficient to conclude a cause-effect relationship. The starting point in empirical research should be the statistical significance of a correlation, association, regression, or difference between the groups. If this significance is lacking, there is no reasonable assurance that the relationship exists, and further exploration is futile. When statistical significance is achieved, one should examine alternative explanations such as bias, inappropriate study design, or confounding factors. If no alternative explanation is available, and at least four of the six criteria above are fulfilled, a cause-effect relationship may be reasonably concluded.

Conflicts of interest: None.

Funding: None.

Conclusion

A relationship between medical factors is statistical when the change in one is accompanied by change in one or more of the others. This relationship can be spurious when it arises due to a third intervening factor. Cause-effect relationship can be concluded when it satisfies most of the criteria listed in this paper.

Abhaya Indrayan. Research Methodology and Biostatistics Series VIII – Statistical Relationship Versus Cause-Effect Relationship Between Medical Factors. MMJ. 2025, December. Vol 2 (4).
DOI: https://doi.org/10.62830/mmj2-04-32c

References

1. Skovlund E. Enkel lineær regresjon. Tidsskrift for Den norske legeforening. 26th October 2020.
2. Louizi C, Khadhraoui E, Lotz J, *et al.* Association of cervical artery stenosis with common cerebral microvascular lesions and coronary artery calcifications. *Front Neuroimaging.* 2025;4:1559481.
3. Bellavite P, Imbriano A. Skin Photoaging and the Biological Mechanism of the Protective Effects of Hesperidin and Derived Molecules. *Antioxidants.* 2025;14(7):788.
4. Liu Y, Qiao K. Dose-response relationship between sleep duration and mediation of phenotypic age acceleration: A cross-sectional study. *Medicine.* 2025;104(40):e44786.
5. Unadkat SV, Padhi BK, Bhongir AV, *et al.* Association between homocysteine and coronary artery disease-trend over time and across the regions: a systematic review and meta-analysis. *Egypt Heart J.* 2024;76(1):29.