Continue

Complex regression procedures like mediation and moderation are best explained with a combination of plain language and a figure. For mediation, a path diagram that illustrates the mediational relationship and indicates beta weights is most useful. The statistical significance of the indirect effect should be tested using bootstrapping (see Hayes [2018], Introduction to mediation, moderation, and conditional process analysis). For moderation, a figure showing conditional/simple slopes at different levels of the moderator (typically 1 SD above, 1 SD below, and the M of the moderator variable for continuous moderators) is most useful. A brief, simulated example of how to report simple mediation:The relationship between math ability and interest in becoming a math major was mediated by math self-efficacy. As Figure 1 illustrates, the regression coefficient between math ability and math self-efficacy was statistically significant, as was the regression coefficient between math self-efficacy and interest in the math major. The indirect effect was (.47)(.36) = .17. We tested the significance of this indirect effect using bootstrapping procedures. Unstandardized indirect effects were computed for each of 10,000 bootstrapped samples, and the 95% confidence interval was computed by determining the indirect effects at the 2.5th and 97.5th percentiles. The bootstrapped unstandardized indirect effect was .17, and the 95% confidence interval ranged from .04 to .32. Thus, the indirect effect was statistically significant. A brief, simulated example of how to report moderation:Negative affect was examined as a moderator of the relation between social support and job burnout. Social support and negative affect were entered in the first step of the regression analysis. In the second step of the regression analysis, the interaction term between negative affect and social support was entered, and it explained a significant increase in variance in job burnout, $\Delta R2 = .03$, $F(1, 335) = 14.61$, $p < .001$. Thus, negative affect was a significant moderator of the relationship between social support and job burnout. The unstandardized simple slope for employees 1 SD below the mean of negative affect was .56, the unstandardized simple slope for employees with a mean level of negative affect was -.08, and the unstandardized simple slope for employees 1 SD above the mean of negative affect was -.72 (see Figure 2). Traditional epidemiological assessments, which mainly focused on evaluating the statistical association between two major components-the exposure and outcome-have recently evolved to ascertain the in-between process, which can explain the underlying causal pathway. Mediation analysis has emerged as a compelling method to disentangle the complex nature of these pathways. The statistical method of mediation analysis has evolved from simple regression analysis to causal mediation analysis, and each amendment refined the underlying mathematical theory and required assumptions. This short guide will introduce the basic statistical framework and assumptions of both traditional and modern mediation analyses, providing examples conducted with real-world data. Keywords: Mediation analysis, Epidemiology, Humans, Logic, Probability In the early days, traditional analytic epidemiological methods mainly focused on the statistical association between two major variables: the exposure (E) and the outcome (Y). However, methods have evolved to explore the “black box” between the E and the Y by investigating the mechanism underlying the association and various pathways. In the same context, the mechanism has also been visualized as being near the center of “Chinese boxes,” or a set of nested boxes. The “black box” is presumed to contain factors, both above and below the level of the individual—the factors above the individual may contain items such as interpersonal dynamics and socioeconomic status, including items related to ethnicity and politics, whereas the factors below the individual level comprise genes, proteins, cells, and organ systems [1]. Mediation analysis was developed to assess this “black box,” and psychologists and social scientists have utilized this framework particularly frequently. Mediation analysis can explore and evaluate biological or social mechanisms, thereby elucidating unknown biological pathways and/or aiding in policy-making [2]. However, because of advances in methodologies, including biostatistics, epidemiological research designs, and causal inference, traditional mediation analysis has evolved and been applied in various fields. In particular, the concept of mediation analysis has been especially appealing in social sciences and psychology. There are several overviews of these topics [3-6], and this study is a guide to the full literature. Mediation was initially hypothesized as a variable in the middle of a causal chain. Previously, most of the epidemiological reports focused on evaluating the simple association between E and Y as in Figure 1A. However, as in Figure 1B, it is shown that an E affects a mediator (M), which in turn affects an Y. The M fully mediates the effect from the E to the Y. However, situations were identified where the M does not fully mediate the effect of E on the Y, which led to the concept of partial mediation, as depicted in Figure 1C. As shown in Figure 1C, the effect of an E can be exerted directly on an Y (direct effect, path c’) or take a detour via a M (indirect effect, paths a and b). Initially, the criteria to be regarded as a M were that E should have a statistically significant association with M, and that M should also have a statistically significant association with Y. The initial criteria also included the condition that the mediation analysis could be performed only if there was a statistically significant association between E and Y; this significant relationship between E and Y should be no longer significant after controlling for the previous paths from E to M and M to Y. However, the latter two conditions were further criticized due to the existence of inconsistent and partial mediation, and were therefore omitted from the essential conditions needed for mediation analysis. A conceptual diagram of mediation analysis (A) traditional epidemiological assessment, (B) full mediation, and (C) partial mediation.In contrast to a moderator or confounder, a M is interpreted as involving a causal pathway between E and Y. A detailed definition of a M is provided in the work of Robins and Greenland [7]. The seminal work on this concept of a M or intervening variable was based on Judd and Kenny [8,9] and Baron and Kenny [10]’s article utilizing the regression method. In Judd and Kenny[8,9]’s difference of coefficients approach, mediation analysis can be conceptualized as utilizing two regressions, as follows. First, we run a simple regression analysis with E on Y without M to estimate path c’. Second, we carry out a multivariable regression with E and M to predict Y. In this case, as the coefficient B reflects the total effect (TE), the direct effect from the E to Y c’ shown in Figure 1C, corresponds to B1 in equation 2. The difference method calculates the indirect effect by subtracting the direct effect (c’) from the TE, as follows: This is a simple and widely used approach to screen for the possible presence of a M. However, the logistic regression method has been criticized for lacking a causal interpretation. The difference method has been used to check for mediation, but non-significant findings using this method do not exclude the chance of possible mediation [11]. The other approach is the product method, which was introduced by Sobel and used by Baron and Kenny [10]. In this method, again, a multivariable regression is conducted with E and M to predict Y. However, the next step is to regress M on X and can be written as In equation 3, B reflects path a in Figure 1C, and B2 in equation 2 reflects b in Figure 1C. The coefficient of the indirect effect, Bindirect, is calculated by multiplying the 2 coefficients, B2 and B. Generally, when there is no interaction between an E and a M, these two methods coincide, except for logistic regression. In particular, for rare Ys (approximately under 10%) with no confounding factors, these 2 estimates will, from a practical standpoint, reflect the natural indirect effect (NIE), which will be discussed in the causal mediation section. The difference method is beneficial because there is no restriction of the M distribution; it can be continuous or categorical (including binary). In contrast, the product method requires a linear model to be applied for the M [11]. In situations with common Ys, especially when they are binary, a log-linear regression model instead of logistic regression is recommended [12]. To calculate the confidence interval (CI) of the indirect effect, 2 approaches have been suggested. The first approach utilizes the Sobel test, which is based on the product of 2 normally distributed values of coefficients. In this case, an assumption should be made about the shape of the sampling distribution of the indirect effect. The second approach uses resampling methods, such as bootstrap testing, which does not require a prior assumption of the sampling distribution. Usually, the bootstrap method involves resampling at least 750 times, for which reason the default resampling setting is 1000 times in many macros (e.g., R and the PROCESS macro in SAS [13,14]). Kim et al. [15] conducted a study to estimate the mediating effect of lifestyle factors on the association between social networks and metabolic syndrome, utilizing the baseline data of the community-based Cardiovascular and Metabolic Diseases Etiology Research Center cohort. In total, 10 103 participants were recruited from 2013 to 2018, and their egocentric social network properties were measured using a social network card that was previously applied and standardized [16]. From the raw data of the social network cards, the authors extracted and calculated the size of the social network and the closeness of the social network, which were used as quantitative E variables. Measurements of blood pressure, the lipid profile, fasting glucose, and waist circumference were made in the initial cohort, and metabolic syndrome was defined based on the National Cholesterol Education Program Adult Treatment Panel III criteria as the presence of 3 or more criteria. As potential Ms, the authors tested 4 domains: physical inactiveness (3 categories: vigorous activities, moderate activities, and walking), alcohol consumption (binary variable: current drinker vs. non-drinker), cigarette smoking (binary variable: current smoker vs. non-smoker), and depressive symptoms (continuous variable: range 0-63 by Beck Depressive Inventory-II score). After conducting the multivariable logistic regression for the E (social network properties, continuous variables) and Y (metabolic syndrome, yes/no), mediation analysis was performed with the ‘mediation’ package developed by Imai et al. [17] in the R software [18]. The analysis was conducted in 3 steps: (1) producing a M model, (2) producing an Y model, and (3) conducting a mediation analysis and sensitivity analysis. In the M model, social network properties and other covariates were regressed to explain lifestyle factors. The metabolic syndrome variable was then regressed on social network properties, lifestyle factors, and other covariates. These two models were grouped with the “mediate” function, which was run to estimate the direct effect, indirect effect, and their 95% CI by a quasi-Bayesian Monte Carlo method, including 5000 simulations per estimate set. As there were 4 potential Ms, the authors applied each M and tested the indirect effect. They found that only physical activity significantly mediated the relationship between social network size and metabolic syndrome in both genders (men: effect size [ES]=5.2×10-3, p=0.024; women: ES=3.1×10-3, p