Sample Size and Statistical Power Calculation in Multivariable Analyses: Development and Implementation of "SampleSizeMulti" Packages in R

Víctor J. Vera-Ponce[1,2,*], Fiorella E. Zuzunaga-Montoya[3], Nataly M. Sanchez-Tamay[1,2], Luisa E.M. Vásquez-Romero[1], Joan A. Loayza-Castro[1], Christian H. Huaman-Vega[1,4], Rafael Tapia-Limonchi[1,2] and Carmen I.G. De Carrillo[1,2]

[1]*Instituto de Investigación de Enfermedades Tropicales, Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM), Amazonas, Perú*

[2]*Facultad de Medicina (FAMED), Universidad Nacional Toribio Rodríguez de Mendoza de, Amazonas (UNTRM), Amazonas, Perú*

[3]*Universidad Continental, Lima, Perú*

[4]*Escuela Profesional de Psicología, Facultad de Ciencias de la Salud (FACISA), Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM), Amazonas, Perú*

Abstract: This paper presents advanced methodological approaches and practical tools for sample size calculation in epidemiological studies involving multivariable analyses. Traditional sample size calculation methods often fail to account for the complexity of modern statistical analyses, particularly regarding the correlation between covariates in multivariable models.

We introduce a series of R packages (SampleSizeMulti) designed to address these limitations. These packages offer two distinct calculation approaches: one based on the multiple correlation coefficient between covariates (rho-based method) and another utilizing standard errors from previous studies (SE-based method). These complementary approaches provide comprehensive solutions for different association measures commonly used in epidemiological research: prevalence ratios, odds ratios, risk ratios, and hazard ratios.

The rho-based method innovatively incorporates the explicit consideration of the multiple correlation coefficient between covariates, significantly impacting required sample sizes in multivariable analyses. The SE-based method leverages information from previous studies through their confidence intervals, offering an alternative when correlation estimates are unavailable but published results exist. Furthermore, both approaches integrate crucial logistical considerations, including rejection rates, eligibility criteria, and expected losses to follow-up, providing researchers with realistic estimates of recruitment requirements and timelines.

Seven detailed case studies covering various epidemiological study designs and analytical scenarios demonstrate the practical application of these methods. These examples illustrate how correlation values, standard errors, and logistical factors influence sample size calculations and study planning.

The implementation in R ensures accessibility and reproducibility, while the incorporation of logistical planning tools bridges the gap between theoretical calculations and practical research requirements. These methods represent a significant advancement in study design methodology, potentially improving the quality and efficiency of epidemiological research by ensuring adequate statistical power while optimizing resource utilization.

Keywords: Sample size, Statistical Inference, Regression Analysis, Epidemiological methods, Software Design, Research Design, correlation coefficient (source: Mesh).

INTRODUCTION

Appropriate sample size calculation is a critical component in the design of biomedical and epidemiological research studies. An adequate sample size ensures that the study has sufficient statistical power to detect relevant effects while avoiding the waste of resources on unnecessarily large samples [1]. However, despite its importance, the methods used for sample size calculation often fail to reflect the complexity of modern statistical analyses, particularly in the context of multivariable analyses, especially in observational studies [2].

Traditionally, researchers have relied on simplified methods for calculating sample sizes, such as those based solely on chi-square or Student's t-tests [1,3]. While useful in certain contexts, these methods may be inadequate for studies involving multiple regression analysis, survival models, or complex designs with multiple confounding variables [4]. Consequently, many studies may be under or over-dimensioned,

*Address correspondence to this author at the Instituto de Investigación de Enfermedades Tropicales, Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM), Amazonas, Perú;
E-mail: vicvepo@gmail.com

compromising the validity of their conclusions or leading to inefficient resource utilization [5].

In recent decades, more sophisticated methods for sample size calculation have been developed that account for the multivariable nature of many modern analyses [4]. These methods offer a more precise and flexible approach to determining sample size across various research scenarios [6-8]. However, adopting these advanced methods in practice has been slow, possibly due to their perceived complexity and the lack of accessible tools for their implementation [4].

This review aims to provide a comprehensive overview of advanced methods for sample size calculation in multivariable analyses, with particular emphasis on their practical application. We will review traditional methods and their limitations, present more sophisticated approaches for various studies, and discuss important practical considerations in their implementation. Additionally, we will introduce a new R package that implements these advanced methods, aiming to make these techniques more accessible to the research community.

By providing this review and practical tool, we aim to promote broader adoption of sample size calculation methods that adequately reflect the complexity of modern statistical analyses, thereby enhancing the quality and efficiency of biomedical and epidemiological research.

TRADITIONAL METHODS OF SAMPLE SIZE CALCULATION

Description of Commonly Used Methods

Traditional sample size calculation methods have been widely employed in biomedical and epidemiological research due to their relative simplicity and ease of application. These methods are generally based on formulas for specific statistical tests and simple research scenarios. The following describes some of the most commonly used methods [1,3,9]:

1. **Comparison of Two Proportions:** This method is used when the objective is to compare the proportion of an event between two independent groups. The basic formula is:

$$n = (Z\alpha/2 + Z\beta)^2 * [p1(1-p1) + p2(1-p2)] / (p1-p2)^2$$

   Where n is the sample size per group, Zα/2 and Zβ are the critical values from the normal distribution for the desired significance and power levels, and p1 and p2 are the expected proportions in each group.

2. **Comparison of Two Means:** To compare the means of a continuous variable between two independent groups, the following formula is commonly used:

$$n = 2\sigma^2(Z\alpha/2 + Z\beta)^2 / \Delta^2$$

   Where σ^2 is the assumed common variance, and Δ is the minimum clinically important difference to be detected.

3. **Case-Control Study Calculations:** In case-control studies, the sample size is calculated using the formula:

$$n = [Z\alpha/2\sqrt{(r+1)p(1-p)} + Z\beta\sqrt{(rp1(1-p1) + p2(1-p2))}]^2 / r(p1-p2)^2$$

   Donde r es la razón de controles a casos, p es la proporción de expuestos en la población, y p1 y p2 son las proporciones de expuestos entre casos y controles, respectivamente.

4. **Cohort Study Calculations:** For cohort studies, a commonly used formula is:

$$n = (Z\alpha/2 + Z\beta)^2 * [p1(1-p1)/r + p2(1-p2)] / (p1-p2)^2$$

   Where r is the ratio of unexposed to exposed subjects, and p1 and p2 are the expected incidences in the exposed and unexposed groups, respectively.

These traditional methods have been widely adopted due to their relative simplicity and the availability of tables and software that facilitate calculations. However, as discussed in the following section, these methods have important limitations, particularly when applied to more complex study designs or multivariable analyses.

Limitations in the Context of Multivariable Analyses

Although traditional sample size calculation methods have been widely used, they present significant limitations when applied to multivariable analyses, which are common when examining associations in observational studies. These limitations can lead to inadequate sample sizes, compromising studies' validity and statistical power [5]. The main limitations are described below:

First, traditional methods generally do not consider multiple variables. They are based on bivariate comparisons and do not account for the effect of

numerous independent or confounding variables. In multivariable analyses, such as multiple regression, covariates can significantly affect statistical power and the precision of estimates [10].

Second, these methods tend to underestimate sample size in regression analyses. In the context of multiple regression, traditional methods do not consider the additional variance introduced by covariates, which can result in studies with insufficient statistical power [11].

Third, they ignore the correlation among predictors. Traditional methods do not account for collinearity between predictor variables, which can significantly affect the precision of estimates and, consequently, the required sample size [12].

Fourth, they do not adapt to different types of outcome variables. Many traditional methods assume continuous or binary outcome variables. Still, in practice, researchers often work with more complex outcome variables, such as survival data or counts, requiring specific sample size calculation approaches [13].

These limitations underscore the need for more advanced and flexible methods for sample size calculation in multivariable analyses. Approaches that address these limitations can provide more accurate estimates of the required sample size, thereby improving the validity and efficiency of research studies.

ADVANCED METHODS FOR SAMPLE SIZE CALCULATION

Calculation for Prevalence/Risk Ratios

Sample size calculation for studies involving prevalence ratios (PR) or risk ratios (RR) is crucial in research that compares event occurrence between exposed and unexposed groups. Advanced methods for this calculation consider the multivariable nature of the analysis and provide more precise estimates than traditional approaches.

In this context, the general formula for sample size calculation is based on the Poisson regression model with robust variance, commonly used to estimate PR or RR in cross-sectional or cohort studies. The formula is expressed as follows [4]:

$$n = (z_{1-}\alpha/_2 + z_{1-}\beta)^2 * (1 / (p_0 * (1-p_0)) + 1 / (p_1 * (1-p_1))) / (\ln(MA))^2 * (1 / (1 - \rho^2))$$

Where: n = required sample size per group; $z_{1-}\alpha/_2$ = critical value from the normal distribution for α/2 (significance level); $z_{1-}\beta$ = crucial value from the normal distribution for β (statistical power); p_0 = prevalence or risk in the unexposed group; p_1 = prevalence or risk in the exposed group; MA = measure of association which can be either PR or RR to be detected; ρ^2 = multiple coefficient of determination between the exposure and other covariates

Sample size calculation in studies using PR/RR requires consideration of several fundamental aspects that affect the study's validity and precision. The adjustment for covariates, represented by the term $(1 / (1 - \rho^2))$, is a crucial element that accounts for the correlation between the main exposure and other covariates in the model. This adjustment, particularly important in multivariable analyses, typically increases the required sample size, ensuring that the study maintains its statistical power even after controlling for confounding variables.

The prevalence or incidence of the event under study also plays a fundamental role in determining sample size—the prevalences p_0 and p_1 in the unexposed and exposed groups significantly impact calculations. In particular, larger sample sizes are required to maintain adequate statistical power when studying rare events, which can have important logistical and budgetary implications for the study.

Another determining factor is the magnitude of the effect to be detected, represented by ln(PR) in the formula's denominator. When attempting to detect small effects, that is prevalence or risk ratios close to 1, considerably larger sample sizes are required. This becomes particularly relevant in studies where weak but clinically significant associations are expected.

The specific study design also influences the interpretation of parameters. Although the basic formula is similar for cross-sectional and cohort studies, the interpretation of p_0 and p_1 varies according to the design: in cross-sectional studies, they represent prevalences, while in cohort studies, they reflect risks or cumulative incidences. This distinction is crucial for proper study planning and interpretation.

Alternatively, sample size calculation can be performed using the standard error method, particularly useful when information from previous studies is

available. This method uses the standard error of the logarithm of the measure of association, allowing for more precise estimation when confidence intervals of earlier studies are available. This approach is particularly valuable in contexts where the correlation between covariates is difficult to estimate but published results with their respective confidence intervals are available.

The ratio of exposed to unexposed subjects in observational studies deserves special attention. The proportion of individuals in exposed and unexposed groups is rarely equal in the study population, making it necessary to adjust the sample size calculation considering this unequal distribution. This adjustment is fundamental to ensure adequate statistical power in both groups.

These advanced methods for calculating sample sizes in studies related to these measures of association allow researchers to plan more robust studies, considering both the complexity of multivariable analyses and the specific characteristics of their research designs. Careful consideration of all these aspects, along with the appropriate choice of calculation method (whether based on correlation between covariates or standard error), significantly contributes to the study's validity and efficiency.

Calculation for Case-Control Studies

Case-control studies are widely used research designs in epidemiology, especially for studying rare diseases or those with long latency periods. Due to their retrospective nature and the use of odds ratios (OR) as the measure of association, sample size calculation for these studies requires specific considerations.

The general formula for sample size calculation in case-control studies is based on the logistic regression model and is expressed as follows [4]:

$$n = (z_{1-}\alpha/_2 + z_{1-}\beta)^2 * [p_0(1-p_0) + p_1(1-p_1)] / (p_1 - p_0)^2 * (1 + 1/c) * (1 / (1 - \rho^2))$$

Where: n = required number of cases; $z_{1-}\alpha/_2$ = critical value from the normal distribution for α/2 (significance level); $z_{1-}\beta$ = critical value from the normal distribution for β (statistical power); p_0 = proportion of exposure among controls; p_1 = proportion of exposure among cases; c = ratio of controls per case; ρ^2 = multiple coefficient of determination between the exposure and other covariates.

The proportion p_1 can be calculated from p_0 and the expected OR using the following formula:

$$p_1 = (OR * p_0) / [1 + p_0(OR - 1)]$$

Sample size calculation in case-control studies requires consideration of various methodological aspects that influence the study's precision and validity. In unmatched designs, a fundamental element is the ratio of cases to controls, represented by the term (1 + 1/c), which adjusts the sample size according to the number of controls per case. While increasing the number of controls can improve statistical power, especially when cases are limited, this benefit shows diminishing returns beyond a ratio of 1:4. This consideration is particularly relevant in rare diseases or when cases are difficult to identify [14].

The frequency of exposure in the population plays a crucial role in determining sample size. The proportion of exposure among controls (p_0) significantly impacts calculations, where rare and common exposures generally require larger sample sizes to maintain adequate statistical power. This aspect must be carefully considered during the study planning phase [15].

The magnitude of the effect to be detected directly influences the sample size through its effect on p_1. When attempting to detect small effects, ORs close to 1, considerably larger sample sizes are required. This inverse relationship between effect magnitude and required sample size is fundamental for realistic study planning.

In multivariable analyses, covariates are adjusted through the term (1 / (1 - ρ^2)), which accounts for the correlation between the main exposure and other covariates in the model. This adjustment is crucial for maintaining adequate statistical power when performing adjusted analyses. Alternatively, when information from previous studies is available, the standard error method provides a valuable approach for sample size calculation, particularly useful when published confidence intervals are available.

Matched designs represent a special case requiring additional considerations. In matched case-control studies, sample size primarily depends on the expected proportion of discordant pairs and the number of matched controls per case. This design can increase statistical efficiency by controlling for important confounding factors but requires specific formulas that account for the matched nature of the data. The decision to employ a matched design must balance the

advantages of confounding control against the additional logistical complexities of matching.

These advanced methods for calculating sample size in matched and unmatched case-control studies allow researchers to plan more robust studies, considering the complexity of multivariable analyses and the specific characteristics of each design. Choosing between matched and unmatched designs and carefully considering all these methodological aspects significantly contribute to the study's validity and efficiency. The availability of multiple calculation methods, including the standard error-based approach, provides flexibility to adapt to different scenarios and available information sources.

THE CALCULATION FOR HAZARD RATIO

Sample size calculation for studies using Hazard Ratio (HR) in Cox proportional hazard models requires specific considerations due to the nature of survival analysis and the complexity of multivariable models. Advanced methods for this calculation provide more precise estimates than traditional approaches.

The formula for sample size calculation in this context is derived from the score test for the Cox model and is expressed as follows [4]:

$$n = (z_{1-}\alpha/_2 + z_{1-}\beta)^2 / (\beta^2 j\, \sigma^2 x,j)\psi(1 - \rho^2 j)$$

Where: n = required sample size; $z_{1-}\alpha/_2$ = critical value from the normal distribution for α/2 (significance level); $z_{1-}\beta$ = critical value from the normal distribution for β (statistical power); $\beta^2 j$ = hypothesized value of the regression coefficient under the alternative (logarithm of HR); $\sigma^2 x,j$ = variance of the predictor of interest; ψ = probability that an observation is not censored; $\rho^2 j$ = multiple coefficient of determination between the predictor of interest and other covariates.

A unique aspect of survival studies is the consideration of censoring, represented by the probability ψ, which indicates the expected proportion of events (non-censored) during the follow-up period. This parameter has a direct impact on the required sample size. A larger sample size is needed to maintain adequate statistical power when there is a higher proportion of censored data (i.e., lower ψ). This consideration is particularly relevant in studies with prolonged follow-up periods or where a high rate of losses is expected.

The magnitude of the effect to be detected, expressed through the coefficient $\beta^2 j$ (the squared logarithm of the Hazard Ratio), is another factor in sample size calculation. When attempting to detect small effects, HRs close to 1, considerably larger sample sizes are required. The study planning phase must consider this inverse relationship between the effect magnitude and the sample size needed.

The distribution of the predictor of interest, characterized by its variance ($\sigma^2 x,j$), also directly influences the required sample size. Interestingly, predictors with greater variability allow for detecting the same effect with smaller sample sizes, which can be advantageous in certain research contexts. Alternatively, when information from previous studies is available, the standard error method provides a valuable approach for sample size calculation, particularly useful when confidence intervals for published hazard ratios are available.

Follow-up time is a critical consideration in survival studies, as it must be sufficient to observe the necessary number of events. This aspect is directly related to the probability of censoring (ψ) and has statistical and logistical implications. Proper planning of the follow-up period is essential to ensure that the required number of events is observed while maintaining study feasibility.

These advanced methods for sample size calculation in studies using the Cox model allow researchers to plan more robust studies, considering both the inherent complexity of survival analyses and the multivariable nature of the models. Explicit consideration of censoring and correlation between covariates is particularly important in this context, as is the availability of alternative methods, such as the standard error-based approach. Careful integration of all these aspects in study planning significantly contributes to its validity and efficiency, allowing for a more precise estimation of the effects of interest in survival analysis.

Practical Considerations Effect of Multiple Correlation Coefficient

The multiple correlation coefficient, generally denoted as R^2 or ρ^2, is crucial in advanced sample size calculation for multivariable analyses. This coefficient represents the proportion of variance in the dependent variable explained by the set of independent variables in the model. Its inclusion in sample size calculations has important implications that must be carefully considered in study planning [16].

For the estimation of ρ², it is fundamental to understand its calculation process. A regression model must be performed where the exposure variable acts as the dependent variable (Y), while the covariates must be considered independent variables in the final model function. It is important to note that at this stage of calculation, the main dependent variable of the original analysis is not included. This procedure allows evaluation of the variance explained by additional covariates in the exposure, a fundamental aspect for correctly adjusting the final model and controlling possible confounding factors.

The impact of the multiple correlation coefficient on sample size materializes through the term 1 / (1 - ρ²). This adjustment typically increases the required sample size. For example, with a ρ² of 0.3, the necessary sample size will be multiplied by approximately 1.43, representing a 43% increase. The effect of this adjustment varies significantly according to the value of ρ²: when it is close to zero, it indicates that the covariates explain little of the variability and the adjustment will have a minimal effect; however, when it is high, it suggests that the covariates explain a significant proportion of the variability, resulting in a substantial increase in the required sample size. As rho increases, the required sample size will also increase. Indeed, Figure (**1**) illustrates how the necessary sample size increases non-linearly (exponentially) as the correlation between covariates increases. Specifically, it shows that when correlation is low (rho < 0.25), sample size remains relatively stable, but as correlation increases, especially after 0.5, the required sample size grows more pronouncedly. This relationship reflects the need to compensate for the loss of statistical efficiency caused by collinearity between study variables.

Precise determination of ρ² represents a significant challenge, especially during study planning. Researchers can employ strategies to address this difficulty: reviewing similar literature to obtain ρ² estimates, conducting pilot studies to estimate it in a small sample, or performing sensitivity analyses by calculating sample size for different ρ² values. Each approach has advantages and limitations; the choice will depend on the specific study context.

The implications for study design are substantial and require careful balance between various factors. On the one hand, the balance between precision and feasibility must be considered: a high ρ² can result in very large sample sizes, which may affect the study's practical viability. On the other hand, the selection of covariates directly influences ρ². Although including relevant covariates may increase ρ² and, consequently, the required sample size, it can also significantly improve the precision of estimates.

The study context also significantly influences the value of ρ². Observational studies tend to present higher ρ² values due to the natural correlation between variables in uncontrolled populations. In contrast, experimental studies may show lower ρ² values due to randomization, although these can still be significant, particularly in clinical trials involving multiple baseline risk factors.

Finally, certain limitations and precautions must be considered when using the multiple correlation

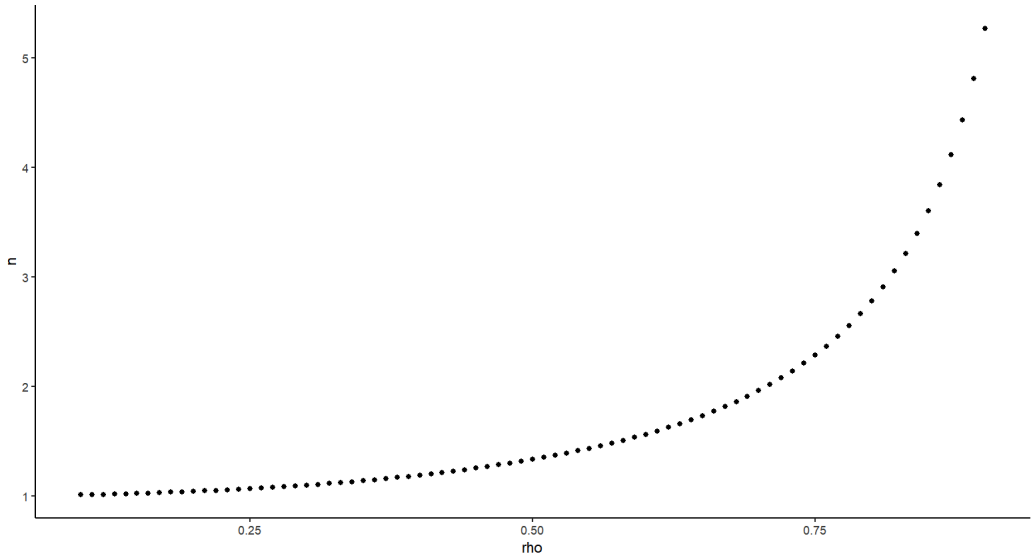

Figure 1: Relationship between correlation (rho) and required sample size.

coefficient. A very high ρ^2 may indicate a risk of model overfitting, especially in small samples, and it is important to remember that a high ρ^2 does not necessarily imply causal relationships between variables. The interpretation and application of the multiple correlation coefficient must be performed carefully, considering the specific study context and objectives.

Exposed/Unexposed Ratio in Observational Studies

In observational studies, particularly cross-sectional and cohort studies, the ratio of exposed to unexposed subjects represents a crucial factor in sample size calculation. Unlike experimental studies, where the investigator can control this ratio, it is determined by the natural prevalence of exposure in the study population. This fundamental characteristic of observational studies has important implications for their design and analysis.

The impact of the exposed-to-unexposed ratio on sample size is direct and significant. An unbalanced ratio substantially different from 1:1 generally requires a larger sample size to maintain the same statistical power. Precise estimation of this ratio can be performed through various methods, including reviewing previous studies in similar populations, analyzing epidemiological surveillance data, or conducting pilot or feasibility studies. Each estimation method has advantages and limitations, and the choice will depend on the specific study context.

The impact of the ratio varies according to the study design. In cohort studies, it directly affects the number of subjects needed in each group, while in case-control studies, it influences matching efficiency and power to detect interactions. Although a 1:1 ratio generally provides maximum statistical efficiency, this distribution is not always feasible or desirable in observational studies, where practical and logistical constraints may dictate other configurations.

Practical considerations play a fundamental role in determining the optimal ratio. In the case of rare exposures, it may be necessary to oversample the exposed group to achieve adequate statistical power. Additionally, economic aspects can significantly influence the decision. In some cases, recruiting more controls than cases may be more cost-effective, leading to unbalanced but economically viable ratios. Considering the study's statistical validity and practical feasibility, these decisions must be made.

The planning process should include sensitivity analyses, varying the exposed to unexposed ratio to evaluate its impact on sample size and statistical power. This exercise provides valuable information about the robustness of the proposed design and can help identify optimal configurations that balance statistical and practical considerations.

The ethical implications of the exposure ratio cannot be ignored. An unbalanced ratio may have significant ethical consequences in certain contexts, particularly when including disproportionate subjects in risk situations. Researchers must consider these ethical implications carefully when designing their studies and obtaining the corresponding approvals.

Design flexibility is crucial in long-term studies. If the observed ratio differs significantly from the expected, recruitment strategies may need to be adjusted during the study. This adaptability is essential to maintaining study validity and achieving the proposed research objectives.

Thus, carefully considering the exposed-to-unexposed ratio is fundamental in designing observational studies. Its impact extends beyond the required sample size, affecting statistical efficiency, study costs, and results interpretation. Researchers must seek an optimal balance between statistical, practical, and ethical considerations when determining the most appropriate sampling strategy for their study.

POWER VERSUS SAMPLE SIZE CALCULATION

Power calculation and sample size calculation are two sides of the same coin in research study design. Both are intrinsically related and based on the same statistical principles but are used at different stages and for other purposes in study planning [17].

While sample size determines the number of subjects or units needed to conduct a study with sufficient validity, statistical power calculation works with a fixed sample size, as frequently occurs with pre-existing databases collected for other purposes. In this context, power is evaluated to determine if the available sample is sufficient to answer the research question adequately. That is, it is calculated to determine if the existing sample has the necessary statistical power to detect a significant effect, thus ensuring the validity of the results.

Furthermore, power calculation, particularly the determination of the β value, can be performed using

the same formulas presented for sample size calculation but adapted to a situation where a fixed sample is already available. In this case, the available sample size is used instead of seeking the necessary sample size, and the probability of committing a type II error (β) is calculated, which is the probability of not detecting an existing real effect. This process evaluates whether the available sample has sufficient statistical power to answer the research question reliably.

It is important to emphasize that there is a direct relationship between sample size and statistical power. Generally, a larger sample size means greater statistical power. Indeed, unlike sample size calculation, which we have been discussing throughout the article, power calculation: a) Can be used in both the planning phase or post-hoc. b) Requires specification of sample size, α, and effect size. c) Is useful for evaluating the adequacy of a given sample size or for interpreting non-significant results.

Logistical Considerations in Sample Size Calculation

Sample size calculation involves determining the number of participants needed to achieve the desired statistical power and considering crucial logistical aspects affecting study feasibility. These logistical adjustments are particularly important for determining the number of subjects needing contact or evaluation.

To make these adjustments, three main factors must be considered:

1. Rejection or non-response rate: The basic formula for adjusting for rejection rate is n_adjusted = n / (1 - p), Where 'n' is the calculated sample size and 'p' is the expected proportion of rejections or non-responses. For example, if a rejection rate of 20% is expected (p = 0.2) and the calculated sample size is 100, it will be necessary to approach 125 people (100/0.8 = 125). This correction ensures that, even with expected rejections, the required sample size will be achieved.
2. Eligibility criteria: After adjusting for rejections, it is necessary to consider what proportion of the population will meet eligibility criteria. If only 20% of the evaluated population will be eligible, the number of people to assess is calculated as n_eligibility = n_adjusted / 0.2. Following the previous example, if we need 125 people who accept to participate, and only 20% will be eligible, we must evaluate 625 people (125/0.2 = 625).
3. Losses during follow-up: In longitudinal or cohort studies, it is crucial to consider losses during follow-up. If a loss rate of 10% is expected, the adjustment would be: n_final = n_eligibility / 0.9
4. Recruitment capacity and estimated time: Once the total number of people to evaluate is determined, it is fundamental to consider daily recruitment capacity and calculate the time needed to complete the study:
    - Daily recruitment = Number of people/medical records/registers that can be evaluated per day
    - Working days per month = Typically 20-22 days, depending on context
    - Total time (months) = (Total number to evaluate / Daily recruitment) / Working days per month

For example, if we need to evaluate 625 people and can process 10 people per day, with 22 working days per month:

- Days needed = $625 / 10 = 62.5$ days
- Months needed = $62.5 / 22 \approx 2.8$ months

It is important to note that these adjustments should be applied sequentially, as each affects the number calculated in the previous step. Furthermore, the proportions used for these adjustments should ideally be based on 1) previous pilot studies, 2) existing literature on similar studies, 3) previous experience in the study population, and 4) specific population characteristics and research settings.

Careful consideration of these logistical aspects ensures study feasibility and helps plan the resources and time needed to complete participant recruitment and follow-up.

IMPLEMENTATION IN R

To facilitate the sample size calculation process in epidemiological studies, the authors have developed a series of R packages that address different association measures commonly used in research. These packages have been designed considering statistical and logistical considerations necessary for study planning.

The first package, SamplePrevRatioMulti, is oriented toward sample size calculation in studies using prevalence ratio as the measure of association. This package allows researchers to specify key parameters such as significance level, desired

statistical power, and the prevalence ratio to be detected. A distinctive feature is its ability to consider different correlation values between covariates, thus providing a more comprehensive assessment of the necessary sample size under various scenarios.

For unmatched case-control studies, the SampleCCNoMatchedIMulti package provides two distinct methodological approaches. The Rho-based method considers the correlation between the exposure of interest and other covariates to be included in the multivariable analysis, requiring parameters such as the proportion of exposure in controls and the OR to detect. The Standard Error-based method leverages information from previous studies using the observed OR and its confidence intervals. Additionally, it includes a function for logistical planning that considers practical aspects such as differentiated rejection rates for cases and controls and estimated recruitment time.

In matched case-control studies, the SampleCC MatchedIMulti package specializes in calculations that consider the matched nature of the data. This package requires specification of the expected proportion of discordant pairs and allows different matching ratios (e.g., multiple controls per case). Like the previous packages, it offers correlation-based and standard error-based methods specifically adapted for matched data.

For cohort studies, the SampleRiskRatioMulti package focuses on sample size calculation when using relative risk as the measure of association. This package incorporates specific adjustments for longitudinal studies, including considerations about losses during follow-up and the expected proportion of exposed and unexposed in the study population.

Finally, the SampleHazardRatioMulti package is designed for studies employing the Cox proportional hazards model. This package handles the additional complexities of survival analyses, such as considering censored data and specifying the expected hazard ratio. Like the previous packages, it offers correlation-based and standard error-based methods, allowing researchers to choose the most appropriate according to available information and study objectives.

These packages have been developed considering researchers' practical needs, including the statistical aspects of sample size calculation and the crucial logistical aspects for successful study planning and execution. The flexibility in calculation methods and the inclusion of functions for logistical planning make these packages comprehensive tools for epidemiological study design.

A particularly useful feature of these packages is their ability to handle multiple scenarios simultaneously, especially concerning the correlation between covariates. This allows researchers to understand how different degrees of correlation can affect the required sample size, facilitating more robust study planning.

The developed packages are available for installation through GitHub, facilitating access to the scientific community. For their implementation, users must have previously installed the "devtools" package in R, which allows the installation of packages from external repositories. Installation is performed using the command devtools::install_github("VicVePo/X"), replacing X with the corresponding package name.

Once installation is complete, the packages can be loaded into the R session using the library("X") command, replacing X with the respective package name. This simple installation and loading process allows researchers to immediately access all sample size calculation and logistical planning functionalities in each package, thus facilitating the design and planning of their epidemiological studies.

EXAMPLES OF USE WITH PRACTICAL CASES

To illustrate the application of the developed packages, we present four practical cases representing different measures of association commonly used in epidemiological research. Each example considers different correlation values (rho) between covariates.

Case 1: Prevalence Ratio - Cross-sectional Study (Using Rho)

A group of researchers plans to study the association between remote work and musculoskeletal disorders in administrative workers. Based on previous literature, they expect to find a prevalence of musculoskeletal disorders of 20% in on-site workers (unexposed) and seek to detect a prevalence ratio of 1.5. The team plans to include sociodemographic and occupational variables in the multivariable analysis, estimating a moderate correlation (rho = 0.3) with these covariates. For the calculation, they consider a significance level of 5% and power of 80%, with a 1:1 ratio between comparison groups.

```
> # Prevalence Ratio Examples
> # Sample size calculation
> sample_size <- SamplePrevRatioMultiENG(
+   alpha = 0.05,          # 5% significance level
+   power = 0.8,           # 80% power
+   p0 = 0.20,             # 20% prevalence in unexposed
+   PR = 1.5,              # prevalence ratio of 1.5
+   r = 1,                 # 1:1 ratio between groups
+   method = "rho",        # correlation method
+   rho_values = 0.3       # moderate correlation
+ )
> print(sample_size)
  rho total_size exposed_size unexposed_size
1 0.3        645          323            323
```

This yields a total of 645 participants. Additionally, for the logistical planning of the study, the researchers estimate a rejection rate of 15% and that approximately 80% of evaluated workers will meet eligibility criteria. Considering that they can determine 15 workers per day, during 22 working days per month:

```
> # Logistics calculation
> logistics <- cross_sectional_logistics(
+   final_n = 645,
+   rejection_rate = 0.15,      # 15% expected rejection
+   eligibility_rate = 0.80,    # 80% will be eligible
+   subjects_per_day = 15,      # can evaluate 15 subjects per day
+   working_days_month = 22     # 22 working days per month
+ )

Study logistics summary:
----------------------------------------
Required final sample: 645
Subjects to evaluate: 759 (considering 15 % rejection rate)
Subjects to invite: 949 (considering 80 % eligibility rate)
Days needed: 64 (evaluating 15 subjects per day)
Months needed: 2.87 (with 22 working days per month)
----------------------------------------
```

Considering the established rejection and eligibility rates, approximately 949 workers must be evaluated to achieve the required sample size. With the planned evaluation capacity, a recruitment time of 64 days is estimated.

Case 2: Prevalence Ratio - Cross-sectional Study (Using SE)

Researchers from a primary care center plan to study the association between prolonged social media use (more than 3 hours daily) and the presence of depressive symptoms in adolescents. They base their study on a previous study with 250 participants with a prevalence ratio of 1.8 (95% CI: 1.3 - 2.5), suggesting a significant association. For the new research, they establish a significance level of 5% and power of 80%. Since confidence intervals from the previous study are available, they choose to use the standard error-based method for sample size calculation.

```
> ################################################################
> # Sample size calculation using standard error method
> sample_size <- SamplePrevRatioMultiENG(
+   alpha = 0.05,             # 5% significance level
+   power = 0.8,              # 80% power
+   PR = 1.8,                 # prevalence ratio from previous study
+   CI_upper = 2.5,           # upper 95% CI limit
+   CI_lower = 1.3,           # lower 95% CI limit
+   r = 1,                    # 1:1 ratio between groups
+   n_previous = 250,         # sample size from previous study
+   method = "se"             # standard error method
+ )
Calculated Standard Error: 0.1676072
> print(sample_size)
  total_size exposed_size unexposed_size
1        160          80            80
```

For logistical planning, the researchers consider that they will work with a local secondary school. They estimate a rejection rate of 20% (assuming the need for consent from both parents and adolescents) and an eligibility rate of 85% (excluding adolescents with a previous diagnosis of depression or under psychiatric treatment). They plan to evaluate 18 adolescents per day during school days:

```
> # Logistics calculation
> logistics <- cross_sectional_logistics(
+   final_n = 160,
+   rejection_rate = 0.20,      # 20% expected rejection
+   eligibility_rate = 0.85,    # 85% will be eligible
+   subjects_per_day = 18,      # can evaluate 18 adolescents per day
+   working_days_month = 22     # 22 working days per month
+ )

Study logistics summary:
----------------------------------------
Required final sample: 160
Subjects to evaluate: 200 (considering 20 % rejection rate)
Subjects to invite: 236 (considering 85 % eligibility rate)
Days needed: 14 (evaluating 18 subjects per day)
Months needed: 0.59 (with 22 working days per month)
----------------------------------------
```

Considering the established rejection and eligibility rates, approximately 236 adolescents must finally be invited to achieve the required sample size. With the planned evaluation capacity, a recruitment time of 14 working days is estimated.

Case 3: Statistical Power Calculation in Cross-sectional Study (Using Rho)

A researcher has completed a cross-sectional study on the association between sedentary behavior (more than 8 hours sitting) and chronic low back pain in office workers. They managed to recruit 300 participants (150 per group), finding a prevalence of low back pain of 25% in non-sedentary workers. The observed prevalence ratio was 1.6, and variables such as age, BMI, and years of work were included in the multivariable analysis, estimating a correlation of 0.45 with these covariates. The researcher wants to calculate the achieved statistical power, considering a significance level of 5%.

```
> # Statistical power calculation
> achieved_power <- SamplePrevRatioMultiENG(
+   alpha = 0.05,           # 5% significance level
+   n = 300,                # achieved sample size
+   p0 = 0.25,              # 25% prevalence in unexposed
+   PR = 1.6,               # observed prevalence ratio
+   r = 1,                  # 1:1 ratio between groups
+   method = "rho",         # correlation method
+   rho_values = 0.45       # correlation with covariables
+ )
> print(achieved_power)
   rho     power
1 0.45 0.6997244
```

The results indicate that the study achieved a statistical power of 69.97% with the sample size and the observed correlation with covariates. This value suggests that if there is a true prevalence ratio of 1.6 in the population, the study had a 76% probability of detecting this association as statistically significant. Although this value is slightly below the traditionally desired 80%, it still represents a reasonable capacity to detect the interest association.

Case 4: Unmatched Case-Control Study (Using Rho with 1:3 design)

A hospital research team seeks to study the association between occupational exposure to organic solvents and the development of interstitial lung disease. Based on previous studies, they estimate that the prevalence of exposure in controls (workers without disease) is 15%. They expect to detect an Odds Ratio of 2.5, considering a moderate correlation (rho = 0.4) with other covariates to be included in the analysis, such as age, smoking, and other occupational exposures. To optimize study efficiency, they plan to recruit three controls for each case (1:3 ratio), with a significance level of 5% and power of 80%.

```
> # Unmatched Case-Control Examples
> library(SampleCCNoMatchedMultiENG)
>
> # Sample size calculation
> sample_size <- SampleCCNoMatchedMultiENG(
+   alpha = 0.05,          # 5% significance level
+   power = 0.8,           # 80% power
+   p0 = 0.15,            # 15% exposure prevalence in controls
+   OR = 2.5,             # Odds Ratio to detect
+   r = 3,                # 3 controls per case
+   method = "rho",       # correlation method
+   rho_values = 0.4      # moderate correlation with covariables
+ )
> print(sample_size)
  rho n_cases n_controls total_size
1 0.4      75        225        300
```

The results indicate that 75 cases and 225 controls are needed (300 participants in total). For logistical planning, the researchers anticipate different rejection rates for cases (10%) and controls (15%), given that cases are already in hospital follow-up. They estimate they can identify 2 potential cases per day and select 3 controls per day from the healthy worker's registry:

```
> # Logistics calculation
> logistics <- case_control_study_logistics(
+   n_cases = sample_size$n_cases[1],
+   n_controls = sample_size$n_controls[1],
+   case_identification_rate = 2,     # 2 cases per day
+   control_selection_rate = 3,       # 3 controls per day
+   case_rejection_rate = 0.10,       # 10% rejection in cases
+   control_rejection_rate = 0.15,    # 15% rejection in controls
+   working_days_month = 22
+ )

Study logistics summary:
----------------------------------------
Required cases: 75
Required controls: 225
Cases to contact: 84 ( 10 % rejection)
Controls to contact: 265 ( 15 % rejection)
Days needed: 89
Months needed: 4.05 ( 22 working days per month)
----------------------------------------
```

Considering the different rejection rates, approximately 84 potential cases and 265 potential controls will need to be contacted. With the established identification and selection rates, a recruitment time of 89 days is estimated.

Case 5: Age and Sex-Matched Case-Control Study (Using Rho)

Researchers from an oncology center plan to study the association between regular consumption of ultra-processed foods and the development of gastric cancer. To minimize the effect of potential confounders, they decided to employ a matched design by age (±3 years) and sex. Based on a pilot study, they estimate that the proportion of discordant pairs (where case and control differ in exposure) will be 35%. The researchers expect to detect an Odds Ratio of 2.0, considering a moderate correlation (rho = 0.35) with other variables to be included in the analysis, such as socioeconomic status, alcohol consumption, and family history. They establish a significance level of 5% and a power of 80%.

```
> # Logistics calculation
> logistics <- matched_study_logistics(
+   n_pairs = sample_size$n_pairs[1],
+   m = 1,                                  # 1:1 matching
+   pair_identification_rate = 3,            # 3 pairs per day
+   pair_rejection_rate = 0.20,             # 20% rejection
+   working_days_month = 22
+ )

Matched study logistics summary:
----------------------------------------
Required pairs: 89
Total participants: 178 ( 1 : 1 matching)
Pairs to contact: 112 ( 20 % rejection)
Days needed: 38 ( 3 pairs per day)
Months needed: 1.73 ( 22 working days per month)
----------------------------------------
```

The results indicate that 89 case-control pairs (178 participants total) are needed. They consider that identifying adequately matched pairs requires considerable logistical planning effort. They estimate they can identify and evaluate 3 case-control pairs per day, with a rejection rate of 20% (considering that both case and control must agree to participate):

```
> # Logistics calculation
> logistics <- matched_study_logistics(
+   n_pairs = sample_size$n_pairs[1],
+   m = 1,                                  # 1:1 matching
+   pair_identification_rate = 3,            # 3 pairs per day
+   pair_rejection_rate = 0.20,             # 20% rejection
+   working_days_month = 22
+ )

Matched study logistics summary:
----------------------------------------
Required pairs: 89
Total participants: 178 ( 1 : 1 matching)
Pairs to contact: 112 ( 20 % rejection)
Days needed: 38 ( 3 pairs per day)
Months needed: 1.73 ( 22 working days per month)
----------------------------------------
```

Considering the 20% rejection rate, approximately 112 potential pairs will need to be contacted. With the established identification capacity of 3 pairs per day during 22 working days per month, a recruitment time of 38 days is estimated.

Case 6: Cohort Study (Using Standard Error)

Researchers from an occupational health clinic plan a cohort study to evaluate whether anemia during pregnancy increases the risk of preterm birth. They base their analysis on a pilot study with 180 workers that found a Relative Risk of 1.8 (95% CI: 1.2 - 2.7) after one year of follow-up. The incidence of preterm birth in women without anemia (unexposed) was 12%. For the new study, they establish a significance level of 5% and power of 80%. Since they have confidence intervals from the previous research, they use the standard error-based method.

```
> # Risk Ratio Examples
> library(SampleRiskRatioMultiENG)
>
> # Sample size calculation
> sample_size <- SampleRiskRatioMultiENG(
+   alpha = 0.05,              # 5% significance level
+   power = 0.8,               # 80% power
+   RR = 1.8,                  # risk ratio from previous study
+   CI_upper = 2.7,            # upper 95% CI limit
+   CI_lower = 1.2,            # lower 95% CI limit
+   n_previous = 180,          # sample size from pilot study
+   method = "se"              # standard error method
+ )
Calculated Standard Error: 0.2068738
> print(sample_size)
  total_sample_size n_exposed n_unexposed
1               176        88          88
```

The results indicate that 296 pregnant women are needed (88 workers in the exposed group and 88 in the unexposed group). Additionally, considering a rejection rate of 10%, a loss to follow-up rate of 15%, and an eligibility rate of 80%, the following logistical calculation was made:

```
> # Logistics calculation
> logistics <- cohort_study_logistics(
+   n_exposed = 200,                  # required number of exposed
+   n_unexposed = 200,                # required number of unexposed
+   recruitment_rate = 5,             # can recruit 5 subjects per day
+   rejection_rate = 0.10,            # 10% will reject participation
+   loss_to_followup_rate = 0.15,     # 15% will be lost during follow-up
+   eligibility_rate = 0.80,          # 80% will be eligible
+   working_days_month = 22           # 22 working days per month
+ )

Study logistics summary:
----------------------------------------
Required final sample: 400
Sample considering losses: 471 ( 15 % losses)
Sample to enroll: 523 ( 10 % rejection)
Sample to evaluate: 654 ( 80 % eligibility)
Days needed: 131 ( 5 persons per day)
Months needed: 5.95 ( 22 working days per month)
----------------------------------------
```

The calculation based on the standard error from the pilot study provides a more precise estimate by incorporating previously observed variability information. Approximately 53 cases of preterm birth are expected during follow-up. Considering the loss and rejection rates, 654 pregnant women will need to be contacted initially. With a recruitment rate of 5 people per month, a recruitment period of 9.8 months is estimated. The total study time, including the year of follow-up, will be approximately 5.95 months.

Case 7: Survival Analysis for Diabetes (Using Standard Error)

Researchers plan a study to evaluate whether high consumption of sugary beverages (more than 2 servings daily) increases the risk of developing type 2 diabetes. They base their analysis on a previous study with 250 participants with a Hazard Ratio of 2.2 (95% CI: 1.4 - 3.5). For the new research, they establish a significance level of 5% and power of 80%. Since they have confidence intervals from the previous study, they use the standard error-based method.

```
> # Hazard Ratio Examples
> library(SampleHazardRatioMultiENG)
>
> # Sample size calculation
> sample_size <- SampleHazardRatioMultiENG(
+   alpha = 0.05,              # 5% significance level
+   power = 0.8,               # 80% power
+   HR = 2.2,                  # Hazard Ratio from previous study
+   CI_upper = 3.5,            # upper 95% CI limit
+   CI_lower = 1.4,            # lower 95% CI limit
+   n_previous = 250,          # sample size from pilot study
+   method = "se"              # standard error method
+ )
Calculated Standard Error: 0.236895
> print(sample_size)
  sample_size
1         178
```

The results indicate that 178 participants in total are needed. For logistical planning, the researchers consider that the planned follow-up time will be 36 months, and the estimated recruitment rate will be 25 participants per month. In contrast, the expected loss to follow-up rate will be 15%, and the initial rejection rate will be 12%. Finally, approximately 30% of participants are expected to develop the event (diabetes) during follow-up. Thus, the following steps will be applied:

```
> # Logistics calculation
> logistics <- survival_study_logistics(
+   total_n = sample_size$sample_size,     # use result from previous calculation
+   expected_events = ceiling(178 * 0.30),  # expect 30% events
+   recruitment_rate = 15,                  # 15 persons per month
+   followup_time = 24,                     # 24 months follow-up
+   loss_to_followup_rate = 0.1,           # 10% losses
+   rejection_rate = 0.2                    # 20% rejection
+ )

Study logistics summary:
----------------------------------------
Required sample size: 178
Expected events: 54
Sample size accounting for losses: 198 ( 10 % losses)
Sample size to contact: 248 ( 20 % rejection)
Recruitment months: 17 ( 15 persons per month)
Total study months: 41 (including 24 months of follow-up)
----------------------------------------
```

Expecting that 30% will develop diabetes during follow-up (approximately 53 events), and considering a loss rate of 10% and rejection rate of 20%, it will be

necessary to contact approximately 250 people initially. With a recruitment rate of 15 participants per month and a planned follow-up of 24 months, it is estimated that the complete study will take approximately 41 months.

DISCUSSION

Research Implications

Adequate sample size calculation is a crucial element in the design of epidemiological studies, and its proper implementation has important research implications. The methods and tools presented in this work address several significant limitations of traditional approaches and offer practical solutions for researchers.

One of the most relevant contributions is the explicit incorporation of correlation between covariates in sample size calculation. Traditional methods often ignore this aspect, resulting in studies with insufficient statistical power [18,19]. Considering rho allows for a more realistic estimation of the necessary sample size, especially in observational studies where predictor variables are often correlated. This adjustment is particularly important in epidemiological studies involving multiple confounding factors.

Another significant aspect is the integration of logistical considerations in the planning process. The gap between theoretical sample size and the number of subjects that must be contacted is frequently underestimated in practice. Our approach, which incorporates rejection rates, eligibility criteria, and losses during follow-up, provides a more realistic view of the resources needed to complete a study successfully. This is especially relevant in contexts with limited resources or when working with hard-to-reach populations.

The developed packages' flexibility in handling different measures of association (PR, RR, OR, and HR) reflects the diversity of epidemiological designs in current practice. This versatility allows researchers to select the most appropriate measure of association for their specific design without compromising the precision of sample size calculation.

Implementing these methods in R, with open-source code and detailed documentation, facilitates their adoption by the research community. This promotes transparency in the sample size calculation process and allows other researchers to reproduce and validate calculations.

In the broader context of epidemiological research, these tools improve the methodological quality of studies. More precise sample size calculation increases the probability of detecting significant effects when they truly exist and helps avoid wasting resources on studies with insufficient statistical power. This is particularly relevant when research reproducibility is under intense scrutiny.

Incorporating these methods into routine research practice could help reduce the number of studies with inconclusive results due to inadequate sample sizes. Furthermore, explicit consideration of logistical aspects from the planning phase can improve the feasibility and efficiency of study execution.

Advantages and limitations of advanced methods

The advanced methods presented offer substantial advantages over traditional approaches to sample size calculation. The main strength lies in their ability to incorporate correlation between covariates, an aspect frequently ignored in conventional methods. This feature is particularly valuable in observational studies, where predictor variables are rarely independent. The flexibility to adjust calculations according to different correlation values (rho) allows researchers to evaluate how interdependence between variables affects the required sample size, facilitating more robust planning. Furthermore, integrating logistical considerations in the calculation process provides a more realistic view of the resources needed to execute the study, a crucial aspect of research feasibility.

However, these methods also present important limitations that must be acknowledged. The precision of calculations heavily depends on the quality of initial estimates, particularly of Rho, which can be difficult to estimate accurately without similar previous studies or pilot data. The additional complexity of these methods may also represent a barrier for researchers less familiar with advanced statistical concepts, although the R implementation seeks to mitigate this limitation. Another important consideration is that, while these methods are more precise in theory, their practical advantage may be marginal in situations where the correlation between covariates is low or when other sources of variability (such as measurement errors or losses to follow-up) have a greater impact on study precision.

COMPARATIVE ANALYSIS AND LIMITATIONS

The SampleSizeMulti packages offer several distinct advantages compared to existing sample size calculation tools. While established packages like 'pwr,' 'power analysis,' and 'G*Power' provide robust solutions for basic study designs, they generally do not account for covariate correlations in multivariable analyses. Commercial software such as PASS and nQuery include some multivariable capabilities, but their proprietary nature and cost can limit accessibility. Our implementation bridges this gap by providing a free, open-source solution specifically designed for multivariable analyses in epidemiological research.

However, it is important to acknowledge certain limitations of the current implementation. The packages assume complete data and may require modification when dealing with substantial missing data. While the current version handles continuous and binary covariates effectively, complex interactions between multiple categorical variables may require additional consideration. Furthermore, the accuracy of the rho-based method depends heavily on the quality of correlation estimates, which may be challenging to obtain in the planning phase of some studies.

The packages currently focus on the most common epidemiological measures of association (PR, OR, RR, HR). Future developments will expand this scope to include additional statistical analyses frequently employed in epidemiological research. Specifically, we plan to incorporate sample size calculation methods for binary and multinomial logistic regression, simple and multiple linear regression, multilevel analysis and mixed models, longitudinal data analysis, models for ordinal dependent variables, and mediation and moderation analyses.

Regarding package maintenance and user support, we have established a systematic approach to updates and improvements. Package updates will be released quarterly, with bug fixes addressed more frequently as needed. Error reports and user suggestions are managed through our GitHub repository, with a structured protocol for review and implementation. Critical issues affecting calculation accuracy receive immediate attention, while feature requests are evaluated and prioritized during quarterly development reviews.

These enhancements and expansions will be implemented gradually to maintain package stability and reliability while expanding its utility for diverse epidemiological research scenarios. The modular development approach allows users to access new features while retaining the simplicity and accessibility of core functionalities.

FUTURE DIRECTIONS AND DEVELOPMENT ROADMAP

While the current implementation of the SampleSize Multi packages addresses critical needs in sample size calculation for common epidemiological designs, we acknowledge several important areas for future development and expansion.

The next phase of development will incorporate support for more complex sampling designs. This includes calculations for multi-stage sampling, cluster sampling adjustments, stratified sampling design considerations, complex survey design power calculations, and weighted analysis accommodations. These additions will enhance the packages' utility for researchers working with more complex study designs while maintaining the user-friendly approach that characterizes the current implementation.

Technical enhancements are also planned to expand the packages' capabilities. These include integrating popular R statistical packages, expanded analytical approaches for specialized study designs, advanced reporting capabilities, interactive visualization features, and automated report-generation tools. These improvements will streamline researchers' workflows and provide more comprehensive analytical support.

We have established several key infrastructure elements to ensure robust development and maintenance. A public GitHub repository has been created for issue tracking, comprehensive unit testing protocols are being implemented, detailed documentation with practical examples is under development, and systematic validation procedures for new features are being established. These measures will ensure the continued reliability and effectiveness of the packages as they evolve.

User engagement and support are crucial elements of our development roadmap. We are implementing a structured system for collecting user feedback, planning regular webinars and training sessions, developing a dedicated website for documentation and resources, and establishing quarterly newsletters to keep users informed of updates and developments. These initiatives will facilitate user adoption and ensure

future developments align with the research community's needs.

These planned developments aim to enhance the utility and accessibility of the packages while maintaining their core focus on practical application in epidemiological research. The implementation will follow a modular approach, ensuring that basic functionality remains straightforward while advanced features are available for more complex needs. Through these improvements, we aim to continue supporting the evolving needs of epidemiological researchers while maintaining the packages' commitment to methodological rigor and practical utility.

CONCLUSIONS AND RECOMMENDATIONS

The advanced methods for sample size calculation presented in this work represent a significant improvement over traditional approaches, especially in the context of epidemiological studies involving multivariable analyses. Incorporating correlation between covariates and logistical considerations in the calculation process allows for more realistic and robust research planning. Implementing these methods through R packages facilitates their practical application and promotes reproducibility in epidemiological research.

The developed experience suggests that researchers should consider the correlation between covariates from the initial stages of study planning. In this regard, conducting pilot studies when possible or utilizing data from similar previous studies is fundamental for a more precise estimation of the multiple correlation coefficient. This preventive approach allows for better analysis of the necessary sample size and reduces the risk of obtaining inconclusive results due to insufficient statistical power.

It is crucial to systematically incorporate logistical considerations in the sample size calculation. Adjustments for expected rejection rates, eligibility criteria, and losses during follow-up are fundamental elements for realistic resource and timeline planning. These practical aspects, often underestimated in traditional calculations, can significantly impact study feasibility and success.

The selection of the most appropriate measure of association for the specific study design, whether prevalence ratio, odds ratio, relative risk, or hazard ratio, should be carefully made using the calculation methods developed for each case. This specificity in the methodological approach ensures greater precision in sample size estimation and contributes to the soundness of the research design.

Maintaining detailed documentation of all parameters used in sample size calculation is essential, including justifications for selected values and sources of information used. This documentation process not only facilitates research transparency and reproducibility but also allows for sensitivity analyses when specific correlation values or expected loss rates are uncertain.

The adoption of these methods can significantly contribute to improving the methodological quality of epidemiological studies. By providing more precise and realistic planning that considers both statistical and logistical aspects of research, these advanced methods represent an important step toward resource optimization and strengthening scientific evidence in epidemiology.

ACKNOWLEDGEMENTS

A special thanks to the members of Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM), Amazonas, Peru, for their support and contributions throughout the completion of this research.

FINANCIAL DISCLOSURE

This study was financed by Vicerectorado de Investigación de la Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

INFORMED CONSENT

This study is an article review; therefore, informed consent is not required.

DATA AVAILABILITY

All R packages described in this manuscript are freely available through GitHub (https://github.com/VicVePo/). The complete source code, documentation, and example files for SamplePrevRatioMulti, SampleCCNoMatchedlMulti, SampleCCMatchedlMulti, SampleRiskRatioMulti, and SampleHazardRatioMulti packages can be accessed and downloaded from this repository. No additional data are required as this is a methodological paper presenting statistical tools and their implementation.

AUTHORS' CONTRIBUTION

Víctor Juan Vera-Ponce: Conceptualization, Investigation, Methodology, Resources, Writing - Original Draft, Writing - Review & Editing

Fiorella E. Zuzunaga-Montoya: Conceptualization, Methodology, Software, Data Curation, Formal analysis, Writing - Review & Editing

Nataly Mayely Sanchez-Tamay: Investigation, Methodology, Writing - Original Draft, Writing - Review & Editing

Joan A. Loayza-Castro: Investigation, Project administration, Writing - Original Draft, Writing - Review & Editing

Luisa Erika Milagros Vásquez-Romero: Investigation, Project administration, Writing - Original Draft, Writing - Review & Editing

Christian Humberto Huaman-Vega: Investigation, Project administration, Writing - Original Draft, Writing - Review & Editing

Rafael Tapia-Limonchi: Validation, Visualization, Supervision, Writing - Original Draft, Writing - Review & Editing

Carmen Inés Gutierrez De Carrillo: Methodology, Supervision, Funding acquisition, Writing - Review & Editing

REFERENCES

[1] García-García JA, Reding-Bernal A, López-Alvarenga JC. Cálculo del tamaño de la muestra en investigación en educación médica. Investig En Educ Médica. 2013; 2(8): 217-24. https://doi.org/10.1016/S2007-5057(13)72715-7

[2] Biau DJ, Kernéis S, Porcher R. Statistics in brief: the importance of sample size in the planning and interpreting medical research. Clin Orthop. 2008; 466(9): 2282-8. https://doi.org/10.1007/s11999-008-0346-9

[3] Noordzij M, Tripepi G, Dekker FW, Zoccali C, Tanck MW, Jager KJ. Sample size calculations: basic principles and common pitfalls. Nephrol Dial Transplant Off Publ Eur Dial Transpl Assoc - Eur Ren Assoc. 2010; 25(5): 1388-93. https://doi.org/10.1093/ndt/gfp732

[4] Vittinghoff E, Glidden DV, Shiboski SC, McCulloch CE. Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models [Internet]. Boston, MA: Springer US; 2012 [citado el 21 de octubre de 2024]. (Statistics for Biology and Health). https://doi.org/10.1007/978-1-4614-1353-0

[5] Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, *et al*. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013; 14(5): 365-76. https://doi.org/10.1038/nrn3475

[6] Althubaiti A. Sample size determination: A practical guide for health researchers. J Gen Fam Med. 2022; 24(2): 72. https://doi.org/10.1002/jgf2.600

[7] Hanley JA. Simple and multiple linear regression: sample size considerations. J Clin Epidemiol. 2016; 79: 112-9. https://doi.org/10.1016/j.jclinepi.2016.05.014

[8] Qin X. Sample size and power calculations for causal mediation analysis: A Tutorial and Shiny App. Behav Res Methods. 2024; 56(3): 1738-69. https://doi.org/10.3758/s13428-023-02118-0

[9] Statistical Methods for Rates and Proportions, 3rd Edition | Wiley [Internet]. Wiley.com. [citado el 21 de octubre de 2024]. Disponible en: https://www.wiley.com/en-in/Statistical+Methods+for+Rates+and+Proportions%2C+3rd+Edition-p-9780471526292

[10] Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol. 2007; 165(6): 710-8. https://doi.org/10.1093/aje/kwk052

[11] Demidenko E. Sample size and optimal design for logistic regression with binary interaction. Stat Med. 2008; 27(1): 36-46. https://doi.org/10.1002/sim.2980

[12] Marill KA. Advanced statistics: linear regression, part II: multiple linear regression. Acad Emerg Med Off J Soc Acad Emerg Med. 2004; 11(1): 94-102. https://doi.org/10.1197/j.aem.2003.09.006

[13] Zurakowski D, Staffa SJ. Statistical power and sample size calculations for time-to-event analysis. J Thorac Cardiovasc Surg. 2023; 166(6): 1542-1547.e1. https://doi.org/10.1016/j.jtcvs.2022.09.023

[14] Ury HK. Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data. Biometrics. 1975; 31(3): 643-9.

[15] Modern Epidemiology [Internet]. [citado el 21 de octubre de 2024]. Disponible en: https://www.wolterskluwer.com/en/solutions/ovid/modern-epidemiology-4634

[16] Cohen J, Cohen P, West SG, Aiken L. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, Third Edition [Internet]. Taylor and Francis; 2013 [citado el 21 de octubre de 2024]. https://doi.org/10.4324/9780203774441

[17] Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2a ed. New York: Routledge; 1988; 567 p. https://doi.org/10.4324/9780203771587

[18] Jenkins DG, Quintana-Ascencio PF. A solution to minimum sample size for regressions. PloS One. 2020; 15(2): e0229345. https://doi.org/10.1371/journal.pone.0229345

[19] Shieh G. Precise confidence intervals of regression-based reference limits: Method comparisons and sample size requirements. Comput Biol Med. 2017; 91: 191-7. https://doi.org/10.1016/j.compbiomed.2017.10.015

Received on 28-09-2024 Accepted on 26-10-2024 Published on 25-11-2024

https://doi.org/10.6000/1929-6029.2024.13.24