Choosing Exploratory, Predictive, or Causal Multivariable Models in Biomedical Research: A Practical Methodological Guide

Victor Juan Vera-Ponce* and Jhosmer Ballena-Caicedo

Facultad de Medicina (FAMED), Universidad Nacional Toribio Rodriguez de Mendoza de Amazonas (UNTRM), Amazonas, Peru

Abstract: *Background*: Multivariable regression is widely used in biomedical research, but models built for different purposes are often treated as if they were interchangeable. This confuses variable handling, covariate adjustment, model evaluation, and interpretation.

Objective: To provide a practical guide for clinicians, biomedical researchers, and collaborating statisticians on how to choose and report multivariable models according to whether the aim is exploratory, predictive, or causal.

Methods: We prepared a narrative methodological tutorial using a targeted search of PubMed, Scopus, and Google Scholar, together with key textbooks and reporting guidance (STROBE, TRIPOD+AI, and PROBAST+AI). We prioritized seminal papers and recent methodological references (2021-2025) on variable prespecification, continuous predictors, validation, calibration, and causal diagrams. Illustrative examples are simulated and are used only for didactic purposes.

Results: The first step is to state the analytic objective explicitly. Exploratory models are used to describe adjusted associations and generate hypotheses; predictive models aim to estimate individual risk and therefore require attention to discrimination, calibration, and internal/external validation; causal models aim to estimate an effect and should rely on temporality, substantive knowledge, and directed acyclic graphs (DAGs) to define adjustment sets. Across objectives, arbitrary dichotomization and univariable screening are discouraged. Continuous predictors should usually be kept on their original scale, with flexible functions such as restricted cubic splines when nonlinearity is plausible. Penalization is generally preferable to stepwise procedures when overfitting is a concern in prediction, whereas full theory-based models are often preferable in causal analyses.

Conclusions: The research question should determine the model, not the reverse. A practical workflow is to define the objective first, prespecify candidate variables, choose a functional form that preserves information, and evaluate the model with objective-specific criteria. Clear separation of exploratory, predictive, and causal aims improves transparency, interpretability, and clinical usefulness.

Keywords: Models, Statistical, Regression Analysis, Causality, Confounding Factors, Epidemiologic, Calibration, Decision Support Techniques.

INTRODUCTION

Multivariable regression models are central tools in biomedical and epidemiologic research because they can summarize adjusted associations, estimate individual risk, and, under explicit assumptions, support causal inference. The problem is that these objectives are often treated as if they required the same model-building strategy, even though they do not. The research question should therefore be specified before choosing covariates, functional forms, model-comparison criteria, and performance measures [1].

In applied work, the consequences of conflating objectives are substantial. Cross-sectional studies frequently report adjusted associations as if they were causal effects, despite limited temporality and persistent risks of confounding, selection bias, and reverse causation [2-4]. Likewise, regression coefficients for adjustment variables are often interpreted as 'independent effects,' even when those variables entered the model only to control confounding or stabilize estimation; this is the classic Table **2** fallacy [3].

Prediction studies have a different failure mode. Many still rely on univariable screening or automated stepwise procedures, and then report only discrimination, often the area under the receiver operating characteristic curve (AUC), without adequate assessment of calibration, optimism, or external validity [5,6]. Causal studies fail in another direction: indiscriminate adjustment for all available variables can block mediating pathways, open collider bias, or remove clinically relevant total effects from the estimand of interest [7,8].

Existing tutorials usually focus on a single framework, for example prediction-model development, causal inference, or variable selection, rather than helping applied investigators decide which framework is appropriate for their question. This manuscript is intended for clinicians, biomedical researchers, and collaborating statisticians who need a practical starting point. Its novelty is not a new algorithm, but an integrated decision framework that contrasts exploratory, predictive, and causal modeling in the same article, adds early visual decision aids, and illustrates recommendations with clearly simulated examples [5,8,53-60].

Accordingly, the aim of this tutorial is to provide a practical methodological guide to multivariable

*Address correspondence to this author at Facultad de Medicina (FAMED), Universidad Nacional Toribio Rodriguez de Mendoza de Amazonas (UNTRM), Amazonas, Peru; E-mail: victor.vera@untrm.edu.pe

modeling in biomedical research. We emphasize four linked decisions: defining the analytic objective, prespecifying candidate variables, handling continuous predictors without unnecessary information loss, and choosing evaluation criteria that are coherent with the intended use of the model.

METHODS

This article is a narrative methodological tutorial rather than a systematic review. It was designed to translate statistical principles into applied decisions that are relevant to clinicians, biomedical researchers, and statisticians involved in observational and prognostic studies.

To develop the tutorial, we performed a targeted search in PubMed, Scopus, and Google Scholar from database inception to January 2026, with an update focused on 2021-2025 to capture recent methodological guidance. Search terms combined concepts related to multivariable regression, variable prespecification, penalization, stepwise selection, restricted cubic splines, fractional polynomials, calibration, internal validation, external validation, confounder selection, directed acyclic graphs, and g-methods. We also consulted major methodological textbooks and reporting guidance, particularly STROBE, TRIPOD+AI, and PROBAST+AI [5,10,52,60]. References were selected purposively for methodological authority, relevance to biomedical applications, and complementarity across exploratory, predictive, and causal objectives. Because the goal was practical synthesis rather than evidence grading, no formal risk-of-bias assessment was undertaken.

All numerical examples in the present article are simulated. Sample sizes, coefficients, and performance measures were chosen to resemble plausible biomedical scenarios, but they do not come from real participants or original datasets. These examples are intended only to demonstrate the consequences of different modeling decisions.

A Practical Starting Point: Define the Model Objective

Before deciding how to build a model, investigators should answer a simpler question: what decision should this model support? If the aim is to describe adjusted associations or generate hypotheses, the model is exploratory. If the aim is to estimate an individual's future risk or prognosis, the model is predictive. If the aim is to estimate what would happen under an exposure or intervention strategy, the model is causal. The same dataset can sometimes support more than one objective, but not with the same specification, adjustment strategy, or interpretation [1,5,8].

Figure **1** provides a simple starting workflow, and Table **1** summarizes the main dos and don'ts for each objective. These distinctions are practical, not semantic. Once the objective is stated clearly, choices about variable pre-specification, shrinkage, validation, and interpretation become more coherent.

Prespecification of Candidate Variables

We use the term prespecification consistently to mean the definition, before inspecting outcome associations, of which variables are candidates for the analysis and how they will be measured. Prespecification reduces unnecessary analytic flexibility and keeps model building aligned with substantive knowledge rather than post hoc convenience.

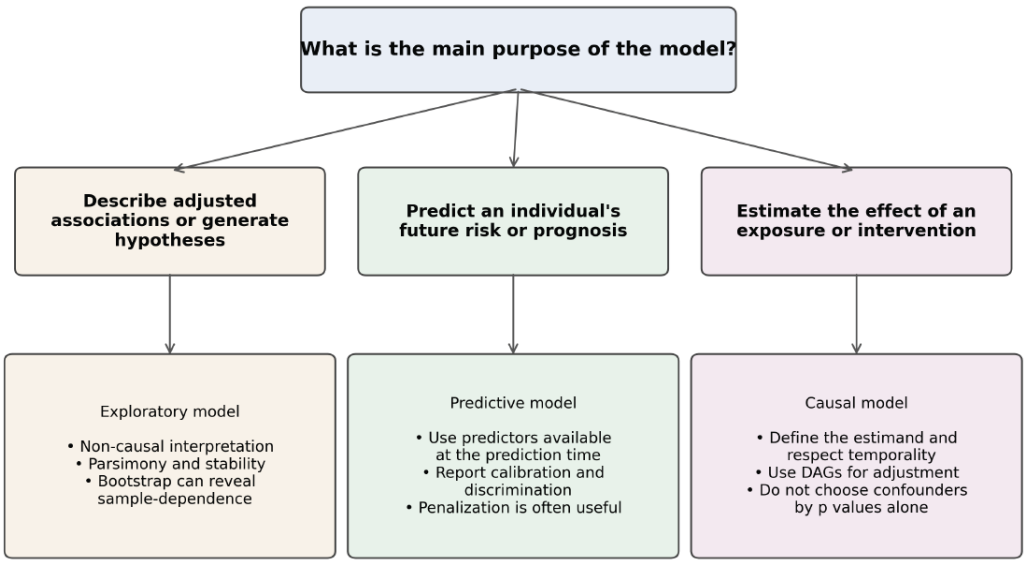

Figure 1: Practical Flow Diagram for Choosing the Model Objective.

Table 1: Practical dos and don'ts According to the Modeling Objective

Objective	Typical question	Priorities (do)	Common errors to avoid
Exploratory	Which variables are associated with the outcome in this setting?	Prespecify plausible candidates; use parsimonious models; report confidence intervals and stability; interpret coefficients as associations.	Causal language; univariable screening; presenting unstable data-driven findings as definitive.
Predictive	Can I estimate an individual's future risk or prognosis?	Use predictors available at the time of prediction; preserve continuous scales; report calibration and discrimination; perform internal and external validation.	Relying only on p values or AUC; using random split-sample validation as the only check; ignoring calibration.
Causal	What is the effect of X on Y?	Define the estimand and temporality; use DAGs to choose adjustment covariates; keep confounders on informative scales; perform sensitivity analyses.	Choosing confounders by significance or AIC/BIC; adjusting for mediators or colliders; causal claims when temporality is unclear.

Abbreviations: AIC, Akaike information criterion; AUC, area under the receiver operating characteristic curve; BIC, Bayesian information criterion; DAG, directed acyclic graph.

In practice, prespecification should consider temporality, measurement quality, missing-data burden, feasibility of routine measurement, and sufficient variability. Rare categories, sparse combinations, or poorly measured variables can produce unstable estimates even before any formal selection procedure is attempted. For binary outcomes, investigators should anticipate separation problems and overly sparse cells when the sample or event count is limited [15].

The content of prespecification depends on the objective. In causal analyses, candidate covariates should be guided by the assumed data-generating mechanism and by the need to control confounding without conditioning on mediators or colliders [8,16]. In predictive analyses, candidate predictors should be available at the time of prediction, reproducible in practice, and proportionate to the amount of information in the data [5,17,57]. In exploratory analyses, prespecification can be broader, but it should still be clinically motivated and explicitly presented as hypothesis generating.

Variable Selection in Practice

Once candidate variables have been prespecified, investigators must decide whether the final model should retain them all or apply some form of reduction. There is no universally best strategy; the preferred approach depends on the model objective, the amount of information in the data, and the cost of overfitting.

Full or theory-based models are often preferable when the candidate set is small to moderate, when predictors were chosen a priori for substantive reasons, or when the analysis is causal and the adjustment set is defined by a DAG or other explicit causal theory. In these settings, the main risk is usually bias from omitting an important variable rather than inefficiency from keeping it [8,16,19].

Penalization methods, especially ridge, least absolute shrinkage and selection operator (LASSO), and elastic net, are most useful when the model is predictive, the candidate set is relatively large, or substantial correlation among predictors is expected. These methods shrink coefficients, reduce overfitting, and typically outperform classic stepwise procedures when generalization is the goal [19,22,23,39,59]. Hyperparameters should be chosen through internal validation, usually bootstrap resampling or cross-validation.

Classic univariable screening and stepwise procedures remain common but should be used with caution. Univariable filters ignore joint confounding and correlation structures; stepwise procedures amplify instability, produce biased coefficients, and encourage an illusion of certainty. Recent critiques of kitchen sink regression emphasize that automated selection based only on p values or information criteria is rarely compatible with a meaningful causal interpretation and often performs poorly for applied prediction as well [18,19,58].

Bootstrap resampling is helpful because it makes model instability visible. Selection frequencies should be interpreted as descriptive evidence of robustness rather than rigid pass/fail criteria: very high frequencies suggest that a result is comparatively stable in the available data, intermediate frequencies indicate meaningful sample dependence, and low frequencies signal that any associated finding should remain clearly hypothesis generating [19,21].

Continuous Predictors: Keep the Scale When Possible

Continuous predictors should usually remain continuous in multivariable models. Dichotomization and other coarse categorization strategies are attractive because they appear simple, but they replace a potentially smooth biological relationship with abrupt artificial jumps, reduce statistical power, and can leave important residual confounding when a continuous confounder is categorized [9,25,26].

For non-technical readers, the practical distinction between the two most common flexible approaches is simple. Restricted cubic splines (RCS) draw a smooth curve by joining polynomial segments between prespecified knots; they are often the easiest flexible option to prespecify and are particularly attractive when the goal is confounding control or transparent applied modeling. Fractional polynomials (FP) search a small set of transformations and can achieve parsimonious nonlinear fits with few parameters, but they are typically more outcome driven and therefore require careful internal validation [27-29].

Categorization can still have a role in communication after the model is fit, for example reporting age groups in a descriptive table, using established public-health categories such as body mass index groups, or translating predicted risks into low/intermediate/high risk bands for decision-making. That communicative role should not be confused with the analytic representation used inside the regression model. Figure **2** and Table **2** summarize the practical implications.

Applied use of AIC and BIC

Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) are tools for comparing competing models fitted to the same outcome and, ideally, the same analytic sample. Smaller values indicate a more favorable trade-off between fit and complexity, with BIC penalizing additional complexity more strongly than AIC [31-36].

In applied biomedical work, these criteria are most useful for comparing reasonable exploratory or predictive specifications, for example a linear term versus a spline, or two prespecified predictor sets. They do not replace internal validation, calibration assessment, or subject-matter reasoning. Most importantly, they should not be the primary basis for choosing confounders in causal models, because a causally necessary variable can be omitted simply for failing to improve global outcome fit enough to satisfy the criterion [8,16,30].

Exploratory Modeling

Exploratory models are appropriate when the purpose is to describe adjusted associations or

Table 2: Practical Options for Handling Continuous Predictors.

Approach	When it is most useful	Main advantage	Main caution
Linear term	When a roughly linear effect is plausible or information is limited.	Easy to interpret; parsimonious.	Can miss important nonlinearity.
Restricted cubic splines	Default flexible option in many applied analyses, especially when transparency or confounding control matters.	Smooth fit with prespecified flexibility; usually easy to explain graphically.	Consumes extra degrees of freedom and should still be validated.
Fractional polynomials	When a parsimonious nonlinear representation is desired and internal validation is feasible.	Flexible with few parameters.	Usually more outcome driven; instability is possible in smaller samples.
Categorization for communication only	After model fitting, to present results in clinically familiar groups or decision bands.	Useful for communication and bedside interpretation.	Should not replace continuous modeling in the main analysis.

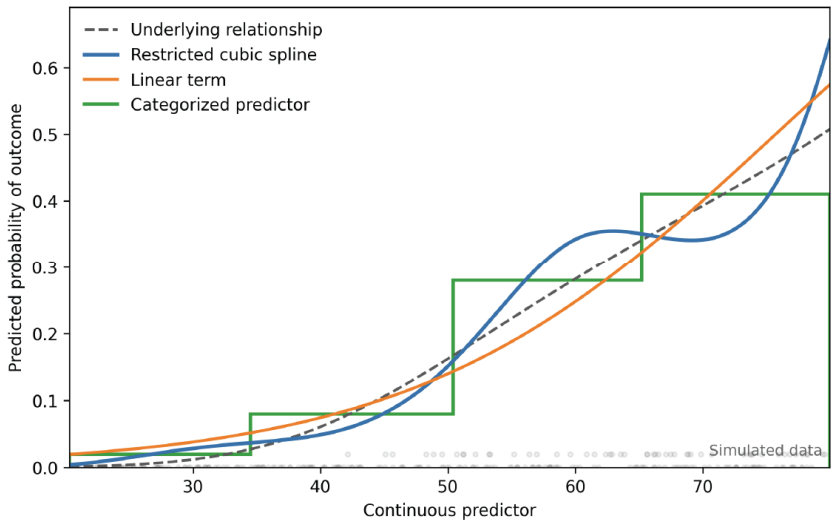

Figure 2: Simulated Illustration of a Continuous Predictor: Categorization Forces Abrupt Risk Jumps, whereas a Spline Preserves the Continuous Pattern.

generate hypotheses in a setting where uncertainty genuinely remains. They can be useful in understudied populations, with new measurements, or when the main value of the study lies in identifying patterns that warrant confirmation in later research. Their key limitation is interpretive: associations from these models should not automatically be treated as causal effects [1-3].

This point is especially important in cross-sectional studies. Even when many variables are entered simultaneously, the resulting coefficients do not become independent causal effects merely because mutual adjustment was performed. Temporality may remain ambiguous, and the same model cannot guarantee the correct confounding set for every coefficient shown in the Tables **2**,**3**.

Model complexity should be planned in relation to the available information. The traditional rule of at least 10 events per variable (EPV) is best treated as a rough warning sign, not a universal threshold. Contemporary sample-size work emphasizes that performance also depends on the total sample size, the number of estimated parameters, outcome prevalence, predictor distributions, and the degree of shrinkage used in the model [17,37,38,57]. When EPV is limited, the safest strategy is often a small full model with clinically prespecified variables. When the information content is larger, carefully supervised reduction can be considered, but the results should remain explicitly hypothesis generating.

Illustrative Simulated Example 1

Consider a cross-sectional survey of hospital physicians and nurses designed to describe factors associated with moderate-to-severe depressive symptoms. Because the purpose is exploratory rather than causal or prognostic, candidate variables are prespecified from the literature, a parsimonious initial full model is fitted, and bootstrap resampling is used to describe which associations are most stable across resamples. Table **3** presents a simplified simulated result summary.

Predictive Modeling

Predictive modeling aims to estimate an individual's future outcome risk at a clearly defined prediction moment and over a clearly defined time horizon. Here, the relevant question is not whether a single coefficient is statistically significant, but whether the model produces useful and transportable predictions. That requires attention to both discrimination and calibration, together with internal and external validation [5,6,39,40,53-60].

A useful practical distinction is whether the study is evaluating the incremental contribution of a new prognostic factor or developing a full multivariable prediction model. In the first situation, a clinically justified base model should usually be prespecified, and the new factor should be evaluated by its contribution to overall performance rather than by its p value alone. In the second situation, greater flexibility is acceptable, but only if accompanied by shrinkage, validation, and transparent reporting [5,39,40,56,59].

When the candidate set is modest and strongly justified clinically, a full model with global shrinkage can be entirely reasonable. When the candidate set is larger, when predictors are correlated, or when flexible terms are needed for several continuous variables, penalization is usually preferable to stepwise elimination. Penalization is therefore most useful when the practical concern is overfitting, whereas full models are most useful when the practical concern is preserving an a priori clinically justified predictor set [19,22,23,39,56,59].

A simple step-by-step workflow is often helpful:

1. Define the prediction moment, the outcome, and the time horizon.
2. Prespecify predictors that would truly be available when the model is used.
3. Handle missing data appropriately and keep continuous predictors continuous whenever possible.

Table 3: Simulated Exploratory Example: Factors Associated with Moderate-to-Severe Depressive Symptoms.

Variable	Full model aPR (95% CI)	Reduced model aPR (95% CI)	Bootstrap inclusion
Age (per 10 years)	1.06 (0.96-1.18)	-	31%
Night shift	1.54 (1.17-2.03)	1.51 (1.16-1.97)	84%
Patients per shift (per 5)	1.11 (0.99-1.24)	1.10 (0.98-1.23)	62%
History of mental disorder	2.39 (1.87-3.06)	2.35 (1.84-3.00)	97%
Risky alcohol use	1.18 (0.91-1.53)	-	43%
Social support (per 10 points)	0.80 (0.72-0.89)	0.81 (0.73-0.89)	91%

Abbreviation: aPR, adjusted prevalence ratio. Bootstrap frequencies are descriptive summaries from 500 resamples and should not be used as rigid cutoffs.

4. Develop the model, and quantify optimism with bootstrap resampling or cross-validation.
5. Report discrimination and calibration together, not one without the other.
6. Evaluate transportability in new data through external validation, and recalibrate or update the model if needed.

Internal validation reuses the development data through resampling to estimate optimism and overfitting. External validation tests the model in genuinely new data, ideally differing by setting, time, or investigators. A random split of a single dataset is usually a weak form of validation because it wastes data for development and does not truly test transportability; temporal or geographic validation is more informative [53,54,56,59].

Calibration deserves explicit attention because a model can discriminate reasonably well and still give systematically wrong absolute risks. Practical assessment should include a calibration plot plus summary measures: the calibration intercept, ideal value 0, where positive values suggest underprediction and negative values suggest overprediction; and the calibration slope, ideal value 1, where values below 1 suggest predictions that are too extreme and values above 1 suggest predictions that are too moderate. The Hosmer-Lemeshow test should not be relied on in isolation because it is strongly sample-size dependent [6,53,54].

Illustrative Simulated Example 2

Suppose a prospective cohort of outpatients with chronic heart failure is used to predict 1-year hospitalization risk. Candidate predictors are prespecified from the literature, continuous variables are represented with restricted cubic splines when needed, and the full predictor set is penalized with a ridge-type Cox model to reduce optimism. Table **4** summarizes the main predictor representations, and Table **5** shows model performance across development and validation samples.

Causal Modeling

Causal modeling aims to estimate the effect of an exposure or intervention on an outcome under explicit assumptions about what would happen under different exposure strategies. The focus is therefore not to predict best and not to screen variables for statistical

Table 4: Simulated Predictive Example: Main Predictor Representations in a Penalized Cox Model.

Predictor	Representation in the model	Illustrative adjusted HR (95% CI)	Practical comment
Age	Restricted cubic spline	1.20 (1.08-1.33) per 10 years	Risk rose more steeply at older ages.
Left ventricular ejection fraction	Restricted cubic spline	0.80 (0.72-0.89) per 10% higher	Lower values were associated with higher risk.
NYHA functional class	Ordinal categories	Class IV vs I: 3.06 (2.05-4.57)	Clear severity gradient.
Serum sodium	Linear term	0.86 (0.78-0.95) per 3 mEq/L higher	Lower sodium indicated worse prognosis.
Creatinine	Restricted cubic spline	1.21 (1.10-1.33) per 0.5 mg/dL higher	Risk increased nonlinearly at higher values.
Loop diuretic use	Binary term	1.36 (1.10-1.68)	Marker of disease severity.
Previous hospitalization	0 / 1 / >=2	>=2 vs 0: 2.44 (1.91-3.12)	Strong recurrence gradient.

Abbreviation: HR, hazard ratio. The complete candidate set was retained and shrunk through penalization; the table displays the clinically most influential predictors for interpretation.

Table 5: Simulated Predictive Example: Discrimination and Calibration in Development and Validation Cohorts.

Metric	Development (n=2000)	Internal validation (n=1000)	External validation (n=850)
Harrell's C (95% CI)	0.762 (0.741-0.783)	0.748 (0.721-0.775)	0.731 (0.704-0.758)
Calibration slope	1.00	0.92	0.89
Calibration intercept	0.00	0.03	0.05
Main interpretation	Apparent performance before optimism correction.	Mild optimism; acceptable calibration.	Reasonable transportability with slight overprediction.

Internal validation was based on temporal splitting plus bootstrap assessment of optimism. External validation corresponds to an independent center.

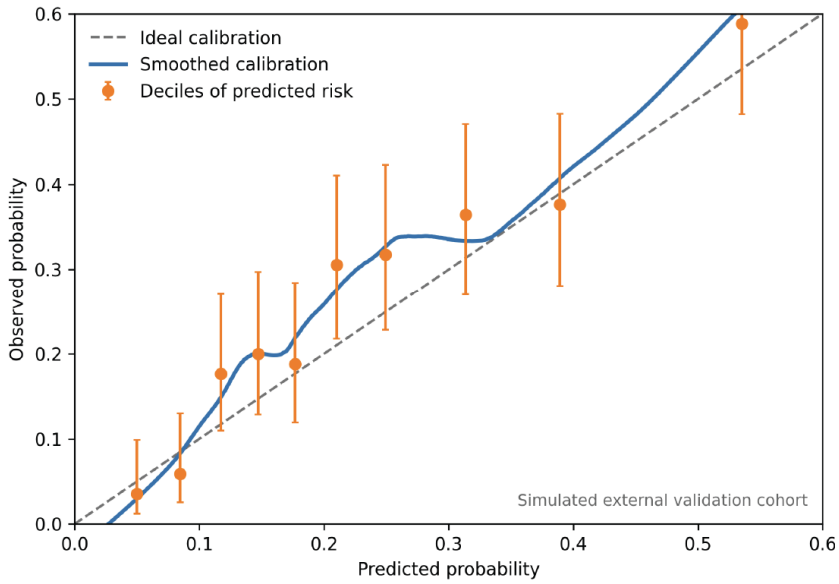

Figure 3: Simulated Calibration Plot for the External Validation Cohort.

significance, but to identify a defensible adjustment set for the causal estimand of interest [4,7,8,16,43].

For applied researchers, a directed acyclic graph can be understood simply as a diagram of assumed cause-effect arrows. Its practical value is that it helps answer three questions before the model is fitted: which variables should be adjusted for because they confound the exposure-outcome relation, which variables should usually not be adjusted for if the total effect is of interest because they are mediators, and which variables can create bias if conditioned on because they are colliders [8,44,45].

Temporality is non-negotiable. The exposure must precede the outcome, and the presumed confounders must be measured before, or at least not after, the exposure-outcome process they are intended to control. Continuous confounders should remain continuous whenever possible, ideally with simple prespecified flexible forms to reduce residual confounding [8,16,26].

When the confounder set is large relative to the information available, the starting point should still be the minimal sufficient adjustment set implied by the causal model. Pragmatic tools such as propensity scores or augmented backward elimination may help with implementation, but they do not replace causal reasoning and should not be confused with proof that a model is causally valid [24,46,47].

Illustrative Simulated Example 3

Consider a cohort study evaluating whether regular night work increases the 5-year incidence of hypertension. Age, socioeconomic status, task type, smoking, alcohol use, baseline body mass index, and Family history are treated as confounders according to

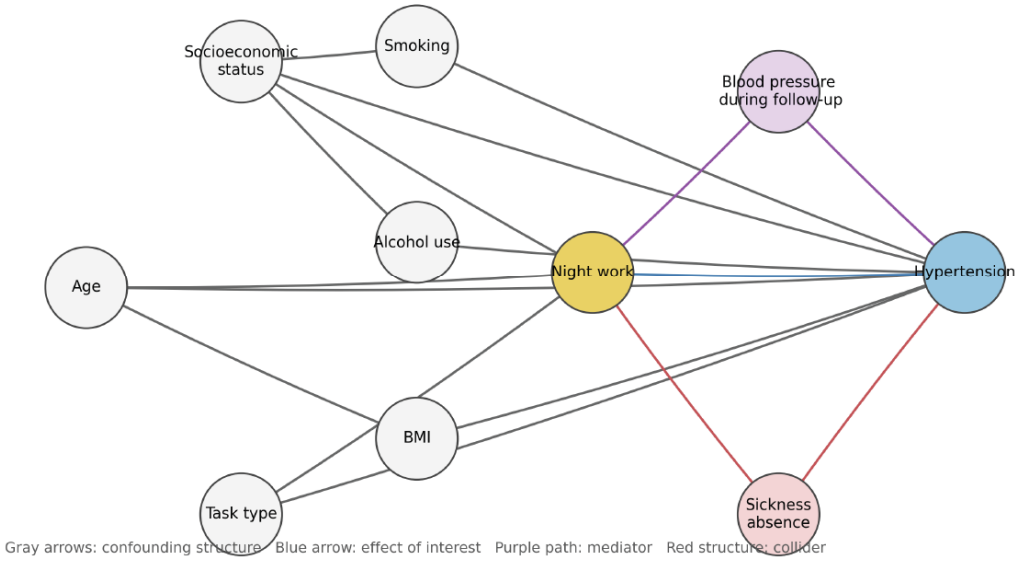

Figure 4: Simplified Directed Acyclic Graph for a Simulated Causal Example Evaluating Night Work and Hypertension.

Table 6: Simulated Causal Example: Effect of Night Work on Incident Hypertension

Model	Adjusted HR (95% CI)	Interpretation
Crude model	1.39 (1.12-1.73)	Unadjusted association.
Age and sex only	1.31 (1.06-1.61)	Residual confounding remains likely.
Full DAG-based confounder set	1.24 (1.01-1.52)	Primary causal estimate for the total effect.
Full set + mediator	1.18 (0.96-1.45)	Likely overadjustment because part of the causal pathway is blocked.
Full set + collider	1.34 (1.09-1.65)	Estimate inflated by collider bias.
Alternative DAG sensitivity analysis	1.25 (1.02-1.54)	Result remained similar under a plausible alternative structure.

Abbreviation: HR, hazard ratio. The E-value for the primary estimate was 1.78, suggesting that a moderately strong unmeasured confounder would be needed to fully explain the association.

the DAG. Intermediate blood pressure during follow-up is treated as a mediator, and sickness absence is treated as a collider. The primary model therefore retains the DAG-based confounder set regardless of individual p values. Table **6** illustrates how the estimate changes under alternative, and sometimes incorrect, adjustment choices.

Short Note on Time-Varying Confounding

A special challenge arises when confounders vary over time and are themselves affected by prior exposure. In this setting, standard regression adjustment can be biased even when many variables are measured correctly. G-methods, including the g-formula, marginal structural models, and doubly robust approaches, were developed for these scenarios [48-51].

Because these methods require stronger assumptions and more specialized implementation, a full treatment is beyond the scope of this practical tutorial. For most applied readers, the key message is simpler: if exposure and confounding both evolve over time with feedback between them, ordinary regression may be inadequate and dedicated causal-methodology support is advisable [43,49-51].

CONCLUSIONS

This tutorial suggests a simple rule: begin by deciding whether you want to explore associations, predict individual risk, or estimate a causal effect. Once that choice is explicit, most modeling decisions become easier and more transparent.

In exploratory work, prioritize a defensible scientific rationale, parsimony, and stability assessment. In prediction, prioritize calibration, discrimination, shrinkage, and external validation. In causal analyses, prioritize temporality, the estimand, and theory-based adjustment guided by DAGs. Across all objectives, avoid univariable screening and arbitrary dichotomization whenever possible. These steps do not make modeling easy, but they make it more credible, interpretable, and useful for biomedical decision-making.

ETHICAL CONSIDERATIONS

This manuscript is a methodological article on multivariable statistical modeling decisions in biomedical research. It did not involve collection or analysis of primary data from human or animal subjects, nor did it use identifiable personal information. Therefore, approval from a Research Ethics Committee was not required. The examples presented were generated via hypothetical simulations for didactic purposes and should be interpreted as illustrations of the analytic process, not as results of original research.

CLINICAL TRIAL REGISTRATION

Not applicable.

PROTOCOL REGISTRATION

Not applicable. Because this is a methodological article without participation of human or animal subjects or analysis of primary data, no protocol registration was performed.

DATA AND MATERIALS AVAILABILITY

No participant datasets were generated or analyzed. The data shown in tables and figures come from hypothetical simulations.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

FUNDING

Article processing charges, if applicable upon acceptance, will be covered by the Vice-rectorate for Research of Universidad Nacional Toribio Rodriguez

de Mendoza de Amazonas. The funding body had no role in the manuscript's conceptualization, methodological decisions, writing, decision to submit or publish, or interpretation of the content.

ACKNOWLEDGEMENTS

We thank Universidad Nacional Toribio Rodriguez de Mendoza de Amazonas, Amazonas, Peru, for the institutional support provided during the preparation of this manuscript.

AUTHOR CONTRIBUTIONS

Victor Juan Vera-Ponce: Conceptualization; Methodology; Software; Formal analysis; Validation; Visualization; Supervision; Funding acquisition; Writing - original draft; Writing - review and editing. Jhosmer Ballena-Caicedo: Conceptualization; Methodology; Validation; Visualization; Writing - original draft; Writing - review and editing.

REFERENCES

[1] Shmueli G. To Explain or to Predict? Stat Sci 2010; 25(3): 289-310. https://doi.org/10.1214/10-STS330

[2] Grimes DA, Schulz KF. Descriptive studies: what they can and cannot do. Lancet 2002; 359(9301): 145-149. https://doi.org/10.1016/S0140-6736(02)07373-7

[3] Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. Am J Epidemiol 2013; 177(4): 292-298. https://doi.org/10.1093/aje/kws412

[4] Hernan MA. The C-Word: Scientific Euphemisms Do Not Improve Causal Inference from Observational Data. Am J Public Health 2018; 108(5): 616-619. https://doi.org/10.2105/AJPH.2018.304337

[5] Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, *et al.* TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024; 385: e078378. https://doi.org/10.1136/bmj-2023-078378

[6] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Bossuyt P, *et al.* Calibration: the Achilles heel of predictive analytics. BMC Med 2019; 17(1): 230. https://doi.org/10.1186/s12916-019-1466-7

[7] Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. Epidemiology 2009 20(4): 488-495. https://doi.org/10.1097/EDE.0b013e3181a819a1

[8] Tennant PWG, Murray EJ, Arnold KF, Berrie L, Fox MP, Gadd SC, *et al.* Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. Int J Epidemiol 2021; 50(2): 620-632. https://doi.org/10.1093/ije/dyaa213

[9] Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 2006; 25(1): 127-141. https://doi.org/10.1002/sim.2331

[10] Vandenbroucke JP, von Elm E, Altman DG, Gotzsche PC, Mulrow CD, Pocock SJ, *et al.* Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Epidemiology 2007; 18(6): 805-835. https://doi.org/10.1097/EDE.0b013e3181577511

[11] Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Stat Med 2007; 26(30): 5512-5528. https://doi.org/10.1002/sim.3148

[12] Box GEP. Science and Statistics. J Am Stat Assoc 1976; 71(356): 791-799. https://doi.org/10.1080/01621459.1976.10480949

[13] Burnham KP, Anderson DR. Multimodel Inference: Understanding AIC and BIC in Model Selection. Sociol Methods Res 2004; 33(2): 261-304. https://doi.org/10.1177/0049124104268644

[14] Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med 2004; 66(3): 411-421. https://doi.org/10.1097/01.psy.0000127692. 23278.a9

[15] Heinze G, Schemper M. A solution to the problem of separation in logistic regression. Stat Med 2002; 21(16): 2409-2419. https://doi.org/10.1002/sim.1047

[16] VanderWeele TJ. Principles of confounder selection. Eur J Epidemiol 2019; 34(3): 211-219. https://doi.org/10.1007/s10654-019-00494-6

[17] 17. Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE, Moons KGM, *et al.* Minimum sample size for developing a multivariable prediction model: part II - binary and time-to-event outcomes. Stat Med 2019; 38(7): 1276-1296. https://doi.org/10.1002/sim.7992

[18] Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. J Clin Epidemiol 1996; 49(8): 907-916. https://doi.org/10.1016/0895-4356(96)00025-X

[19] Heinze G, Wallisch C, Dunkler D. Variable selection - A review and recommendations for the practicing statistician. Biom J 2018; 60(3): 431-449. https://doi.org/10.1002/bimj.201700067

[20] Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996; 15(4): 361-387. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361:AID-SIM168>3.0.CO;2-4

[21] Austin PC, Tu JV. Bootstrap methods for developing predictive models. Am Stat 2004; 58(2): 131-137. https://doi.org/10.1198/0003130043277

[22] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc B 1996; 58(1): 267-288. https://doi.org/10.1111/j.2517-6161 1996.tb02080.x

[23] Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc B 2005; 67(2): 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[24] Dunkler D, Plischke M, Leffondre K, Heinze G. Augmented backward elimination: a pragmatic and purposeful way to develop statistical models. PLoS One 2014; 9(11): e113677. https://doi.org/10.1371/journal.pone.0113677

[25] Greenland S. Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. Epidemiology 1995; 6(4): 356-365. https://doi.org/10.1097/00001648-199507000-00005

[26] Desquilbet L, Mariotti F. Dose-response analyses using restricted cubic spline functions in public health research. Stat Med 2010; 29(9): 1037-1057. https://doi.org/10.1002/sim.3841

[27] Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. J R Stat Soc C 1994; 43(3): 429-453. https://doi.org/10.2307/2986270

[28] Sauerbrei W, Meier-Hirmer C, Benner A, Royston P. Multivariable regression model building by using fractional polynomials: description of SAS, STATA and R programs. Comput Stat Data Anal 2006; 50(12): 3464-3485. https://doi.org/10.1016/j.csda.2005.07.015

[29] Royston P, Sauerbrei W. Stability of multivariable fractional polynomial models with selection of variables and transformations: a bootstrap investigation. Stat Med 2003; 22(4): 639-659. https://doi.org/10.1002/sim.1310

[30] Harrell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. 2nd ed. Cham: Springer 2015. https://doi.org/10.1007/978-3-319-19425-7

[31] Akaike H. A new look at the statistical model identification. IEEE Trans Autom Control 1974; 19(6): 716-723. https://doi.org/10.1109/TAC.1974.1100705

[32] Burnham KP, Anderson DR. Model Selection and Multimodel Inference. 2nd ed. New York: Springer 2004. https://doi.org/10.1007/b97636

[33] Vrieze SI. Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Psychol Methods 2012; 17(2): 228-243. https://doi.org/10.1037/a0027127

[34] Nagelkerke NJD. A note on a general definition of the coefficient of determination. Biometrika 1991; 78(3): 691-692. https://doi.org/10.1093/biomet/78.3.691

[35] Schwarz G. Estimating the dimension of a model. Ann Stat 1978; 6(2): 461-464. https://doi.org/10.1214/aos/1176344136

[36] Kass RE, Raftery AE. Bayes factors. J Am Stat Assoc 1995; 90(430): 773-795. https://doi.org/10.1080/01621459.1995.10476572

[37] Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol 2007; 165(6): 710-718. https://doi.org/10.1093/aje/kwk052

[38] Van Smeden M, de Groot JAH, Moons KGM, Collins GS, Altman DG, Eijkemans MJC, *et al.* No rationale for 1 variable per 10 events criterions for binary logistic regression analysis. BMC Med Res Methodol 2016; 16(1): 163. https://doi.org/10.1186/s12874-016-0267-3

[39] Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J 2014; 35(29): 1925-1931. https://doi.org/10.1093/eurheartj/ehu207

[40] Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. 2nd ed. Cham: Springer 2019. https://doi.org/10.1007/978-3-030-16399-0

[41] Ogundimu EO, Altman DG, Collins GS. Adequate sample size for developing prediction models is not simply related to events per variable. J Clin Epidemiol 2016; 76: 175-182. https://doi.org/10.1016/j.jclinepi.2016.02.031

[42] Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. BMJ 2009; 338: b605. https://doi.org/10.1136/bmj.b605

[43] Hernan MA, Robins JM. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC 2020.

[44] Shrier I, Platt RW. Reducing bias through directed acyclic graphs. BMC Med Res Methodol 2008; 8: 70. https://doi.org/10.1186/1471-2288-8-70

[45] Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology 2004; 15(5): 615-625. https://doi.org/10.1097/01.ede.0000135174.63482.43

[46] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983; 70(1): 41-55. https://doi.org/10.1093/biomet/70.1.41

[47] Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. Am J Epidemiol 2006; 163(12): 1149-1156. https://doi.org/10.1093/aje/kwj149

[48] Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology 2000; 11(5): 550-560. https://doi.org/10.1097/00001648-200009000-00011

[49] Naimi AI, Cole SR, Kennedy EH. An introduction to g methods. Int J Epidemiol 2017; 46(2): 756-762. https://doi.org/10.1093/ije/dyw323

[50] Cole SR, Hernan MA. Constructing inverse probability weights for marginal structural models. Am J Epidemiol 2008; 168(6): 656-664. https://doi.org/10.1093/aje/kwn164

[51] Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics 2005; 61(4): 962-973. https://doi.org/10.1111/j.1541-0420.2005.00377.x

[52] Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, *et al.* PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019; 170(1): 51-58. https://doi.org/10.7326/M18-1376

[53] Collins GS, Dhiman P, Ma J, Schlussel MM, Archer L, Van Calster B, *et al.* Evaluation of clinical prediction models (part 1): from development to external validation. BMJ 2024; 384: e074819. https://doi.org/10.1136/bmj-2023-074819

[54] Riley RD, Archer L, Snell KIE, Ensor J, Debray TPA, Van Calster B, *et al.* Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ 2024; 384: e074820. https://doi.org/10.1136/bmj-2023-074820

[55] Riley RD, Snell KIE, Archer L, Ensor J, Debray TPA, Van Calster B, *et al.* Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study. BMJ 2024; 384: e074821. https://doi.org/10.1136/bmj-2023-074821

[56] Efthimiou O, Seo M, Chalkou K, Debray TPA, Egger M, Salanti G. Developing clinical prediction models: a step-by-step guide. BMJ 2024; 386: e078276. https://doi.org/10.1136/bmj-2023-078276

[57] Dhiman P, Ma J, Qi C, Bullock GS, Sergeant JC, Riley RD, Collins GS. Sample size requirements are not being considered in studies developing prediction models for binary outcomes: a systematic review. BMC Med Res Methodol 2023; 23(1): 188. https://doi.org/10.1186/s12874-023-02008-1

[58] Kuhle S, Brown MM, Stanojevic S. Building a better model: abandon kitchen sink regression. Arch Dis Child Fetal Neonatal Ed 2024; 109(6): 574-579. https://doi.org/10.1136/archdischild-2023-326340

[59] Fung A, Beyene J, Mediratta RP. Principles of Clinical Prediction Model Development and Validation. Hosp Pediatr 2025; 15(6): e280-e285. https://doi.org/10.1542/hpeds.2024-008218

[60] Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro CL, Dhiman P, *et al.* PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ 2025; 388: e082505. https://doi.org/10.1136/bmj-2024-082505

Received on 14-03-2026 Accepted on 12-04-2026 Published on 07-05-2026

https://doi.org/10.6000/1929-6029.2026.15.16