PERSPECTIVE

Unaccounted Variations Can Surreptitiously Spoil the Validity of "Good" Biostatistical Models

Abhaya Indrayan
Biostatistics Consultant, Max Healthcare, New Delhi

CORRESPONDING AUTHOR
Abhaya Indrayan, MSc, MS, PhD(OhioState), Biostatistics Consultant, Max Healthcare, New Delhi
Email: a.indrayan@gmail.com

CITATION
Indrayan A. Unaccounted Variations Can Surreptitiously Spoil the Validity of "Good" Biostatistical Models. Journal of the Epidemiology Foundation of India. 2024;2(4):205-210.
DOI: https://doi.org/10.56450/JEFI.2024.v2i04.009
ARTICLE CYCLE
Received: 05/12/2024; Accepted: 21/12/2024; Published: 31/12/2024

Abstract

Most biostatistical models have enormous underlying uncertainties despite being a good fit for the data. These uncertainties arise not just by variations due to sampling fluctuations in the results but also in estimating its components. The clinical application of these models at the individual level aggravates these variations because the models are obtained and tested for groups, not individuals. Such variations can cause a large imprecision but almost invariably remain unaccounted for due to a lack of awareness amongst the researchers. Add to this are variations due to restricting to a limited number of predictors for achieving parsimony, considering a simple model such as linear in place of, say, quadratic, measurement errors, non-random sampling, and other variations such as between observers and instruments. These also are mostly ignored at the time of developing and interpreting the results of a model. Thus, the models generally fail to give correct results in applications. We illustrate the enormity of the unaccounted variations in models with the help of two examples and suggest ways to minimize them.

Keywords

Validity of Models; Variations; Prediction; Uncertainties

Introduction

Biostatistical models are popular for studying the relationship between two or more variables and are frequently used either for prediction (prediction models) or for understanding the mechanism of the outcome (explanatory models). Knogcold et al(1) developed a model for the prediction of device-measured sitting time based on self-reported sitting time in adults in Norway, and George et al (2) reported a statistical model for the outcome of CoViD-19 cases in different waves of south Indian states. Whether predictive or explanatory, these types of models often fail in applications because many researchers seem unaware of the nuances of the enormous underlying variations that can spoil the validity of the model despite a good fit for the data. Such variations are hardly ever considered in medical research results. We explain these unaccounted variations and their effect with the help of two simple examples that show how the biostatistical models' validity can be surreptitiously seriously affected. This can increase the awareness among medical researchers about the unaccounted variations and may help them to take steps to minimize the effect of these variations. Suggestions for developing improved models are also presented. The first example is on a quantitative outcome and the second is on a qualitative outcome.

Example 1 Statistical model for predicting systolic blood pressure by age and BMI

Consider the possibility of predicting the level of systolic blood pressure (SysBP) (mmHg) in healthy male obese adult residents of a town based on their age and body mass index (BMI). Suppose a survey was conducted on a random sample of 200 healthy male adult (age 30 to 49 years) overweight (BMI≥25) kg/m^2 residents. No other factor was considered in the selection of subjects. The complete data is in the Supplementary Material. The linear regression model obtained for this data is as follows:

SysBP = 96.1 + 0.72(Age) + 0.27(BMI); 30≤Age≤49 years; BMI≥25 kg/m^2.

The square of the multiple correlation coefficient R^2 is 0.89 and this would be considered an extremely good fit. As an explanatory model, this says that each year of age adds an average of 0.72 mmHg to the SysBP and each unit of BMI adds 0.27 mmHg. For prediction, the variables selected as predictors do not matter much although age and BMI are biologically relevant to SysBP in our example. Thus, few will question this model. Statistically, this means that the model is an excellent representation of the data and should be adequate for prediction. However, by considering this model as adequate, we are ignoring 11% uncertainty at the outset because the R^2 is only 0.89. A model with a perfect fit is almost impossible in medicine but this unaccounted variation can not be ignored at the time of interpreting the model adequacy. Oh et al.(3) considered even a low R^2 of 0.71 adequate for their model for the prediction of DXA BMD by routine CT scan. Secondly, the SysBP model under our consideration is based on a specific sample and another sample may give a different R^2. The sampling variation is expressed by the 95% confidence interval (CI), which is 0.87 to 0.92 for a sample size of 200 in this case for R^2 = 0.89. Most papers on models stop at the CI and rarely consider several other hidden variations that play the spoilsport as mentioned next.

Consider a person of age 45 years with BMI = 30 kg/m^2. The predicted SysBP for this person by this model is 136.7 mmHg. However, there are many questions regarding this predicted value. Confidence interval (CI) for mean SysBP for specific age and BMI can be straightaway obtained by using the properties of the Gaussian distribution because of a fairly large sample size. For age = 45 years and BMI = 30 kg/m^2, the 95% CI for mean SysBP is 136.4 to 137.0 mmHg. This is quite narrow because of the relatively large n = 200 in this example. However, this CI is for the population mean for persons of age 45 years and BMI 30, and not for individual values. The prediction interval for an individual(4) of this age and BMI would be necessarily relatively large –133.6 to 139.8 in this case. This is hardly ever considered. What is almost invariably additionally ignored is that the regression coefficients are estimates and subject to sampling fluctuation themselves. The 95% CI for the intercept, age coefficient, and BMI coefficient, which are 96.1, 0.72, and 0.27 in this equation, are 93.6-98.5, 0.68-0.76, and 0.20-0.35, respectively. The latter is really large in this example which can happen due to collinearity between age and BMI. When these lower and upper limits are used, the prediction interval for SysBP becomes 130.2 to 143.2 mmHg for an individual aged 45 years with BMI 30 kg/m^2. Note how quickly the prediction interval inflated in this case when the sampling errors in estimates of the intercept and the regression coefficients are considered. We could not locate any publication where these variations are considered for evaluating the model adequacy. This interval would further enlarge if the possibilities of inadvertent random errors in the measurement of age and BMI are admitted. Both may be correctly assessed but if age is measured as on the last birthday and BMI to the nearest integer, the implied range is 45.0 to 45.9 for age and 29.5 to 30.4 for BMI. These small-looking variations made a difference of nearly 1 mmHg in the predicted SysBP on either side. If inherent variation in measuring SysBP is also admitted, the range for predicted SysBP could finally be 128 to 145 mmHg. This is the uncertainty interval attached to the predicted level of systolic blood pressure for a person whose age and BMI are known. This interval is unexpectedly large, showing the imprecision of the predicted value, and delineates only the aleatory uncertainties(5). These results are summarized in Table 1. Such a large interval in a way shows the limitation of the conventional CI, as well as the inadequacy of the prediction by the statistical model used in this example, and explains why models so commonly fail to give correct prediction. A large and truly representative sample is the only way to minimize this variation. All these calculations are based on Gaussian (Normal) distribution and the uncertainty interval will be different if the distribution is different. These calculations also assume a simple random sample which may or may not be true. There are other caveats also as explained next.

Table 1 Inflation of prediction interval due to aleatory and epistemic uncertainties in Example1 with high R^2 (0.89)

Particulars	**SysBP level (mmHg) for a person aged 45 years with BMI 30 kg/m2**	
	Lower limit	**Upper limit**
Aleatory Uncertainties		
95% confidence interval for mean	136.4	137.0
95% prediction interval for one person	133.6	139.8
Effect of sampling variation in the estimate of the intercept and the regression coefficient for age and BMI	130.2	143.2
Effect of rounding off of age and BMI and inherent variations	128	145
Epistemic Uncertainties still not considered		
Factors other than age and BMI not considered in this model		
Other measures of obesity in place of BMI		
SysBP measured once or the average of 3 readings, and in a restive position or not		
Form of regression – linear, quadratic, or any other		
White coat effect Diurnal variation in SysBP		
Nonresponse and Digit preference		
Inter-observer variation		
Inter-instrument variation		

Now consider epistemic uncertainties(5) associated with such prediction. These arise from a lack of knowledge and avoidable errors. The question is whether age and BMI are the adequate determinants of physiological levels of BP in adult males. If these two factors are not adequate, what other predictors should be considered? These simple-looking questions do not have simple answers and point to the limitation of knowledge on this aspect. Depending on how these questions are answered, the predicted SysBP would change when new variables are included in the model.

Even if age and BMI are considered appropriate determinants, errors may occur because BMI is used as a surrogate for obesity. There are suggestions that waist-hip ratio, skin-fold thickness, waist circumference, index of conicity, and weight-height ratio can also be used. There is no universally accepted criterion to measure obesity. Some data on height and weight measurements may be wrong either because the instruments have unknown errors, or the observer is not careful. The age may not be correctly known. All these are mentioned here in the context of the development of the model but can also occur at the time of applications when age and BMI are measured for a new person whose SysBP is to be predicted.

On the outcome side, SysBP can be just one reading or can be an average of three readings and can be taken in sitting, lying, or any other position. Accordingly, the results could vary, although the variation may not be large in these instances. Because of diurnal variation in SysBP(6), all measurements have to be taken at a specific time of the day for all the subjects and in similar posture and surroundings. The prediction too would be valid for this setting – a caveat quite often forgotten. It is sometimes not possible to adhere to this strictly. Some subjects may not be fully relaxed when BP is measured. There might also be some 'white-coat effect' that occurs while facing a doctor(7).

The regression model in this example is linear. This is the most common and the most preferred form because of its simplicity. However, it is not known what functional form best expresses the level of systolic blood pressure in terms of age and BMI. Various other forms, such as quadratic and logarithmic, can be tried and the one that provides the best empirical fit can be adopted. Most will consider such an exercise redundant since $R^2 = 0.89$ for the linear model is high. However, there is a scope for improvement by considering other forms of regression. Another problem is that a very large number of options are available for the form of regression, and it may not be possible to try all of them. Each model may give different values of predicted levels of SysBP and different uncertainty intervals.

This survey was intended on a random sample of subjects from an area. If the selection strategy actually adopted was different from simple random, such as cluster random, an adjustment in the CI would be required. The selection process should be examined to assess whether the sample was indeed random or not. A non-random sample can give completely misguided results. Then is the question of cooperation of the subjects. Nonresponse, if any, would also affect the results.

In any survey of this type, there could be other non-sampling errors. Digit preference in blood pressure readings is known(8). Hopefully, the instruments used for measuring SysBP, height, and weight are standardized and accurate. Errors in recording and data entry to the computer also have to be ruled out. If a sphygmomanometer is used, the hearing acuity of the observer and the care adopted in deflating the cuff can affect the reading. In the case of electronic equipment, the replicability has to be ascertained. If there is more than one observer, the inter-observer differences may not be negligible. Thus, a large variety of sources of uncertainties exist that put a question mark on the results, and the validity of the model. Unfortunately, no method is available yet to quantify the effect of such unaccounted variations.

All these clearly show that a perfectly valid predictive model may not be able to predict BP near the truth unless a large number of precautions are taken, and all the variabilities are properly accounted for. This is almost never done.

Example 2: Model for estimating the incidence of adverse effects of rimegepant for acute treatment of migraine

Consider rimegepant 75 mg given to 600 patients for acute treatment of migraine by a consortium of hospitals. The drug can cause adverse effects as reported by Gao et al(9). Suppose a total of 4.7 percent of cases report drug-related adverse effects assessed by nausea, dizziness, urinary tract infection (UTI), and liver injury within a month. Suppose the antecedent factors of interest in the model for estimating the adverse effects are the age of the patient, sex, and general health condition visually categorized as good, fair, or poor. The statistical objective is to ascertain the uncertainties associated with the adverse effects when a new patient is prescribed rimegepant. With this large sample, the results are expected to be precise, and the estimated 4.7 percent of the incidence of adverse effects believable!

The most obvious aleatory uncertainty is the sampling fluctuation as in the case of SysBP model. Another group of 600 patients may reveal adverse effects in 4.8 or 5.2 percent of cases. If the sample is random from a specific target population, a CI can be built around it. This is not possible with a non-random sample. In this example, the CI is expected to be quite narrow since the sample size is large but it is not so narrow. The 95% CI in this case is 3.0 percent to 6.4 percent. This itself is quite wide. The sample estimate may be 4.7 percent but it is not unlikely to be as low as 3.0 and as high as 6.4 percent in the target population.

The CI is valid only for the population represented by the sample. If age, sex, and general health in another population are different, the incidence could also be different. For example, the incidence of side effects could be lower in young patients than in old patients. When these factors are varied within the plausible range, different CIs would be obtained. This variation will depend on how these factors affect the incidence of side effects. For an individual patient, his or her age, sex, and general health condition have to be considered. For our example, we assume for illustration that this variation makes a minor difference of 0.3 percent on either side, and the new limits for the incidence of adverse effects become from 3.0-6.4 to 2.7-6.7 percent. These limits are with the natural variations but there might be errors in assessing the antecedents, particularly the general health condition. These errors would vary from individual to individual, and this can inflate the limits of incidence of side effects to 2.5 to 6.9 percent.

This kind of variation can also occur with the outcome – in this case, the adverse effects in terms of nausea, dizziness, UTI, and liver injury. While sample-to-sample variation in these is accounted for in the CI, the subjective variation in their reporting and assessment is not included in the CI. When this is considered, the limits for the incidence of side effects may expand to 2.3 to 7.1 percent. This is the uncertainty interval and can be computed when the complete information on the variation is available.

The uncertainty level would further increase when other epistemic variations are considered. For example, factors other than age, sex, and general health condition, such as food and alcohol intake, are not considered in this model. These can affect the incidence of adverse effects. Some people may not report side effects, and some may exaggerate. This will depend on the skill of the interviewer. While the aleatory variations can be minimized by studying a large representative sample, control of epistemic uncertainties requires extreme care in obtaining the correct data and its proper collation.

Besides the stage of developing model, similar errors can occur at the time of application to individuals coming to a clinic. These may alter the predicted value even when the model is accurate. The correct probability assessment by the model depends on the accuracy of the measurement of age and general health condition in this example. If there are errors, such as incorrect age or incorrect assessment of general health, the prediction loses accuracy. The model already has in-built variations as explained in the preceding paragraphs, application to individuals has its own perils. For correct prediction, the conditions underlying the model development must be accurately replicated in the individuals besides that the model should have minimal underlying variations.

Conclusion

The basic message from Examples 1 and 2 is that the uncertainty around an estimate provided by a biostatistical model is much more than what is made out by the conventional statistical confidence interval. Consideration of aleatory variations with different components of the model may provide an enormously large uncertainty interval, and epistemic bottlenecks put a further question mark on the validity of this interval. Many such variations go unnoticed and uncared for, leading to unexpected results in many cases. Thus, all precautions must be taken to adjust for such variations while building up and reporting biostatistical models. These include adequate predictivity, large sample size, appropriate selection of the predictors, correct and unbiased measurements of the antecedents and the outcome, adjustment for errors such as non-response, building up a comprehensive model, and, above all, a humble conclusion that leaves scope for improvement. At the time of application of the models to individuals coming to a clinic, the variations from the conditions under which the model was developed must also be considered.

Financial Support and Sponsorship

Nil

Conflict of Interest

There are no conflicts of interest.

Declaration Of Generative Ai And Ai Assisted Technologies In The Writing Process

The authors haven't used any generative AI/AI assisted technologies in the writing process.

References

1. Kongsvold A, Flaaten M, Logacjov A, Skarpsno ES, Bach K, Nilsen TIL, Mork PJ. Correction: Can the bias of self-reported sitting time be corrected? A statistical model validation study based on data from 23 993 adults in the Norwegian HUNT study. Int J Behav Nutr Phys Act. 2023 Dec 18;20(1):147.
2. George N, Prasad JB, Verma P. Statistical Model for COVID-19 in Different Waves of South Indian States. Dialogues Health. 2022 Dec;1:100016.
3. Oh J, Kim B, Oh G, Hwangbo Y, Ye JC. End-to-end semi-supervised opportunistic osteoporosis screening using computed tomography. Endocrinol Metab (Seoul). 2024 May 9.
4. Kutner MH, Nachtesheim CJ, Neter J, Li W. Applied Linear Statistical Models Fifth Edition. McGraw-Hill, 2005.
5. Indrayan A, Malhotra RK. Medical Biostatistics, Fourth Edition. CRC Press, 2018.
6. Kawano, Y. Diurnal blood pressure variation and related behavioral factors. Hypertens Res 2011;34:281–285.
7. Cai P, Peng Y, Wang Y, Wang X. Effect of white-coat hypertension on arterial stiffness: A meta-analysis. Medicine (Baltimore). 2018 Oct;97(42):e12888.
8. Hessel PA. Terminal digit preference in blood pressure measurements: effects on epidemiological associations. Int J Epidemiol. 1986 Mar;15(1):122-5.
9. Gao B, Yang Y, Wang Z, Sun Y, Chen Z, Zhu Y, Wang Z. Efficacy and safety of rimegepant for the acute treatment of migraine: Evidence from randomized controlled trials. Front Pharmacol. 2020 Jan 24;10:1577.