Why Most Biostatistical Models Fail to Provide Valid Results

Abhaya Indrayan

(Author:

Medical Biostatistics, 3rd Ed, CRC Press, 2012

Concise Encyclopedia of Biostatistics for Medical Professionals, CRC Press, 2016)

Statistical models are popular for studying relationship between two or more variables and frequently used either for prediction (prediction models) or for understanding the mechanism of the outcome (explanatory models). However, both these types of models many times fail because we do not fully realise the nuances of the enormous underlying uncertainties despite a good fit to the data. We explain these uncertainties with the help of following two examples.

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) in healthy male adult obese residents of hypothetical Townsland. Two important correlates of SysBP are age and obesity. Su[[pose a survey was conducted on a random sample of 200 apparently healthy male adult (age 30 to 49 years) overweight (BMI≥25) residents. No other factor was considered in the selection of subjects. Suppose the regression model obtained is as follows.

SysBP = 96.6 + 0.72(Age) + 0.26(BMI); 30≤Age<49 years; BMI≥25.

Let this have an extremely good fit with the value of square of the multiple correlation coefficient R^2 = 0.87. Note that it does not matter much for prediction which variables are used as predictors although in our example, age and BMI are biologically relevant. Thus few will question this model. This model leaves out 13% variation in the sysBP uncovered because the R^2 is only 0.87 – thus there is inherent deficit. Yet, statistically this means that the model is an excellent representation of the features of the data collected for the development of the model, and should be adequate for prediction. By considering this model as adequate, we are ignoring 13% variation at the outset. Secondly, this model is based on a sample where individuals have variations, and another sample may give different result. When such sampling fluctuations are counted, it is not unlikely that the value of R^2 may dip to just 0.81. Most statisticians will consider even this value good enough to proceed with prediction on the basis of this model. We are starting with an uncertain note already but there are a large number of other uncertainties that play the spoil sport.

Confidence interval (CI) for *mean* SysBP for specific age and BMI can be straightaway obtained by using this equation and properties of Gaussian distribution in view of a fairly large sample size. For age = 45 years and BMI = 26, suppose 95% CI for mean SysBP is 135.1 to 136.2 mmHg. Confidence intervals are for sample mean and not individual values. Prediction interval for an *individual* of this age and BMI would be relatively large, say, 133.1 to 138.2. Statistical theory tells us that CI would be relatively narrow when age and BMI are close to the respective averages of the group. The regression coefficients are estimates and subject to sampling fluctuation

themselves. Simultaneous 95% CI for the age coefficient, which is 0.72 in this equation, could be 0.65 to 0.78, and for BMI coefficient, which is 0.26, it could be 0.16 to 0.36. The latter is really large in this example that can happen due to collinearity between age and BMI. When these lower and upper ends are used, the prediction interval for SysBP becomes 128.5 to 142.8 mmHg for an individual of age = 45 years and BMI = 26. Note how quickly the interval has enlarged in this case when errors in estimates of regression coefficients are considered. This would further enlarge if the possibilities of inadvertent random errors in measurement of age and BMI are admitted. Both may be correctly assessed but if age is measured as on last birthday and BMI to nearest integer, the implied range already is 45.0 to 45.9 for age and 26.5 to 27.4 for BMI. These apparently small looking variations can also make a difference of 1 mmHg in the predicted SysBP. If inherent variation in measuring SysBP is also admitted, the range could finally be something like 126 to 145 mmHg. This is the uncertainty interval attached to the normal level of systolic blood pressure for a person whose age and BMI are known. This interval delineates the **aleatory uncertainties**. But such a large interval in a way shows a limitation of the conventional CI as well as inadequacy of the statistical model used in this example.

Now consider **epistemic uncertainties** associated with such prediction. The question at the outset is whether normal level is person specific, or there is some absolute normal valid for all adults. You may be aware of a debate on what is hypertension. If various body functions indeed work in synchronisation with each other to attain dynamic homeostasis, is it specific to the person? The next question is whether age and BMI are the adequate determinants of physiological levels of BP in adult males. Rise in SysBP with age and BMI can partly transgress into pathological domain. If these two are not adequate, what variables 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 normal SysBP would change.

Even if age and BMI are considered as the appropriate determinants, epistemic uncertainties arise 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. On the outcome side, SysBP can be just one reading or can be average of three readings. Accordingly the results could vary, although the variation may not be large in these instances.

The regression model in this example is linear. This is the most common and most preferred form because of its simplicity. But it is not known what functional form best expresses normal level of systolic blood pressure in terms of age and BMI. Various other forms such as quadratic and inverse can be tried and the one that provides best empirical fit can be adopted. Most will consider it redundant since $R^2 = 0.87$ but a very large number of options are available and it may not be possible to try all of them. Then it needs to be externally validated. Each model may give different values of normal level of SysBP and different uncertainty interval.

Because of diurnal variation in SysBP, all measurements have to be taken on specific time of the day for all the subjects and in a similar posture and surrounding. It is sometimes not possible to adhere to this strictly. Some subjects may not be fully relaxed when measured. There might also be some ‘white-coat effect’ that

occurs while facing a doctor.

This survey was intended on a random sample of subjects from an area. If the design actually adopted were different from simple random, an adjustment in the CI would be required. The selection process should be examined to assess that the sample was indeed random or not. Then is the question of cooperation of the subjects. Nonresponse, if any, would also affect the results.

There would be other nonsampling errors. Digit preference in blood pressure readings is known. Hopefully the instruments used for measuring SysBP, height and weight are standardized and accurate. Errors in recording and in data entry to the computer also have to be ruled out. If a sphygmomanometer is used, 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 have to be ascertained. If there are 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.

All these clearly show that a perfectly valid predictive model may not be able to predict BP to anywhere near the truth.

Example 2: Estimating sexual adverse effects of finasteride

Consider finasteride given to a group of 800 patients 1 mg daily for 12 months for male-pattern hair loss. This drug can cause side effects. Suppose a total of 5.1 percent cases report drug related sexual adverse effects assessed by decreased libido, erectile dysfunction, or decreased volume of ejaculate. Suppose the antecedent factors of interest are age of the patient, general health condition and extent of hair loss at the time of the start of the treatment. General health is categorised as good, fair, or poor, and extent of hair loss as mild, moderate or severe. No other information is available. What kind of uncertainties does this express for sexual adverse effects when a new patient is confronted?

Most obvious **aleatory uncertainty** is the sampling fluctuation. Another group of 800 patients may reveal sexual adverse effects in 4.9 or 5.2 percent. If the sample is random from a specific target population, a statistical confidence interval (CI) can be built around it. This is not possible for a nonrandom sample. In this example, this would be quite narrow since the group size is large. Suppose this is 4.8 to 5.4 percent. But the patients of age 24 years may not have same incidence of side effects as of age 39 years. For this it is necessary that the probability of sexual adverse effect is obtained as a function of the antecedents – in this case age of the patients, general health condition, and extent of hair loss. When these are varied, different CIs would be obtained. Only those values of the antecedents can be considered that have been adequately observed. The new limits so obtained for sexual adverse effects may be from 4.1 to 6.0 percent. When the estimates of the parameters (such as coefficients) of functional form of the relationship are varied within the plausible limits as provided by the sampling error around them, the minimum of the lower limits of such CIs may be 3.8 and the maximum of the upper limits 6.3 per cent. This assumes that the antecedents and the outcomes have been exactly measured with no error. Among the antecedents, only age could be considered error-free. Assessment of general health and of extent of hair loss can be rarely exact.

This is true for the outcome also that is assessed in this example by decreased libido, erectile dysfunction, and decreased volume of ejaculate. Assume for the moment that there are no biases but minor random errors in such assessment can never be ruled out. When these are considered, the limits for incidence of sexual adverse effect could become 3.7 to 6.4 per cent. This is the uncertainty interval that a clinician should work with. Note that this interval has been built with a certain confidence level, say 95%—thus this uncertainty in any case is inherent. The interval can be narrowed down for practical purposes for a patient whose age, level of health, and extent of hair loss is exactly known. Then the variation in these antecedents need not be considered.

While presenting the uncertainty interval in the preceding paragraph, we have assumed that there is no dropout, and no patient took any medication during the follow-up that could have altered the sexual functioning. If yes, that will introduce bias. It also assumes that valid methods were used to ascribe side effects to the drug. Further assumption is that no case is missed or misdiagnosed. Note that the effect could be possibly, probably, or definitely drug related. This distinction may be important to explain the consequences to the patient. Uncertainty interval also assumes that the sexual dysfunction could be affected only by age, level of health, and extent of hair loss and nothing else. Or, at least the effect of other factors averages out. This obviously won't be true for any individual patient. Perhaps the factors such as heredity, work stress, conjugal harmony, comorbidities, and diet should be also examined. There might be other antecedent factors that are not known yet – can it be pre-existing testosterone level? The example also assumes that the antecedents are exactly measured. This may not be true for a variable such as level of health or extent of hair loss. Inter-observer variation in their assessment can be high. There might be an element of clinician's bias in assessing such variables. This is true for the outcome variables also – in this case for decreased libido and erectile dysfunction. The clinician may not be fully able to explain such outcome to the patient and the patients would evaluate them in their own subjective way. Validated tool may not be available to measure these outcomes. There might be error in measuring the volume of ejaculate before and after the treatment. It may not be clear what sort of decrease is to be considered as a real decrease since some variation is natural. Some patients may have missed intake of drug once in a while. All these form part of the **epistemic uncertainties**.

Further assumptions in this example are that the sexual adverse effect can be measured by decreased libido, erectile dysfunction, and decreased volume of ejaculate, and there is no need to assess anything else such as hormone level or breast tenderness. Other type of sexual adverse effects may occur that are not visualised but can emerge later as the science progresses. An important source of epistemic uncertainty in this example is the nature of the relationship between the antecedent and the outcome. Biological model is not available. A statistical model based on logistic regression is almost invariably worked out by assuming linearity. This may or may not be true. If quadratic or any other functional form is assumed the results could be very different. Sensitivity analysis is done to find the effect of all such variations. In addition, the model needs external validation.

This trial has gone on for only one year in this example. Thus nothing could be said about long- term effects. If the age of the patients varied from 20 to 39 years, the findings may not be indicative of what happens to a new patient of age 48 years. If the general health of most patients included in the study is good, it is not easy to

extrapolate the findings to a patient whose health is very poor. In him the chances of adverse sexual effects could be very high—or very low (who knows!).

Extrapolation of the results requires that the subjects included in the trial are truly representative of the target population and the new patients are from this target population. Also that there is no dropout, or else the dropout effect is properly adjusted. Possibility of bias in the sample introduces another component to the epistemic uncertainty. The method of analysis of data and their interpretation should be complete and free from bias. In this particular example, the possibilities of these positives are bright but the situation may not so nice in other setups.

Basic message from Examples 1 and 2 is that the uncertainty around an estimate is much more than what is made out by the conventional statistical confidence interval. Consideration of aleatory uncertainties may provide an enormously large uncertainty interval, and epistemic uncertainties put a further question mark on the validity of this interval. Many of such uncertainties go unnoticed and uncared for, leading to unexpected results in many cases.