Models in Epidemiology and Biostatistics
Gordon Hilton Fick

Terminology For These Sessions

In the health care literature, there are many terms, namings and notation differences and sometimes there are complicated differences depending on the textbook, monograph, journal, website or blog. It seems unlikely that there will be a consensus established any time soon.
It then seems best, then, to just make clear choices and to note, when appropriate, when there are duplicates or alternates elsewhere in the literature.

Cause and Effect

Many writers speak of cause and effect. We have Bradford's Hill's set of nine criteria for causality. Debates rage on about them. Many authors argue that causality is a largely theoretical matter and that, in the real world, causality is largely incomplete, speculative or inappropriate. Nevertheless, there is a research area called Causal Inference. There are some very strong proponents for this research. There are many nay sayers as well. The book by David Freedman : 'Statistical Models and Causal Inference' is an important contribution.
I have made the decision to avoid speaking of cause and effect. I will not use either word. I will speak of association. So, for example, the risk difference, the risk ratio and the odds ratio are measures of association [and they will not be called 'effect' measures]

Confounding

Confounding can have some quite elaborate variations. For many circumstances, we will have something(s) that can be viewed as adjusted and something that can be viewed as crude. How crude and adjusted are compared can lead to statements of the apparent presence of confounding. There are circumstances where we have an incomplete adjustment and then we have a more complete adjustment. This will lead to statements that may make a confounding statement appropriate.
We will see that statements about the presence or absence of confounding can require a number caveats. Confounding detection is a function of the measure of association. For example, it may be that a confounding observation is detected and is appropriate when the measure of association is a risk ratio but such an observation might be irrelevant when the measure is an odds ratio.
If the assessment of confounding is in any way based on the use of an adjusted measure, one must determine if the adjustment is based on a weighted average of measures that are deemed to be 'close' together. Averaging things that differ by 'more than error' is meaningless. Such an average refers to no one. Inevitably, this leads to the need for an assessment of modification. So modification detection precludes confounding assessment unless one is prepared to assume the meaningful absence of such modification.

Modification

One sees the term modification on its own and with many different adjectives. There is [just] modification and there is measure modification, effect modification, effect measure modification, association measure modification. In part, these adjectives are noting that modification detection is a function of the measure of association. One rarely hears adjectives of these types used with confounding even though confounding is function of the measure as well.
One also sees the term interaction used sometimes as a synonym for modification. It is perhaps best to keep the use of the term interaction separate from the use of the term modification although it is true that sometimes the distinction is far from clear, in context.
I have made the decision to use the term 'modification' without any of the adjectives mentioned above.

Just as there are elaborate variations of confounding, so too there are elaborate variations of modification. Typically, when the elaborate forms are in play, both confounding and modification require quite detailed discussion and interpretation. Some authors try to use brief phrases involving 'double', 'triple' or 'joint' but such descriptions are typically incomplete, oversimplified or misleading. For example, when an author writes that there is 'joint confounding' one must inevitably ask which form of joint confounding is detected. When there is more than one potential confounder/modifier, one can have that an identification of modification by one variable leads to confounding found for another variable and vice versa.

Interaction

In health research, we might have an interaction of two exposures. The identification of this interaction may enabled by the additional detection that such interaction is seen only after correctly seeing confounding. For example, it may be that this interaction is seen by adjusting and is not seen without adjustment. Here then, interaction detection requires confounding assessment. One can also have interaction only seen by a modification analysis.

Certain writers borrow the terms main effect and interaction effect from the analysis of variance and apply these terms incorrectly to regression models. It is only with a careful assessment and interpretation of the relevant regression coefficients that one can see the faulty application of 'main'; for example.

Sometimes one sees phrases like : 'the interaction of exposure and confounder'. A very confused state, indeed.

Estimation

The distinction between the population characteristics and the estimates of them can be very blurry in certain parts of the health research literature. To avoid confusion, one must use different symbols for population and estimate. Most of time, population parameters are symbolized with Greek letters [but not always]. One can present the estimates with hats over the Greek letters or use the Latin equivalents. For example, the letter beta [usually with a subscript] used for the population characteristic, then one uses that beta with a hat for the estimate or one uses the letter b for the estimate. Writing the letter beta [alone without a hat] for the estimate is confusing. Which is which?

Sometimes, we have a population characteristic that has no widely used Greek symbol. Then one must use a hat for the estimate. For example, the odds ratio OR is a function of 2 population parameters and so the odds ratio is a population characteristic. The estimate of OR needs a hat like : $\hat{OR}$ to clearly separate this number from its population version.

Some authors use the Greek letter π for a probability and then the Latin equivalent p for the estimate of π . I decided [a long time ago] to use p for a probability and then $\hat{p}$ for the estimate of p .

Some authors use all upper case letters for a probability and then mainly lower case letters for the estimate. For example : PREVALENCE for the probability and Prevalence for the estimate.

Prediction

The distinction between estimation and prediction can be blurry in the literature too. We will presume that a prediction is a subset of the possible outcomes. With a dichotomous outcome, and Logistic Regression, for example, one can estimate a log odds, an odds or a probability. The prediction of an outcome for an individual is either 1 [for the presence of the outcome] or 0 [for the absence of the outcome]. A prediction would not be a probability. One can estimate a probability and one can predict an outcome.

One now sees 'risk prediction' in the literature. With 'risk prediction', one is, in fact, predicting an

outcome or one is estimating a risk.
Some authors separate classification from prediction. Sometimes, classification is presented in groups based on the predictions or the estimates. One sees classification rules; usually determined by being above or below a threshold.

Models, Regression Coefficients and Terms
In these sessions, we are modelling health outcomes and the models are all 'Generalized Linear Models' [GLMs]. These models contain equations made up entirely of population characteristics. The equation is 'left hand side = right hand side'.
The left hand side is called the link and it is a function of the response variable. The right hand side is a function of the explanatory variables
This function of the explanatory variables is said to be linear in the sense that it is a sum of terms and each term is the product of a regression coefficient and a function of 'explanatory variables'. The explanatory variables are exposures, confounders [sometimes], modifiers [sometimes], products of expressions, other functions of expressions and more.
We will see three link functions, the logit link, the log link and the identity link.
In the health literature, the response variable is more commonly called the 'health outcome' [or just the outcome].
Once the data is at hand, we can construct a fit. We must keep clear the distinction between the model and the fit. The model is the population and the fit is the estimate(s).
With the fit, the estimates of the regression coefficients determine the 'fitted values'. We can have fitted values for each observed outcome and, in addition, we can construct fitted values for any set of the explanatory variables.
Most of the time, we will use the Greek letter, beta, with subscripts to denote the regression coefficients. The estimates of these regression coefficients will be symbolized with hats over the corresponding beta with subscript or will be symbolized with the letter 'b' with subscript for the corresponding beta.
Somewhat recently, the expression 'beta coefficient' has become vogue. Unfortunately, it is often unclear whether the 'beta coefficient' is a population characteristic or an estimate. We will not use the term 'beta coefficient' in these sessions. There are other reasons why we will avoid this expression. For example, in some literature, the beta coefficient is a 'standardized' regression coefficient. We will not have use for this standardization. In any case, this naming can be yet another source of confusion.
So the right hand side in a model equation is made up of a sum of terms. The terms are products of regression coefficients with functions of the explanatory variables. Some authors try to attach adjectives to some of the terms. One sees 'confounding term', 'modification term' and sometimes more elaborate expressions like 'joint confounding term'. Unfortunately, these adjectives often do not apply to the specific model under consideration.
For example, if age is being considered as a confounder, the term with age is sometimes called a confounding term. But this is misleading. A determination of confounding here involves a comparison of two models; one with age and one without age. Confounding may or may not be seen.
As another example, one may have a model to assess age as a modifier. Such a model might include terms for the exposure, age and the product of age and exposure. With this model it would be incorrect to call the term involving age alone a confounding term. Indeed, the interpretation of this term is different precisely because of the term including the product of age and exposure.
Use of the expression 'joint confounding term' is often misleading too. A joint analysis [model and fit] would need to be compared with one or more 'one-at-a-time' analyses [models and fits] in order to determine the perhaps complex nature of confounding in play.

Assumed Common

It is important to know when a regression coefficient is interpreted with the phrase 'assumed common' to something. For example, if we are considering whether age may confound, we would compare two models. Model 1 has a regression coefficient that is, say, the log odds ratio assumed common to each of the age groups. In other words, this regression coefficient is the log odds ratio for the young, for the middle aged and for the old. This regression coefficient applies to all three age groups. It is common to all three age groups. Model 2 has a regression coefficient for which age is not considered in its description. This regression coefficient is not assumed common to the age groups.
With Model 1, some authors use phrases like 'holding age fixed' or 'controlling for age'. Both of these phrases are incomplete, miss the crucial issue [assumed common] and require further explanation.

Circles and Arrows

Some authors include abstract visuals with 'circles and arrows'. Sometimes such visuals are supposed to aid in interpretation. Most of the time, [maybe ALL of the time] far more detailed descriptions [than just 'circles and arrows'] are essential to explaining the issues in play. Such descriptions require detailed sentences carefully prepared in context.

Independence

Back in the 1960's, some authors referred to the response variable as the 'dependent variable' in so far as the response was dependent [conditioned on] a collection of explanatory variables. Unfortunately, at this time, some authors then referred to the explanatory variables as the independent variables [because they weren't the dependent variable]. Many statisticians protested the use of 'independent' here and they then presented other namings. [like explanatory] There remains a considerable inertia to this day regarding this naming. It gets worse. The literature is now filled with phrases like 'independent factors' and/or 'independent predictors' and more and more muddle...

It would seem that 'most' of the time, when a researcher refers to the 'independent factors', they usually mean that such factors are included in a model in an additive way (i.e. no interactions). However, there does not seem to be clear guidance on these matters and the cynical reviewer needs to dig deep these days to determine what is actually intended.
To make matters even more confusing and ill-formed, we now have 'Independent Risk Factors'.

Continuity and Linearity

Continuity has a precise mathematical definition. [have a look in your favourite calculus text]
Informally, a continuous variable is one for which, within the limits that the variable ranges over, any value is possible. Age, weight, height and duration of illness are examples of continuous variables.
A 7 point “Likert” variable is not a continuous variable. The number of return visits during a study is not a continuous variable. A variable that is not continuous is called “discrete”.
The adjective “continuous” has crept into constant usage in regression analysis. Often, there is a decision to be made as to whether to use an actual variable as a response variable or to use a version of this variable with two or more levels based on cutoffs/thresholds. This is an issue concerning the left hand side of the model and not the the right hand side.
For the right hand side, the issue is often whether the explanatory variable affects the response in the linear way. If this is a plausible assumption, then such a use of the actual variable may be warranted. If the relationship is not linear, then one option is to set up a set of indicator variables based on sensible thresholds and to then study the nature of the variable-response relationship. Keeping with the actual variable, one can add polynomials in this variable or possibly more elaborate expressions like restricted cubic splines to address non-linearity.
Unfortunately, authors now speak of the use of a 'continuous' variable if the actual values of a variable

are used. The continuity of the variable is in fact irrelevant to the issue at hand. The real issue is the nature of the variable-response relationship. Indeed, it is certainly possible and reasonable that an explanatory variable can clearly have only a discrete set of values and yet nevertheless has at least an approximately linear relationship with the response. Such a variable can, then, with advantage, be included in the model even though the variable is most clearly not 'continuous'. It is far more helpful to refer to the possible linearity of a such variable rather than to merely to say it is 'in the model' as a continuous variable. The continuity or discreteness of a variable is relevant when such a variable is being considered as a response variable however.

Distribution Form

The probability distribution of the response variable is a crucial part of the list of assumptions implicit in a chosen model.

For a dichotomous outcome, one usually requires that the marginal distribution is a binomial distribution [for the sets of outcomes with the same explanatory variables]. If each outcome has a unique set of explanatory variables then this special binomial distribution is called the Bernoulli distribution. For most of these sessions, we require an assumption of statistical independence. The joint distribution of the response variables can then be written as a product of Bernoulli distributions. This crucial assumption is relaxed in the next course.

For an ordinal/nominal outcome, we usually require that the marginal distribution is the multinomial distribution. Statistical independence is still a crucial assumption.

For a count outcome, the marginal distribution is sometimes the Poisson distribution or the Negative Binomial distribution.

For a measured outcome, the classic assumption is that the marginal distribution is Normal or at least approximately Normal. The crucial part to the approximation is the symmetry of the distribution so that the centre of this distribution can be estimated using methods based on averages. If the distribution is clearly asymmetrical [typically with skewness], then one usually needs to use one a vast collection of methods to address this asymmetry. There are many other issues here.

For the methods described in these sessions, the distribution of the explanatory variables is not relevant. For example, we may know that the distribution of some or all of the explanatory variables is skewed. If an explanatory variable is an indicator variable, then the distribution is most certainly not symmetrical [being a Binomial distribution]. If an explanatory variable is a count, its distribution is skewed. If an explanatory variable is measured, it may be skewed. Duration is typically skewed. One may contemplate a transformation of an explanatory variable but the reason to consider such is to address linearity with the response and not, per se, the likely skewness.

In fact, it may be one does not have a meaningful probability distribution for some [or all] of the explanatory variables. For example, with a case-control study, the cases are determined sometimes by availability and then one might have an equal number of controls selected to be representative. There is no real probability distribution for the indicator for cases.

Now we come to the terminology issue here. We often speak of the conditional distribution for the response variable given the explanatory variables. However, there may not be a meaningful joint distribution for all the variables [response and explanatory together] and by implication no meaningful marginal distribution for the explanatory variables. So the term 'conditional' is not meant to be seen in the usual complete probability world. Nevertheless, we do speak of conditional means, conditional log odds and so on. Maybe other names should used here; like, for example, the log odds specific to a set of explanatory variables.

Exact versus Approximate

One might think that biostatisticians and epidemiologists would agree that an exact method is preferred to an approximate method. With 2x2 tables, Fisher made the case for what is now called Fisher's Exact

Test [FET] making it clear that Pearson's Chi-Squared test was an approximation to it. It is true that, with large samples, the two methods give the same result. FET is certainly a 'computationally intensive' method. Years ago, there were 'rules of thumb' for determining when the hard work needed with FET was necessary. These days, FET is computed easily. Other exact tests are becoming possible with advances to computing algorithms and computing speeds. Exact Logistic Regression, Exact Poisson Regression, Exact Binomial tests are now available. Computer memory and computing times can be substantial though.
Student's t test is certainly an exact test as well. The degrees of freedom for t can be important for small samples.
Most investigators would agree that the "exact vs approximate" issue is arguably less important than many other issues with small studies.

Accept / Reject - Significance - p-values
Health research is now filled with p-values and confidence intervals. This not a bad thing, per se. Unfortunately, both p-values and confidence intervals are widely misunderstood. Part of this confusion stems from remnants of the use of decision rules to 'accept' or 'reject'. Notions of 'evidence' date back at least 200 years. For our purposes here, significance can be separated [at least historically, perhaps] by the research led by Ronald Fisher and the research led by Jerzy Neyman and Egon Pearson. Fisher developed the measure of evidence that we now call the p-value and Neyman developed the decision rule approach that led to 'accept or reject'. It is fair to say that the Neyman/Pearson approach dominated all research [including health research] well into the 1960's and in spite of protests from Fisher and others. For a whole complex of reasons, in the 1970's, health researchers began to advance the reporting of p-values. For another set of complicated reasons, health researchers clung to the number 1/20 = 5% as a 'dividing line' and kept notions of 'rejecting'. In a way, while the reporting of p-values caught on, aspects of the decision ethic continued [and to this day].
The name 'p-value' is hard to date with assurance. Into the 1970's, researchers spoke of 'pure significance tests', 'levels of significance', 'observed levels of significance'. The books by Fraser [1976, chapter 7] ; Cox and Hinkley [1974, chapter 3] and others detail some of these names. For example, Fraser spoke of the 'observed level of significance' [OLS] while Cox & Hinkley spoke of the 'level of significance' [p_{ols}] . Some authors call the 'level of significance' the dividing line [like 5%] so this can be confusing. These days, some authors use p-value or OLS interchangeably.
So where did the name 'p-value' come from? In his Statistical Methods For Research Workers [SMRW], Fisher refers [in several places] to : ' the values of P '. Here, though, in one place, for example, he is detailing the values of the probability distribution for χ^2 : the values of χ^2 and the values of P where P = Pr(χ^2 or greater). He does not call an 'observed' probability a 'p-value', rather he just uses the capital letter P.
So when did the name p-value come into common usage ? I am quite unsure about this. Who or what groups started this usage? Dunno.

Confidence - Fiducial
Around the same time that Fisher and Neyman & Pearson were first presenting statistical tests, they were also developing intervals. Neyman & Pearson spoke of 'Confidence' intervals while Fisher spoke of 'Fiducial' intervals. For almost all the methods used in health research, the calculation of these intervals is the same. The interpretations [Confidence versus Fiducial] are different though. Many [most?] health researchers describe their intervals as confidence intervals but they often [incorrectly] interpret them as though they are fiducial intervals.

Bayesian - Frequentist

Bayes Rule is covered in most first courses in Probability. Its most direct application to health research comes with taking sensitivity, specificity and prevalence and then using Bayes rule to present predictive value. See, for example, Haynes, Sackett, Guyatt & Tugwell [2005]. Here, the pre-test probabilities are seen as 'prior' probabilities. Bayes rule then provides post-test probabilities that are often called 'posterior' probabilities. [really!].
Folks calling themselves Bayesians then present considerably more elaborate contexts for the application of Bayes rule. Fisher argued against most [maybe all?] Bayesian methods. In any case, there are now [staunch] Bayesians and there are the rest of us [non-Bayesians [yuck], Frequentists [nah!], part-time Bayesians [worse still?]....]
I would say that most Biostatisticians and most Epidemiologists can appreciate some parts to the Bayesian world. The landscape here is changing constantly though.
Many universities now offer entire courses in Bayesian methods. Nevertheless, papers based on Bayesian methods remain in the minority in Health Research.
There are intervals based on Bayesian methods. They are usually called Credible intervals. Fisher was adamant that Fiducial intervals were not to be seen as Credible intervals. Some authors present what they call Confidence intervals but then [incorrectly] interpret them as though they are Credible intervals. [sigh]

Marginal / Conditional - GEE / Mixed : Random : Fixed
When one is modelling without the assumption of independence, it is best to distinguish between two types of models: those with subject specific components [Conditional] and those without subject specific components [Marginal]. Conditional models can then be seen to be have regression coefficients that are assumed common to subjects and hence the estimates are 'adjusted' for subjects.
It is best not to refer to the model type by the methods used to carry out the fit. Nevertheless, many authors continue to refer to GEE models, REML models, ... It is important to determine whether the intended models are marginal or conditional.
One also sees wide use of terms like 'fixed', 'random' and 'mixed'. Unfortunately, these terms are incomplete descriptors. Some use of these terms appears to be remnants of older software formulations [SAS, in particular]. Several key contributors to this area of research [like Nelder...] are advising the descriptions that start with the adjectives marginal or conditional.

Multivariable Analysis
Some health researchers continue to use the name 'multivariable analysis' when discussing model based methods [where the right hand side of the regression equation has more than 2 terms]. This name is often misunderstood and confused with 'multivariate analysis' [where the outcome {the left hand side} is multivariable]. These days, interesting new research has led to 'joint' models with two or more variables as outcomes. This area is growing rapidly and there are many advances. Notably, there has already been considerable success with joint models where one has a longitudinal variable and a dropout variable as a 'bivariate' outcome.

Rules Of Thumb
Gerald Van Belle has written an interesting book on 'Rules of Thumb'. Before adopting any 'rule of thumb' it is important to understand its origins and merits/demerits. Some of these rules are fading from importance as computationally intensive methods are becoming viable. For example, one still sees the application of obsolete rules with certain approximate methods even though the corresponding exact method is available.

Very Small Studies : N of 1?

Now that 'Individualized Medicine' and 'Precision Medicine' have captured so much attention, how do Biostatisticians and Epidemiologists respond to the 'N of 1 trials' seemingly front and centre? Perhaps the many aspects of 'Exploratory Data Analysis' as developed by John Tukey and others could be the first direction. There may be sensible aspects to be taken from Time Series Analysis. In both areas, considerable caution and qualification will be essential.
There are attempts to borrow notions from the Analysis of Longitudinal Studies and from the Analysis of Cross Over Studies. Such attempts may aid in 'interpretation'.
There is an 'N of 1' literature from Education and Psychology that has quite a history. There is a book by Todman and Dugard [2001, 2011] entirely devoted to these matters.
Perhaps the most famous [maybe 'notorious'] 'N of 1' is the 'Lady Tasting Tea'. One person [the lady] is tested. Fisher [in the Design of Experiments] provides a test of significance. Historians of science have been debating Fisher's intentions here. Did he really mean to present this test for an 'N of 1'? or is the context merely apocryphal.

Power - Confidence Interval Width
It is certainly true that accept/reject is fading from view in health research. Nevertheless, power calculations remain dominant in sample size determinations. Power is based around the decision rule 'accept/reject'. Since confidence intervals now have gained considerable usage, I now argue that sample size determinations should be based on confidence interval width. While there is a certain equivalence between power and confidence interval width, it can be illuminating to explore the implications of sample sizes on the confidence intervals that might result. I think that sample size matters need to be based on very simple analyses since [even when one starts simple] the number of unknowns can be quite demoralizing.

Quantitative - Qualitative
Many researchers using qualitative methods refer to all methods that they see as non-qualitative as quantitative. An obvious clumping together of many very different methods. No one outside of the Qualitative world calls such methods Quantitative.
Qualitative 'Inquiry' ranges from valuable, academic and enriched to very suspicious and simplistic, and, in some cases, to absolute nonsense.
It appears that a high proportion of 'Qualitative' health researchers have almost zero understanding of Epidemiology or Biostatistics. This is a very serious matter. It seems to me that the interactions between Qualitative people and anyone else are becoming less and less viable. With certain Qualitative people, it is an aversion to anything involving the simplest of Math. One sees this aversion among some Epidemiology folk as well. Not a valid or promising circumstance.
In the Qualitative world one sees so-called Mixed methods. Most of this literature is trivial at best and troubling as well. There is an odd simplistic bundling of all things supposedly not Qualitative usually involving a highly superficial use [and understanding] of elementary statistics and an even more primitive understanding of elementary epidemiology.
Nevertheless, over my career, I have seen a number of worthy MSc theses using decently applied Qualitative approaches. The decent ones involved an enormous amount of time and effort to do things in a sensible way. Further, the stronger students did have at least an introductory understanding of modification and confounding and so would avoid the obvious fallacies seen in so much Qualitative nonsense.

Abbreviations
I do not abbreviate The Analysis of Variance, The Analysis of Covariance ....
So...which abbreviations are OK and which ones are not OK? I prefer to avoid abbreviations and acronyms almost all the time.

Hierarchically Well Formulated

Some authors say that all models must be hierarchically well formulated. This is not true. The correct reasoning is elaborate and based on centring matters.

Prospective - Retrospective

Retrospective means looking back, looking back on, contemplating, or directed to the past, looking or directed backward, contemplative of past situations, events, ... looking or directed backward.
Prospective means expected or expecting to be something particular in the future; likely to happen at a future date; concerned with or applying to the future; relating to or effective in the future; likely to come about... likely to be or become.
One sometimes sees the name : 'retrospective cohort study' even though one is looking forward. With this type of study, one sets the clock back in time usually to when exposure was determined but then one views the disease [in the future] as the outcome. Perhaps the better name here is 'historical cohort study'.
A case-control study is retrospective. Case-control status is determined [usually in present time]. Then ones looks back in time for the exposure status. Here, exposure is the outcome and case-control status is explanatory.

Mediation, moderation

There is a complex and demanding literature concerning intermediates [aka mediators]. One needs to be cautious with certain oversimplified methods. In particular, one sees the 'product method' mainly in the social science literature. The 'product method' is not recommended.
Some authors refer to moderation. Moderators are often not ideally formulated. In some contexts, moderation is a synonym for modification. One needs to be cautious when reading about the assessment of moderation. Often, such assessments are incomplete.