Using logistic regression in perinatal epidemiology: an introduction for clinical researchers. Part 1: basic concepts

R. Brand and M.J.N.C. Keirse

Pediatric and Perinatal Epidemiology 1990; **4:** 22-38.

Department of Medical Statistics, Medical Faculty and Department of Obstetrics, Academic Hospital, State University, Leiden, The Netherlands

Summary. Logistic regression *is* a statistical modelling technique which may be applied to estimate the simultaneous effect of a set of predictors (e.g. gestational age. birthweight) on the risk of a certain outcome variable (e.g. neonatal death) which can take either one of two possible values (yes/no, alive/dead) or in the situation where one wants to estimate the effect of a particular risk factor (e.g. sex) while adjusting (correcting) for the effect of other risk factors (e.g. gestational age). Since this situation often occurs both in medical or epidemiological research and in daily practice it is important to have a flexible and readily interpretable technique to predict risk of mortality and morbidity. Since the logistic regression technique is a powerful and widely applicable tool which is appearing more and more often in the epidemiological literature, a basic understanding of this technique becomes necessary for the clinical researcher. In this paper we explain logistic regression to medical researchers who do not have any particular statistical background. Part 1 covers the basic concepts. Part 2 will describe the actual representation of the basic concepts in a logistic framework.

1.0 Introduction

Logistic regression can be applied to study phenomena that can be expressed categorically as 'yes/no' situations (e.g. dead or alive; preterm or not). Clinicians often want to know the likelihood that a particular outcome will occur when one or more predictors of this outcome are present or absent. Ideally, they also like to know what weight can be given to a particular predictor even if its effect is confounded by the presence of other factors that influence outcome. It is in such situations that logistic regression is especially useful in that it allows quantification of the effect of a particular risk factor adjusting for the confounding effects of other variables.

In this paper logistic regression will not be introduced in the usual way of presenting a general theory illustrated with examples. Rather, the main part of this paper is the example itself: a simple, hypothetical cohort for which appropriate cross-tabulations are made to introduce and illustrate the basic concept of logistic regression. The numbers in the example are chosen so that they facilitate the clarification of the statistical arguments and display some sense of reality despite their fictitious character. When appropriate, more general formulations of the conclusions reached in the example are stated separately (these general formulations may be skipped until the end, when the entire example should have familiarised the subject and provided a basis for generalisation).

In section 1.1 the small fictitious cohort study is introduced and some characteristics are fixed for the rest of the paper. Section 1.2 covers the concepts of odds and odds ratio. In sections 1.3 and 1.4 confounding and interaction are explained. The way in which they are presented leads inevitably and naturally to the concept of logistic regression. The reader will become aware that what he is viewing as 'common sense' in dealing with the problem of confounding (or adjusting) is exactly the way these problems are tackled in the framework of logistic regression.

Having established this 'link', the logistic equation itself is introduced in Part 2, the mathematics of which are in this paper confined to addition, multiplication and taking logarithms.

1.1 Description of the example population

First, we consider a cohort of 400 infants that will form the basis for the remainder of this paper. We assume that the cohort has been obtained by selecting *all liveborn infants with a birthweight less than 1500 g (very low birthweight infants) in a particular population and during a particular time period.* In this cohort we study three variables: **'neonatal mortality'**, as the outcome variable of interest and **'infant sex'** and **'gestational age'** as predictors of this outcome. We will later generalise our example and consider other potential predictors as well. The three variables involved thus far can take the following values:

Neonatal mortality: alive, dead

Infant sex: girl, boy

Gestational age: 24, 25, ..., 29 completed weeks

Our cohort can now be fitted into appropriate cross-tabulations with frequencies that will *remain the same* throughout the example, even when later on we vary the mortality risks to illustrate the notions of odds, odds ratio, confounding and interactions (Figure 1).

	Gestational age in completed weeks						
Sex	24	25	26	27	28	29	Total
girls	9	25	42	60	39	25	200
boys	18	45	60	40	27	10	200
Total	27	70	102	100	66	35	400

Figure 1.

The actual numbers have been chosen to facilitate subsequent calculations. It can be seen in Figure I that the distribution of gestational age in boys compared with girls is shifted to the left by I week. Although the effect has been exaggerated here, this reflects the 'real life' situation in which at the same gestational age boys tend to be heavier than girls. A selection criterion based on birthweight alone will lead, and has indeed led,1 to an over-representation of boys in the lower gestational age groups. This phenomenon will play an important role in our discussion of the concept of confounding.

1.2 Odds and odds ratio

The main purpose of our hypothetical cohort study is to examine neonatal mortality in this group of infants. Neonatal mortality is a perfect example of a 'dichotomous variable' or 'dichotomy'; it can assume only two values, 'dead' or 'alive'.

Suppose that the distribution of neonatal mortality in our cohort is as shown in Figure 2. The likelihood that any individual infant in our cohort will die can be expressed in two ways: as a **risk** (probability) of dying or as the **odds** of dying. The **risk** or **probability** is defined as 227/400=0.57 or 57% which simply means that 227 out of 400 or, equivalently, 57 *out of* every 100 infants will die. It should be noted that the word **'probability'** has two meanings: a general meaning ('chance') and a more specific meaning (synonymous for 'risk').

One can also express the chances of dying as an **odds.** In our example this quantity is 227:173= 1.31; it indicates that for 227 infants who die there are 173 infants who survive. The chance that an infant will die is

227 against 173, that is 1.31 *against 1*. More generally speaking, the mortality odds are defined as the 'risk' of dying **divided** by the 'risk' of staying alive (in this case, 227/400 divided by 173/400).

Mortality	
alive	173
dead	227
Total	400
Risk	57%
Odds	1.31

227/400 = 0.57
227/173 = 1.31= 0.57/(1-0.57) - 0.57/0.43

Figure 2.

$\text{Odds} = \frac{\text{Probability}}{1 - \text{Probability}}$	$\text{Probability} = \frac{\text{Odds}}{1 + \text{Odds}}$ Probability = Risk

Figure 3.

Presently, the choice between risk and odds is a matter of personal preference of the investigator. It is useful to know that odds can be converted to risk and vice versa using the formula shown in Figure 3. Thus an odds of 1.31, for example, can be calculated to provide a risk of 1.31/(1 + 1.31)=0.57. Similarly a risk of 0.57 will yield an odds of 0.57/(1 - 0.57)=1.31. Risk is not a better measure for 'probability' than odds. Both are merely different measures to express the same thing, in the same way as degrees Celsius and degrees Fahrenheit are both measures of temperature.

In its most general form, the clinical questions we shall address are the following: **What is the association, if any, between gender and mortality in this cohort? How strong is it? Should we consider gestational age in this context?** To do so we need to consider the chance of neonatal death separately for both boys and girls. Again this is expressed in a cross-tabulation (Figure 4).

In figure 4 we have assumed for the time being that the **'crude'** odds of dying are 1.08 (= 104/96) for girls and 1.60 (123/77) for boys. The term 'crude' refers to the fact that we simply calculated the odds across the table and ignored the distribution of other factors, such as gestational age. In our cohort boys thus have a higher overall probability of dying than girls.

	Sex		
Mortality	girls	boys	Total
alive	96	77	173
dead	104	123	227
Odds	1.08	1.60	1.31

Figure 4.

A measure to express the amount by which boys and girls differ in their chances of dying is the **odds ratio.** The odds ratio of neonatal mortality for boys versus girls is defined as the ratio between the two odds: 1.60/1.08=1.47. The odds ratio tells us how the chances for boys compare with those for girls when both are

expressed as odds. The odds ratio can also be computed in a familiar way as a 'cross product': (96 x 123)/(104 x 77)= 1.47.

Another measure to compare the chances of dying in a cohort study is the **risk ratio** or **relative risk,** which can be defined as the ratio between the two risks: (123/200):(104/200)= 1.25.

Note that for rare outcomes (occurring in less than 10% of the infants) the risk is nearly equal to the odds. It follows that the risk ratio is also nearly identical to the odds ratio. This is due to the fact that the odds is defined as

risk: (1 - risk)

and (1—risk) approaches 1 when the risk is low. Hence the distinction between odds and risk is blurred for small risks. As a numerical example: suppose the risk for boys to be 8% and for girls 6%. The odds for boys become 0.08:0.92=0.087 and for girls 0.06:0.94=0.064. This yields an odds ratio of 0.087:0.064=1.36 while the risk ratio is 0.08:0.06=1.33.

Also note that an odds ratio of 1 implies that the odds are identical; the risks will thus also be identical.

In the remainder of this paper we will describe **differences in risk or probability** only in terms of **odds ratios.** As will become obvious later, this is because the odds ratio (sometimes denoted by **OR**) is the natural choice when the logistic model is used. The parameters in the regression model then receive a direct clinical meaning in terms of odds ratios. Interpretation of the OR is as straightforward and as easy as the interpretation of the **RR** (relative risk), but it has a number of desirable mathematical properties to facilitate statistical significance testing.

It is important to note that in a **case control** study design the population risk ratio **cannot** be computed from such a 2x2 table - it is not estimable - but the population odds ratio **can** be estimated, again as the cross product. This stems from the observation that in a case control study the number of cases and controls is fixed in advance and the number, of for example, boys and girls, is observed afterwards. Hence the **risk ratio** for a boy compared to a girl is dependent on the number of controls selected; on the other hand the **odds ratio** is always the same, no matter what ratio between cases and controls has been used. We will not pursue this subject any further but the reader is cautioned to bear this limitation of the risk ratio in mind.

Use of the odds ratio need not be confined to dichotomous variables such as gender. It is also meaningful in the case of continuous variables such as gestational age. In reality, this can be translated as follows: assign a code of 0 (zero) to the variable sex in the case of a girl and a code of I (one) in the case of a boy. The odds ratio then measures the increase or decrease in odds (on a multiplicative scale) when the predictor variable shifts one **unit,** from 'girl' to 'boy' (from 0 to 1). This may sound artificial, but it suddenly becomes logical when we consider gestational age.

	Gestational age in completed weeks						
Mortality	24	25	26	27	28	29	Total
alive	3	14	34	50	44	28	173
dead	24	56	66	50	22	7	227
Odds	8	4	2	1	1/2	1/4	
Odds ratio	1/2	1/2	1/2	1/2	1/2		

Figure 5.

Sex	'0'	'1'	
Mortality	girls	boys	Total
alive	96	77	173
dead	104	123	227
Odds	1.08	1.60	1.31
Odds ratio	1.47		

Figure 6.

For example, Figure 5 describes the association between gestational age and neonatal mortality. For each gestational age category there is a separate calculation of the mortality odds; these odds decrease with increasing gestational age as would be expected. How much do the odds increase when we increase gestational age by one unit, i.e. l week? In our example (Figure 5) the odds always decrease by a factor of two (from 8 to 4, from 4 to 2, from 2 to 1, etc.). Hence the **ratio** of two odds in two adjacent categories is 1/2 and thus constant over all categories in this simple cohort. We thus have a quantity which measures the decrease in odds when the predictor increases by one unit: this quantity is the **odds ratio**.

As Figure 6 illustrates, this situation is similar to that in which gender is considered. It is important to realise that the situation is identical irrespective of whether we consider a **continuous** variable, such as gestational age, or a **discrete** one, such as gender.

With the discrete variable the odds ratio measures the ratio of the odds in two distinct categories: it indicates *by what factor the odds in the first group should be multiplied to obtain the odds in the second group* if we were to 'move' an individual case from the first to the second group.

In exactly the same way for the continuous variable, the odds ratio is the *multiplication factor* for the odds when we 'move' from one group to another *provided that these groups differ* **one** *unit on the scale used to measure the continuous variable.*

Of course the ratio *need not* be constant over the different levels of the predictor variable. In practice, there will usually not be a 'typical' odds ratio that applies to all levels of the predictor variable; often the odds ratio will vary when we move among different levels. This will be discussed in more detail later.

The analogy between the cases of discrete and continuous variables exists simply because we can assign arbitrary 'codings' to the two categories in the discrete case and because 'O' and '1' can be chosen to represent these categories, without loss of generality.

In Figure 7 the definitions from this section are summarised.

Discrete variables comparing group 2 (G2) with group 1 (G1):

$$\text{Odds ratio G2 vs. G1} = \frac{\text{Odds in group 2}}{\text{Odds in group 1}}$$

$$\text{Risk ratio G2 vs. G1} = \frac{\text{Risk in group 2}}{\text{Risk in group 1}}$$

Continuous variables (assuming a constant odds ratio):

$$\text{Odds ratio per unit} = \frac{\text{Odds at level L + 1}}{\text{Odds at level L}}$$

Figure 7.

1.3 Confounding and risk factors

In this section we will introduce the notion of confounding which is essential to a proper understanding of logistic regression analysis. We will analyse the relation between neonatal mortality and gestational age and the influence of gender on our conclusions. Each step is illustrated by a cross-tabulation (T) and a diagram (D) corresponding to that tabulation. The diagrams are logistic regression models, based on the data displayed in the cross-tabulations.

First we consider the association between neonatal mortality and gestational age in our cohort, shown in Figure 8T.

	Gestational age in completed weeks						
Mortality	24	25	26	27	28	29	Total
alive	3	14	34	50	44	28	173
dead	24	56	68	50	22	7	227
Odds	8	4	2	1	0.5	0.25	1.31

Figure 8T.

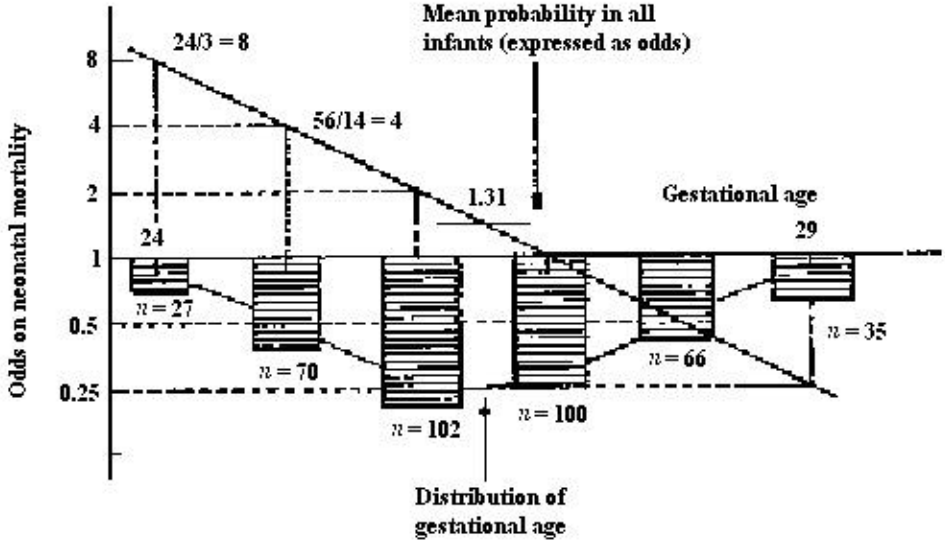

Figure 8D.

In the accompanying diagram (Figure 8D) the relation between the odds of mortality and gestational age is depicted as a regression line: gestational age is shown as the independent and continuous variable in the range between 24 and 29 weeks; the vertical axis corresponds to the associated odds on a multiplicative scale (as discussed later, the multiplicative scale is chosen because the relationship then provides a straight line in the picture).

The straight line in Figure 8D fits the data perfectly (the line connects the points referring to the odds 8, 4, 2, 1, 0.5 and 0.25 belonging to the gestational ages 24, 25, 26, 27, 28 and 29 respectively) and one readily observes the decrease in mortality with increasing gestational age.

While looking at figures like diagram 8D one should bear in mind that throughout this paper we *assume* the relationship depicted to be linear (i.e. representable in a fair way by a straight line) and on the basis of that assumption we draw the straight line which fits the data best. Notwithstanding this, the concepts and arguments which follow can be generalised to the situation where a linear relationship is not assumed, and where the variable is discrete rather than continuous. For clarity here we only describe the linear case, but non-linearity will be considered in Part 2.

The distribution of gestational age within this population is shown in this diagram as a normal curve and by the histograms. We choose a normal curve for illustrative purposes only. Of course, this distribution would not be symmetrical in practice, but it is a plausible description for a cohort of infants weighing less than 1500 8.

The computed overall odds of mortality (1.31) is indicated by a small horizontal line. It is evident that this diagram conveys more information than the simple overall odds, because it depicts the *dependence* on *gestational age* of the chances of dying.

Next we want to study the relationship between gender and neonatal mortality. We could, of course, estimate the chances of dying in both boys and girls in the same way as we estimated the overall chances of dying. This would mean computing the odds in both groups as the ratio between the number of infants who died and the number alive.

However, *does the comparison of the overall odds for boys and for girls inform us as to whether gender in itself is associated with neonatal mortality?*

To answer this question, we consider the cross-tabulation which completely describes the (multivariate) relation of gender, gestational age and mortality in our cohort (Figure 9T).

Two important conclusions can be reached from this crowded table. First, within each **separate** gestational age category the odds for the boys and the girls are exactly the same (the odds ratio is I in each category). Second, the overall odds of mortality in the right part of the table shows the boys to be at a substantially higher risk, the odds ratio being 1.47 (123/77:104/96 = 1.60/1.08 = 1.47). This illustrates the **confounding effect** of gestational age on the **relation** between gender and mortality. Because of the association between gestational age and mortality and the differences in the distribution of gestational age between boys and girls, the overall odds ratio for sex is **biased**. Because *in general* boys tend to have a higher birthweight than girls, they have a lower gestational age than girls of the same weight. In a birthweight defined cohort like ours, this results in a higher overall mortality. However, if we *compare like with like* boys and girls of the same gestational age are at the same level of risk. The corresponding diagram (Figure 9D) visualises this situation in a very simple way.

	Gestational age in completed weeks, broken down by sex																				
	24			25			26			27			28			29					
Mort-ality	G	B	T	G	B	T	G	B	T	G	B	T	G	B	T	G	B	T	G	B	T
alive	1	2	3	5	9	14	14	20	34	30	20	50	26	18	44	20	8	28	96	77	173
dead	8	16	24	20	36	56	28	40	68	30	20	50	13	9	22	5	2	7	104	123	227
Odds	8	8	8	4	4	4	2	2	2	1	1	1	1/2	1/2	1/2	1/4	1/4	1/4	1.08	1.60	1.31
OR (B:G)	1.0			1.0			1.0			1.0			1.0			1.0			1.47		

Figure 9T.

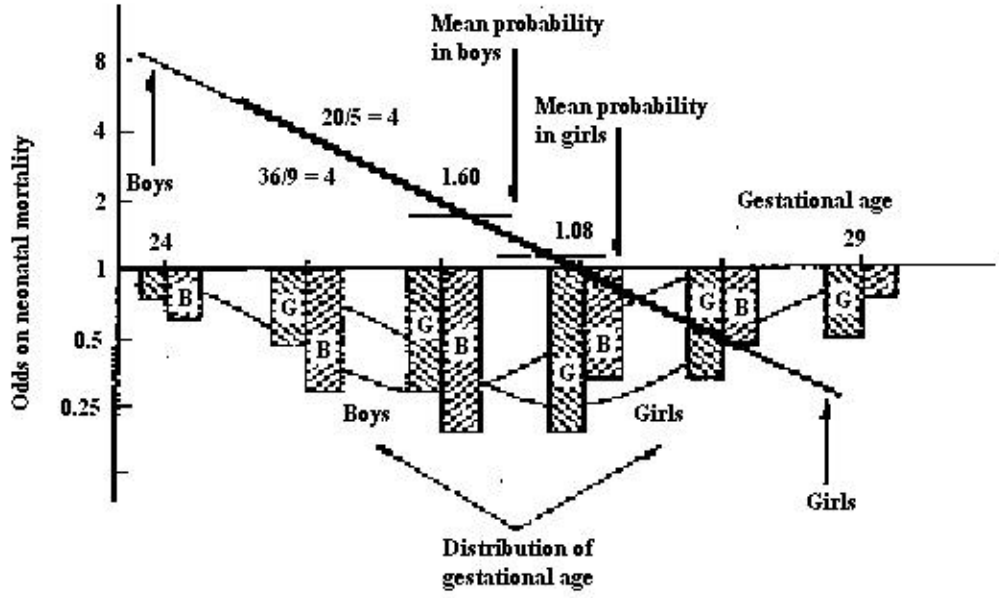

Figure 9D.

The regression lines (representing the **relation** between gestational age and mortality for each group, not just **individual data**) coincide and the difference between the overall odds in each group is only caused by the shift in the gestational age distribution.

Hence, the **relevant** question pertaining to the effect of 'sex' per se is not the difference between the two mean probabilities, but **the distance between the two lines** (which in this case is zero). This is the crux of any regression approach. When studying a factor such as gender in a cohort, we do not just ask whether more boys than girls are dying in our study population. What we really want to know is how much boys and girls differ in their chances of dying if all other known determinants of mortality are the same. In other words, the information that is needed is the level of difference between boys and girls **after adjustments** have been made for various other determinants.

Suppose now that boys are at lower risk of dying than girls of the same gestational age. Even then the overall odds for boys and girls could falsely create the opposite impression. This point is illustrated in Figure 10T and its accompanying diagram 10D.

	Gestational age in completed weeks, broken down by sex														
	24		25		26		27		28		29				
Mortality	G	B	G	B	G	B	G	B	G	B	G	B	G	B	T
alive	1	3	5	11	14	24	30	23	26	20	20	8	96	89	185
dead	8	15	20	34	28	36	30	17	13	7	5	2	104	111	215
Odds	8	5	4	3.1	2	1.5	1	0.74	0.5	0.35	0.25	0.25	1.08	1.25	1.16
OR (B:G)	0.63		0.77		0.75		0.74		0.70		1.0		1.15		

Figure 10T.

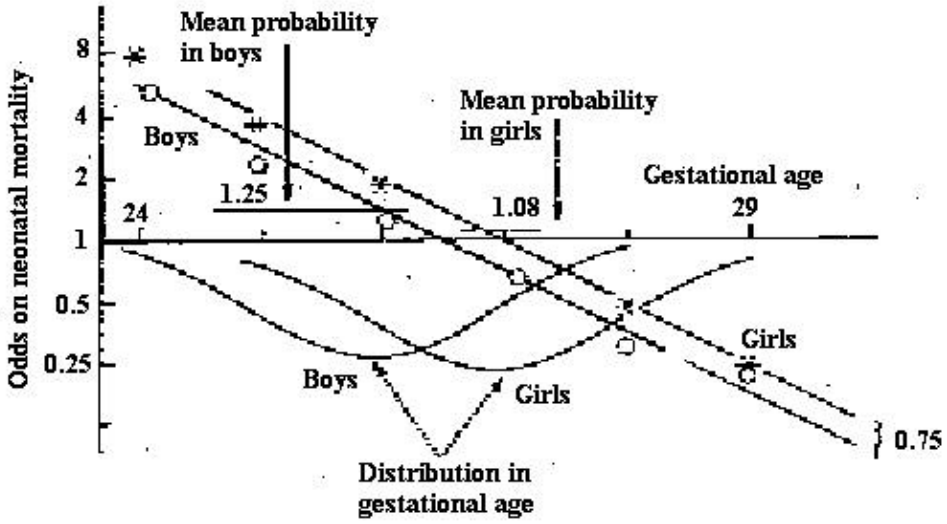

Figure 10D.

As shown in Figure 10T, the odds ratio differs little across the gestational age categories. Because of the low numbers in the extreme categories, it is not possible to choose the data so that the odds ratio remains identical; but this would of course not happen in practice either. Again, the 'regression approach' to visualise the dependency of the odds on sex and gestational age fits the data well (Figure 10D). The lines are again chosen in

such a way as to best fit the series of points determined by the odds of boys and girls in each gestational age category.

Whatever gestational age category we look at, boys are at a lower risk than girls (Figure 10D). Because of the parallelism of the regression lines, it is clear that the ratio of their odds is constant over all gestational age categories (see also the table in Figure 10T) and equal to an estimated 0.75 in this case. This 0.75 estimate results from the 'fitting' of the two lines in Figure 10D to the data in Figure 10T; it is also displayed in Figure 10D indicating the distance between the lines. The 0.75 estimate itself cannot be computed from the table however; it has been obtained by a logistic regression analysis of the data in Figure 10T.

One should remember that in this instance (and of course in practice) the lines do not fit the data perfectly. Nevertheless, 0.75 here is a reasonable estimate of the difference between the sexes, obtained by the regression procedure as a kind of 'mean' over the gestational ages.

Thus there is something like **'the'** odds ratio for boys versus girls, independent of their gestational age. This quantity is in effect the distance between the two regression lines (on a multiplicative scale); it is certainly **not** the ratio between the overall odds (1.25:1.08 = 1.15), which is highly confounded by the effect of gestational age.

The ratio between the overall odds is commonly called the **unadjusted** or **crude** odds ratio because it does not take into account the possible confounding effects of other risk factors. If an odds ratio has been computed taking into account at least one possible confounder, it is often denoted as an **adjusted** odds ratio.

To obtain valid information on the relationship *between* a *specific risk factor and a dichotomous outcome variable, we should correct for the distributions* of *other variables that influence the relationship between the risk factor of interest and the outcome.*

A slight diversion may be necessary to make the concepts of **risk factor** and **confounder** more explicit.

Kleinbaum *et al.*[2] defined a risk factor as:

Any variable that the investigator determines to be 'causally related' and antecedent to illness outcome status on the basis of substantive knowledge or theory and/or previous research findings.

These authors also gave a 'working definition' of a confounder:

A confounder is a risk factor for the disease under study whose control in some appropriate way will reduce or completely correct a bias when estimating the (true) exposure-disease relationship.

Hence in the (multivariate) study of risk factors, some (or all) risk factors should be considered as possible confounders when estimating the association between a specific factor and the outcome under consideration. Regarding the use of risk factors as (possible) confounders, Kleinbaum *et* al. cautioned:

A list of risk factors should be restricted to variables that cannot be characterised in causal terms as intervening in the causal pathway between exposure and disease. A pure intervening variable should not be considered as a potential confounder, since its control can spuriously reduce or eliminate any manifestation of a true association between Exposure and Outcome in the population.

The following example may illustrate the importance of these points. A study is undertaken to investigate the relation between gestational age ('exposure') and perinatal mortality (outcome) in a cohort of pre-term infants

(gestational age less than 37 completed weeks}. At least some of these infants will be admitted to a NICU and treated in an incubator.

Suppose the investigator considers *treatment in an incubator* as a risk factor because infants admitted to an incubator do have a higher mortality risk than those who are not treated in such a way. 'If 'treatment in an incubator' would be considered a **confounder** for the gestational age-mortality relationship and the effect of gestational age would be adjusted for 'treatment in an incubator', the association between mortality and gestational age would probably weaken. This is due to the fact that being in an incubator is in no way - as far as we know - a plausible **causal** factor for mortality but simply a **consequence of the exposure** (gestational age) considered. It is very important to see if and when this so called 'incubator-effect is present in your data.

Another example is the *'yellow fingertips*' effect: when measuring the lung cancer-smoking association, adjusting for yellow fingertips causes this (true) association to disappear: again yellow fingertips cannot be considered as a risk factor and hence cannot be taken as a confounder because causality is lacking.

In conclusion, whether or not a risk factor should be considered as a confounder for the exposure-disease relationship is a matter of clinical judgement, not statistical significance. Given a factor which satisfies the definition of a risk factor, if the exposure-disease relationship as estimated by the odds ratio is not appreciably influenced by whether or not we adjust for it, then that risk factor is not a confounder (even if it is a strong risk factor per se). **In other words, it is not the effect on the outcome itself that matters, but the effect on the relationship between outcome and exposure.**

1.4 Interaction

The concept of interaction is introduced by a final cross-tabulation and diagram. The **previous** situation is depicted in Figure 11 (which shows the same situation as Figure 10D).

Suppose now that in our cohort the relation between the three variables under study is as shown in Figure 12T. The corresponding diagram with the fitted regression lines displaying the relation between gestational age and mortality is shown in Figure 12D.

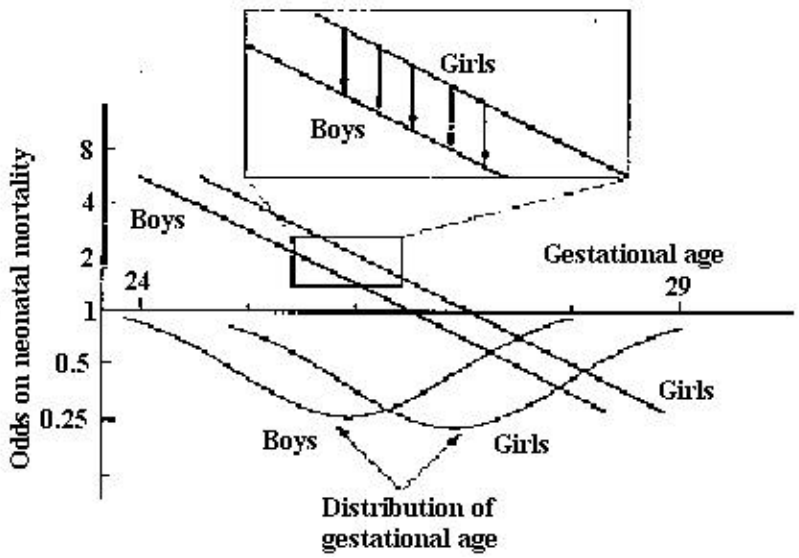

Figure 11.

In Figure 11, the lines that indicate the relationships between neonatal mortality and gestational age in boys and girls are more or less parallel. In Figure 12D, we suppose that there are differences in the way in which gestational age influences outcome between boys and girls.

As mentioned previously, in order to assess the effect of gender per se, we should compare the lines which depict the probabilities in both gender groups. In Figure 11, the answer is straightforward; one can define the distance between two parallel lines and interpret this distance as a difference in chance of dying. Because lines are considered instead of single probabilities, there is an implicit correction for differences in gestational age between boys and girls.

In general, for any given gestational age, the difference in odds between boys and girls can be viewed as the difference between both lines at that specific gestational age. In Figure 11, this difference is the same at each gestational age because the lines are parallel and the result can be expressed by **one** number that is independent of gestational age. In Figure 12D, however, this difference varies with gestational age. Thus, no single estimate can cover the entire range *of* gestation. The estimate varies with the value of the confounder (gestational age) and it becomes necessary to describe how the odds ratio varies with different values of the confounder.

This phenomenon is called **interaction.** This interaction may be due either to a true **effect modification** in the population or merely to the play of chance. The logistic regression technique provides for the appropriate statistical significance test.

Note that in our example, the difference between boys and girls even changes in direction. In a diagram such as Figure 12D, an interaction manifests itself as non-parallelism between the regression lines. Note that this does not necessarily mean that they have to intersect within the observed ranges of the variables involved. In our example, they do in fact intersect, that is, the difference between boys and girls does not just change in magnitude, it also changes in direction. At the lower gestational ages boys have a lower mortality than girls, at higher gestational ages the reverse is observed.

	Gestational age in completed weeks, broken down by sex														
	24		25		26		27		28		29				
Mortality	G	B	G	B	G	B	G	B	G	B	G	B	G	B	T
alive	1	4	5	15	14	26	30	21	26	17	20	7	96	90	186
dead	8	14	20	30	28	34	30	19	13	10	5	3	104	110	214
Odds	8	3.5	4	2	2	1.3	1	0.90	0.5	0.59	0.25	0.43	1.08	1.22	1.15
OR (B:G)	0.44		0.50		0.65		0.90		1.18		1.71		1.13		

Figure 12T.

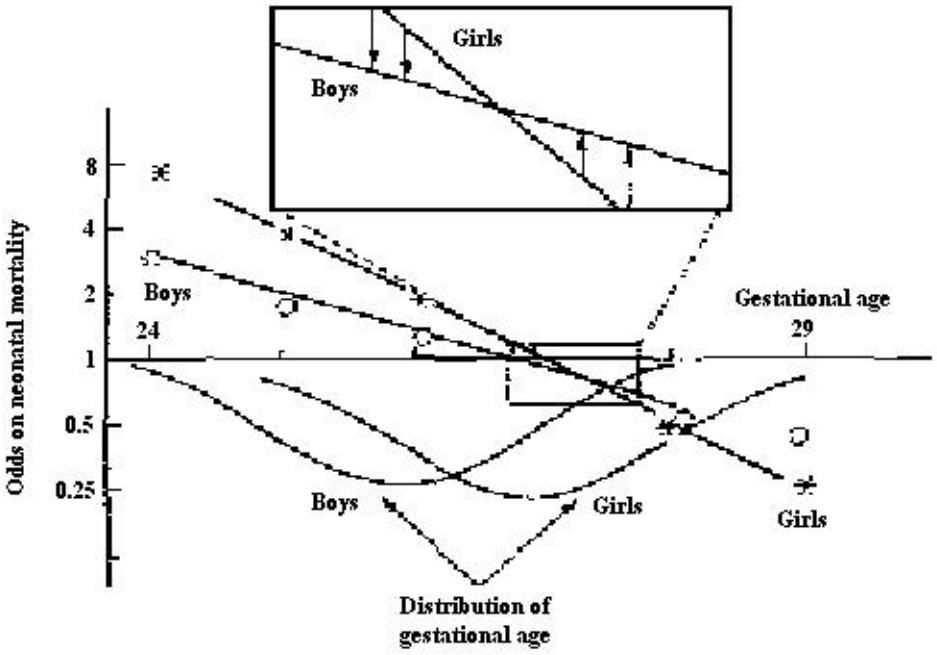

Figure 12D.

When interaction is present in the data, one can still compute a difference between the two lines by 'forcing' them in some way to become parallel and thus provide a kind of 'mean' difference between boys and girls. However, this 'mean' difference is not necessarily a good estimate of the differences in risk between boys and girls at a given value of the confounder. Similarly, an overall mortality rate is not necessarily a good estimate of the risk of dying at a given gestational age (because mortality depends on gestational age).

Nevertheless, the provision of a 'mean' difference may be useful provided that both lines **do not intersect** in the range of clinical interest but are just tending towards or away from each other. When the lines actually cross one should refrain from computing an overall odds ratio. In those circumstances **separate** odds ratios at different values of the confounder(s) are essential.

Exercise: How do you view the use of 'gender' in these sections as an example of an' exposure' variable? Is 'gender' indeed a true exposure in the sense of the definitions as quoted from Kleinbaum *et al.?* Should we always correct sex for a possible confounding effect of gestational age: or never? If 'gender' should not be adjusted for 'gestational age, why not? If so, why is this adjustment still valid in this particular case?

This concludes Part 1 of the introduction to logistic regression. All basic concepts necessary to understand the regression technique have been introduced. In fact, when reading Part 2, one will see that the logistic regression itself has already been introduced, in the framework of confounding and interaction and that the only step needed is the actual 'translation' of these clinical concepts to some - relatively simple - formulae. It is essential to see that, whatever formula will appear in Part 2, it just describes in a straightforward way a situation as depicted in one of the five diagrams displayed so far.

The purpose of Part 2 will be to show the reader how to interpret these regression equations clinically; it will of course not be to teach the reader to carry out the actual analysis himself: to that end one will need a statistician and the appropriate computer programmes.

Acknowledgement

The idea for this paper was born from the first author's participation as a statistician in the nationwide Dutch study on Preterm and Small for Gestational Age Infants (POPS), The Netherlands, 1983. The many fruitful discussions with and help from Pauline Verloove (Department of Neonatology) and Jo Hermans (Department of Medical Statistics) were essential to the present form of this presentation.

References

(x = not referred to in text)

The clinical background for the examples presented came from:

1 Verloove, S.P., Verwey, *R.. Project on Preterm and Small-for-Gestational-Age Infants in the Netherlands, 1983.* Thesis, 1987. Ann Arbor: UMI, 1988; order number 8807276.

General references for logistic regression (in increasing order of statistical detail):

2 Kleinbaum, D.G. Kupper, 11, Morgenstern, *H. Epidemiological Research.* Belmont, California: Lifetime Learning Publications, 1982.

x Matthew, D.E. Farewell. *V.T. Using and Understanding Medical Statistics.* Basel: Karger, 1985.

x Breslow, R., Day, N. *Statistical Methods in Cancer Research.* Lyons: IARC Scientific Publications, No. 82, 1986.