TRUTH IN MUSCULOSKELETAL MEDICINE. III: TRUTH IN THERAPY

Nikolai Bogduk. Newcastle Bone and Joint Institute, University of Newcastle, Newcastle, NSW, Australia.

A speaker announces to you a new therapy; he proclaims a success rate of 80%. Do you believe him? You should not.

The first question that you should ask is - what was N? Given N, you should calculate the confidence interval of the proportion. If N = 10, the success rate is 8/10, whose 95% confidence interval is 55% to 100%. Thus, the true efficacy of the treatment could be as low as 55%. If N= 100, the confidence interval is 72% to 88%, which makes 80% more credible as a representative figure.

The second question that you should ask is - were there controls? Without a control group you cannot be sure that the success rate is actually due to the treatment proclaimed and not to some other, overlooked factor such as the charisma and enthusiasm of the therapist, the setting of the study, the expectation of the patients, and the natural history of the condition treated.

The speaker's claim is depicted graphically in Table 1. Would you regard this treatment as successful and worthy of adoption into your practice?

	SUCCESS	FAILURE
TREATMENT	**80%**	**20%**

Table 1. The reported success of a treatment

Now consider table 2, in which control data are given. Is this a worthwhile treatment?

The data show that it confers no benefit over control. If control happens to be what you are doing already, there is no virtue in changing you practice in order to adopt this new therapy. If control happened to be a placebo therapy, the data show that the success of the treatment is due to non-specific causes.

	SUCCESS	FAILURE
TREATMENT	**80**	**20**
CONTROL	**80**	**20**

Table 2. The reported success of a treatment and success in a control group of patients

Next consider the several sets of data in Table 3. Which of these data do you find compelling; at what stage in the sequence is the difference between treatment group and control group convincingly different? Record your intuitive answer before continuing.

		SUCCESS	FAILURE
1	TREATMENT	80	20
	CONTROL	80	20
2	TREATMENT	80	20
	CONTROL	75	25
3	TREATMENT	80	20
	CONTROL	70	30
4	TREATMENT	80	20
	CONTROL	65	35
5	TREATMENT	80	20
	CONTROL	60	40

Table 3. Five different sets of data on the success of a treatment compared to control therapy.

Having recorded your response, consider the more challenging data in Table 4. Again record your intuitive response to the question - which of these data sets do you find compelling?

		SUCCESS	FAILURE
1	TREATMENT	65	35
	CONTROL	65	35
2	TREATMENT	65	35
	CONTROL	60	40
3	TREATMENT	65	35
	CONTROL	55	45
4	TREATMENT	65	35
	CONTROL	50	50
5	TREATMENT	65	35
	CONTROL	45	55

Table 4. Five different sets of data on the success of a treatment compared to control therapy.

As a greater challenge, consider the data in Table 5. Which is the first set of data to become compelling?

SUCCESS	FAILURE	SUCCESS	FAILURE

1	**TREATMENT**	**8**	**2**	**4**	**TREATMENT**	**8**	**2**
	CONTROL	**7**	**3**		**CONTROL**	**4**	**6**
2	**TREATMENT**	**8**	**2**	**5**	**TREATMENT**	**8**	**2**
	CONTROL	**6**	**4**		**CONTROL**	**3**	**7**
3	**TREATMENT**	**8**	**2**	**6**	**TREATMENT**	**8**	**2**
	CONTROL	**5**	**5**		**CONTROL**	**2**	**8**

Table 5. Six sets of data on the success of a treatment compared to control therapy.

THE CHI-SQUARED (χ^2) TEST

A statistical device is available by which to answer these questions. Known as the chi- squared (χ^2) test, the device determines the probability that the differences between figures in a 2 x 2 contingency table are arbitrary or random. The less likely that the differences are due to chance the more likely that the differences are due to a biological effect, such as the effects of the treatment compared to control.

Without going into details, the statistic - χ^2 , refers to a distribution of differences. In essence, statisticians have laboriously recorded all the possible ways that sets of four figures in a table might differ, and have determined the frequency with which differences of particular magnitudes might occur. All the possibilities have been recorded in tables (or in a computer package) to which reference can be made. By comparing the difference encountered in a particular case to those recorded in the tables, an investigator can determine the probability of that difference being due to chance.

	COLUMN 1	COLUMN 2	
ROW 1 ROW 2	a c	b d	a + b c + d
	a + c	b + d	a + b + c + d = N

Figure 1. The elements of a 2 x 2 contingency table. The columns represent possible outcomes, such as success and failure. The rows represent groups, such as treatment and control. The numbers inside the central square, a,b,c and d, represent the number of cases in each class of results. Outside the square are the sums of the respective columns and rows. N is the total number of subjects.

The principles of operation of the χ^2 test are:

assume that chance is operating;
calculate the result that would be expected if chance alone was operating;
compare the observed results with those expected from chance alone.

Figure 1 depicts the distribution of numbers in a 2 x 2 contingency table. The columns would represent outcome of a therapy, such as success and failure. The rows would represent assignment to treatment or to control. The numbers inside the central square represent the observed results. The cell known as row 1, column 1 (R_1C_1) contains *a* cases, and represents the number of patients who obtained success with the treatment. The cell known as row 1 column 2 (R_1C_2) contains *b* patients who failed the treatment. The cell known as row 2, column 1 (R_2C_1) contains *c* patients who responded successfully to control therapy, and the cell known as row 2, column 2 (R_2C_2) contains *d* patients who failed control therapy. The numbers outside the square represent the sums of the respective columns and rows.

The first step is to calculate what results would have occurred by chance alone. For this, the outside figures are used. On the average, success occurs, regardless of assignment to treatment, in *(a + c)* cases out of *N* (column 1). Similarly, failure occurs, on the average in *(b + d)* cases out of *N* (column 2). Meanwhile, *(a + b)* patients were allocated to treatment (row 1). If this group of patients responded simply in an average manner, the number of patients that would be expected to fall into cell (R_1C_1) would be:

$$R_1C_1 = \frac{(a+c)}{N} x(a+b)$$

Similarly, the number of cases that would be expected to fall into cell R_1C_2 would be:

$$R_1C_2 = \frac{(b+d)}{N} x(a+b)$$

In the same way, *(c + d)* patients were allocated to control (row 2). If this group of patients responded simply in an average manner, the number of patients that would be expected to fall into cell R_2C_1 would be:

$$R_2C_1 = \frac{(a+c)}{N} x(c+d)$$

and the number in cell R_2C_2 would be:

$$R_2C_2 = \frac{(b+d)}{N} x(c+d)$$

If we let

$R_1C_1 = a^*$
$R_1C_2 = b^*$
$R_2C_1 = c^*$
$R_2C_2 = d^*$

the differences between the observed values (a,b,c and d) in the table and the expected values (a*, b*, c* and d*) can be calculated. In order to define the difference, the χ^2 statistic uses the operation:

$$\chi^2 = \frac{(observed - \exp ected)^2}{\exp ected}$$

It compares the square of the difference between the observed and expected results in any cell with the expected result for that cell. For the table as a whole, the χ^2 value is the sum of the χ^2 values of each of its cells. This sum value can then be compared with values recorded in statistical tables to determine the probability of this χ^2 value having occurred simply by chance. (In computerised versions, this comparison is done automatically.)

The figure that is provided from the tables is a probability. By convention, the limit of statistical significance is a probability of 5% (i.e. 0.05). If the probability is less than 0.05 that the differences that occur in a given table could have occurred by chance, one can conclude that the figures in the table are unlikely to have arisen by chance and, therefore, are more likely to be due to a real, biological effect, such as the effect of the treatment.

A short form of the equation for χ^2, suitable for use within pocket calculators [2] is:

$$\chi^2 = \frac{(ad - bc)^2 N}{R_1 R_2 C_1 C_2}$$

where R_1, R_2, C_1, and C_2 are the sums of the rows and columns of the contingency table.

Example

Consider the following data (Figure 6).

	SUCCESS	FAILURE	
TREATMENT	36	10	46
CONTROL	20	16	36
	56	26	82

Figure 6. The reported success of a treatment and success in a control group of patients

The expected value (a*) for cell R_1C_1 = 56/82 x 46 = 31.4
The expected value (b*) for cell R_1C_2 = 26/82 x 46 = 14.6
The expected value (c*) for cell R_2C_1 = 56/82 x 36 = 24.6
The expected value (d*) for cell R_2C_2 = 26/82 x 36 = 11.4

Thus the expected table looks like figure 7. These expected figures are somewhat different from the observed figures. If chance alone were operating one would expect somewhat fewer success from the treatment than what was observed. Thus, it appears that perhaps the treatment is exerting an effect greater than chance. But is it significantly better statistically? The answer to that question is provided by the χ^2 value.

	SUCCESS	FAILURE

TREATMENT	**31.4**	**14.6**	**46**
CONTROL	**24.6**	**11.4**	**36**
	56	**26**	**62**

Figure 7. The expected values for success and failure for the data in figure 6.

For the data in figure 6,

χ^2 value for R_1C_1 = $(36 - 31.4)^2 / 31.4 = (4.6)^2 / 31.4 = 0.67$
χ^2 value for R_1C_2 = $(10 - 14.6)^2 / 14.6 = (-4.6)^2 / 14.6 = 1.45$
χ^2 value for R_2C_1 = $(20 - 24.6)^2 / 24.5 = (-4.6)^2 / 24.5 = 0.86$
χ^2 value for R_2C_2 = $(16 - 11.4)^2 / 11.4 = (4.6)^2 / 11.4 = 1.86$

and the sum of the χ^2 values is 4.8. Upon reference to a table of probabilities of χ^2 values, we find that the probability of such a difference arising by chance alone is less than 0.03. Therefore, there is not more than a 3% probability that the data in figure 6 arose by chance alone. Consequently, we can conclude that the difference between treatment and control most likely arose because of the effects of the treatment.

FISHER'S EXACT TEST

One of the frailties of the χ^2 test is that for mathematical reasons its sensitivity lapses when the numbers in the raw data are small. This applies in general when the values, a,b,c, and d, are less than 10, and definitely when they are less than 5. For data of this size a more rigorous mathematical treatment is required. This is the Fisher's exact test [1].

A complete explanation of the principles of Fisher's exact test is quite demanding. However, in brief, it involves determining all the possible ways that four small numbers in a contingency table could be generated, and the relative frequency of each combination. From such an analysis the probability of a given pattern of numbers arising simply by chance can be calculated.

At an initial stage in their edification, readers should not be required to know or remember how Fisher's exact test should be calculated; that can be left for an advanced level of study. However, it is imperative that consumers of information in musculoskeletal medicine should know to call for a Fisher's exact test instead of a χ^2 test, when the numbers involved are small.

EXAMPLES

Armed with a knowledge of the χ^2 test and Fisher's exact test, readers can now return to the examples from above in which they were invited to record their intuitive response to data (Tables 3, 4 and 5).

In table 6, the χ^2 values show that the first set of data that become statistically significant are those of set 4. Only when the control group scores 65,35 is a score of 80,20 in the treatment group significantly better.
In table 7, the χ^2 values show that only when the control group scores 50,50 is a score of 65,35 in the treatment group significantly better.

		SUCCESS	FAILURE	χ^2	P
1	**TREATMENT**	**80**	**20**	**0.0**	**1.0**

	CONTROL	80	20		
2	TREATMENT	80	20	0.7	0.40
	CONTROL	75	25		
3	TREATMENT	80	20	2.7	0.10
	CONTROL	70	30		
4	TREATMENT	80	20	5.6	0.02
	CONTROL	65	35		
5	TREATMENT	80	20	9.5	0.002
	CONTROL	60	40		

Table 6. Results of a χ^2 test on the data in table 3.

		SUCCESS	FAILURE	χ^2	P
1	TREATMENT	65	35	0.0	1.0
	CONTROL	65	35		
2	TREATMENT	65	35	0.53	0.47
	CONTROL	60	40		
3	TREATMENT	65	35	2.08	0.15
	CONTROL	55	45		
4	TREATMENT	65	35	4.60	0.03
	CONTROL	50	50		
5	TREATMENT	65	35	8.08	0.004
	CONTROL	45	55		

Table 7. Results of χ^2 test on the data in table 4.

The data in table 5 consisted of small numbers. Therefore, a χ^2 test is not appropriate. These data require Fisher's exact test. Table 8 shows the results of statistical analysis of the data sets of table 5. Notice how the first set of data that become significantly different statistically is set 5. It is not until the data read 8,2 vs 3,7 that statistical significance occurs. Compare this observation to the data in table 6. There, a significant difference occurred for 80,20 vs 65,35. Realise that although 8/10 might look like 80%, or be portrayed as such, it is not the same as 80/100. Although 80,20 is significantly different from 60,40, 8,2 is not significantly different from 6,4. It is not until the difference is between 8,2 and 3,7 that statistical significance occurs.

		SUCCESS	FAILURE	χ^2	P	Fisher's Exact Test (P value)
1	TREATMENT	8	2	0.27	0.6	0.5
	CONTROL	7	3			

2	**TREATMENT**	**8**	**2**	**0.95**	**0.3**	**0.3**
	CONTROL	**6**	**4**			
3	**TREATMENT**	**8**	**2**	**1.98**	**0.16**	**0.18**
	CONTROL	**5**	**5**			
4	**TREATMENT**	**8**	**2**	**3.3**	**0.07**	**0.09**
	CONTROL	**4**	**6**			
5	**TREATMENT**	**8**	**2**	**5.1**	**0.03**	**0.04**
	CONTROL	**3**	**7**			
6	**TREATMENT**	**8**	**2**	**7.2**	**0.007**	**0.012**
	CONTROL	**2**	**8**			

Table 8. Results of a χ^2 test and Fisher's exact test on the data in table 5.

A further feature to note in table 8 is the change in probabilities in the χ^2 test and Fisher's exact test. Note how the χ^2 test approaches the critical value of 0.05 more rapidly than does Fisher's test. This indicates how the χ^2 is more "generous" to data that comes as small numbers. Fisher's exact test is more demanding; a greater difference between the numbers is required before statistical significance is achieved. Readers are thereby warned against accepting χ^2 results on small numbers.

ADVANCED TRUTH: SURVIVAL CURVES

A limitation of χ^2 tests is that they operate only on single sets of data. Thus, they can reveal significant differences between the results of treatment and control, but only at a fixed point in time. An investigator might report statistically significant differences but that significance applies only to the data that the investigator chooses to present.

For a treatment to be worthwhile, not only must it be significantly different from control, but it must maintain that difference over time. A single 2 x 2 table about a treatment is insufficiently informative. The best picture of a treatment is provided by survival curves.

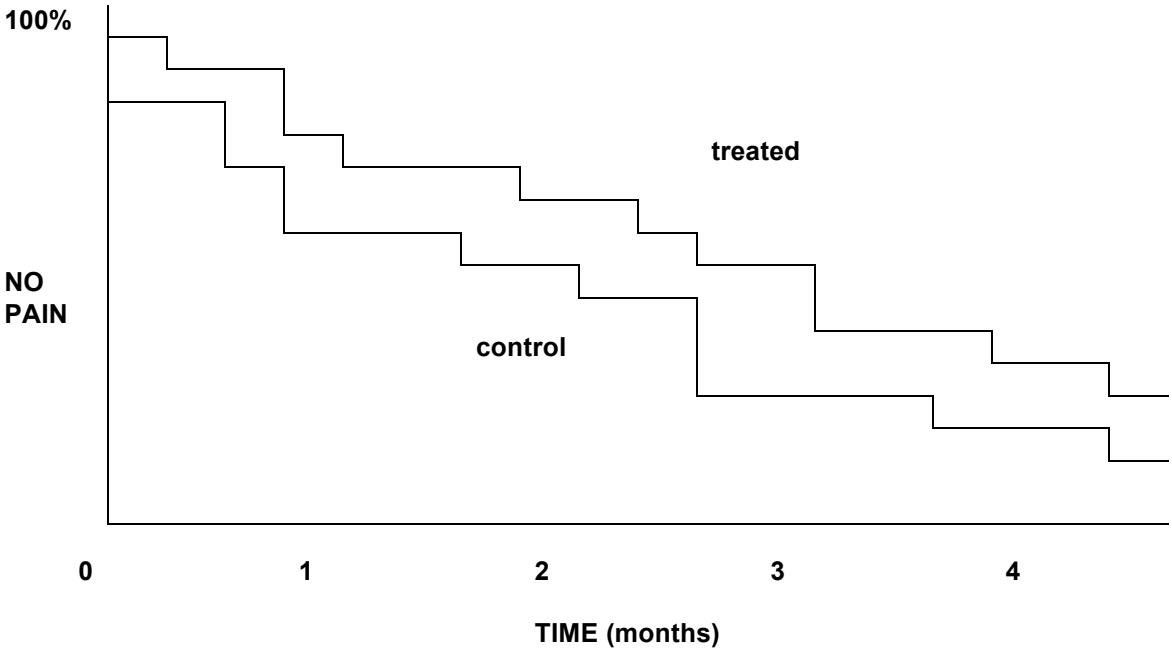


Figure 8. A pair of survival curves between which there is no apparent and no statistical difference.

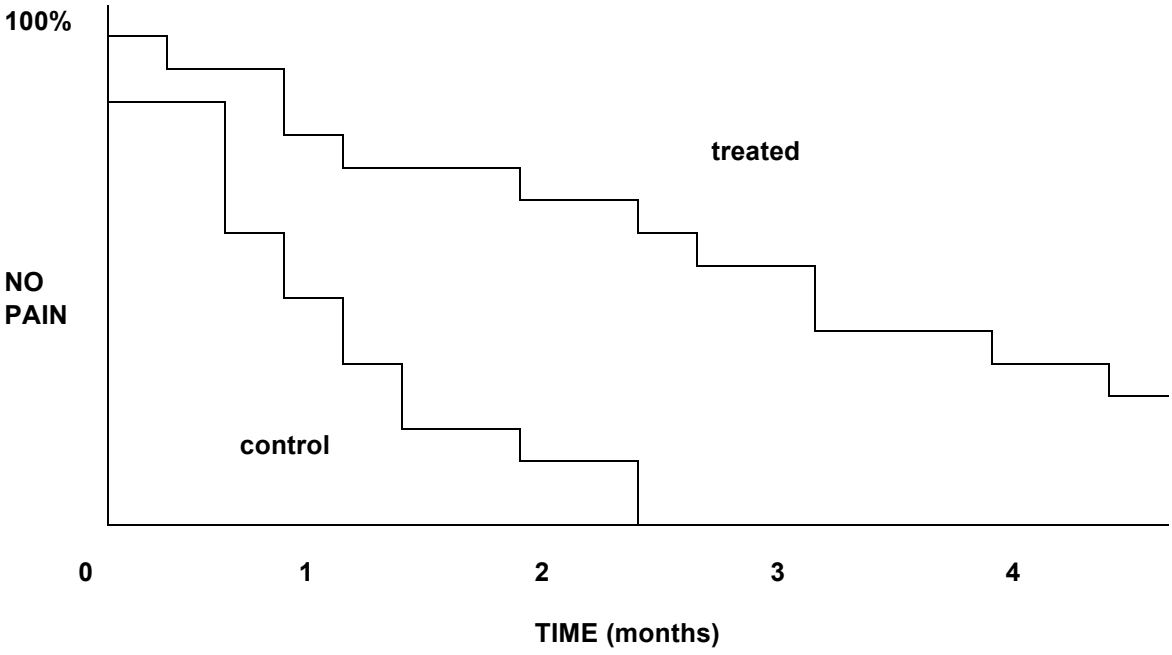


Figure 9. A pair of survival curves between which there is an obvious difference over time in the number of patients remaining with no pain.

In a survival curve, the proportion of patients remaining with a successful result (e.g. relief of pain) is plotted over time. Initially, most, if not all patients, successfully respond to treatment. Progressively, however, some patients may suffer relapses, as the effect of the treatment wanes. A survival curve plots this

decay in response, and provides a picture of the durability of the therapy. If the survival curves of a treated group of patients and their control group are plotted simultaneously, the differences between the two can be monitored over time.

Figure 8 shows an example of such survival curves. Both curves are closely parallel and show the same rate of decay. On inspection alone, without applying any statistical test, one could infer that there was no appreciable difference in the response rates of the two groups of patients, both initially and over time.

Figure 9 shows a different situation. Here, members of the treated group tend to maintain their response over time. Meanwhile, the control group relapses rapidly. In the first two weeks, there is no appreciable difference between the two groups, but quite obviously, by the second and subsequent months, there is a marked difference between the two.

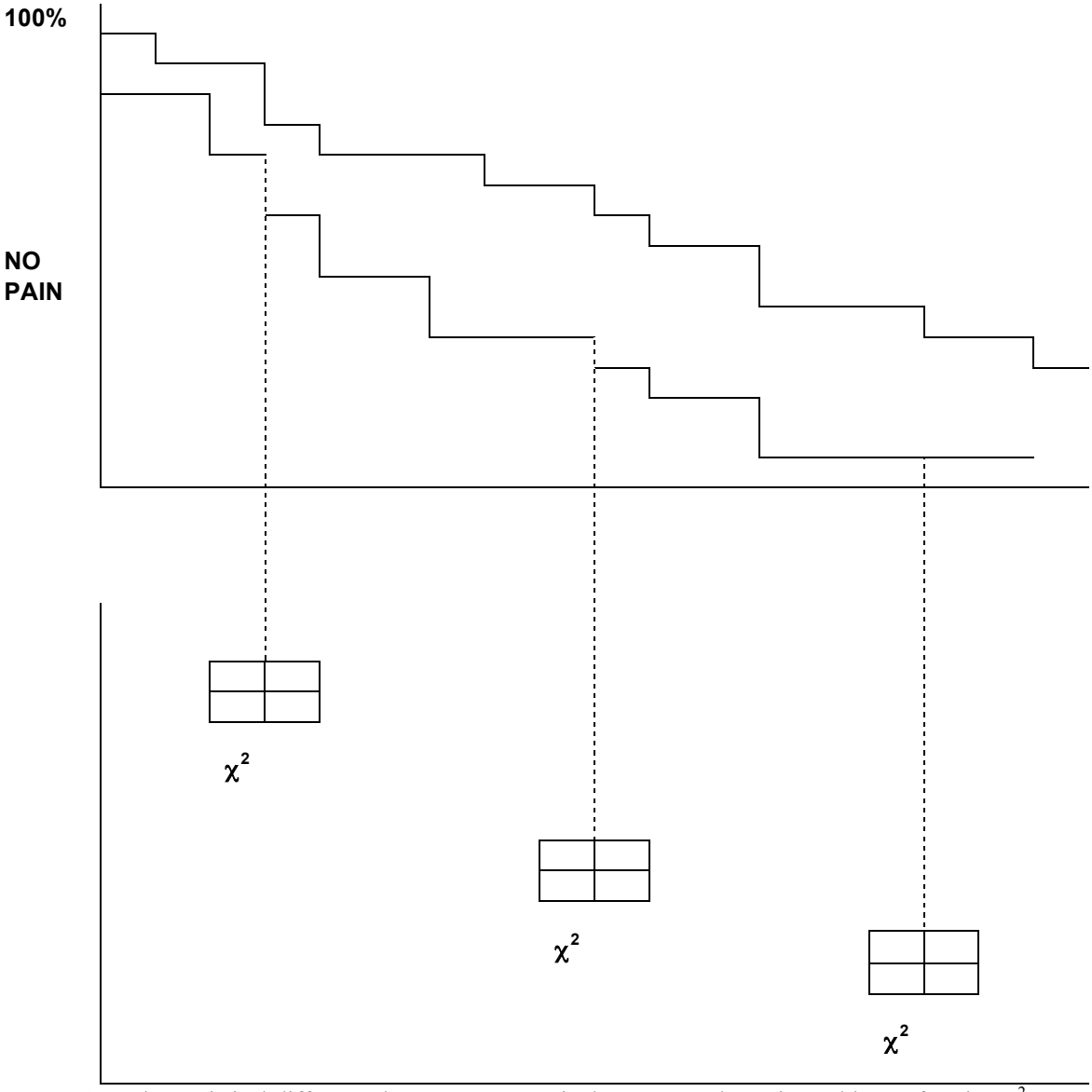


Figure 10. The statistical difference between two survival curves can be estimated by performing χ^2 tests at selected points in time along the curves.

Statistical tests maybe applied to survival curves in order to check mathematically whether or not the apparent difference between two curves is statistically significant. The mathematics involved is relatively complex, but in principle amounts to applying a χ^2 test to the data at selected points along the two curves

(Figure 10). At any given point in time, there will be successes and failures both in the treatment group and in the control group. These respective numbers can then be entered into a χ^2 table, and the significance of any difference can be calculated, for that point in time.

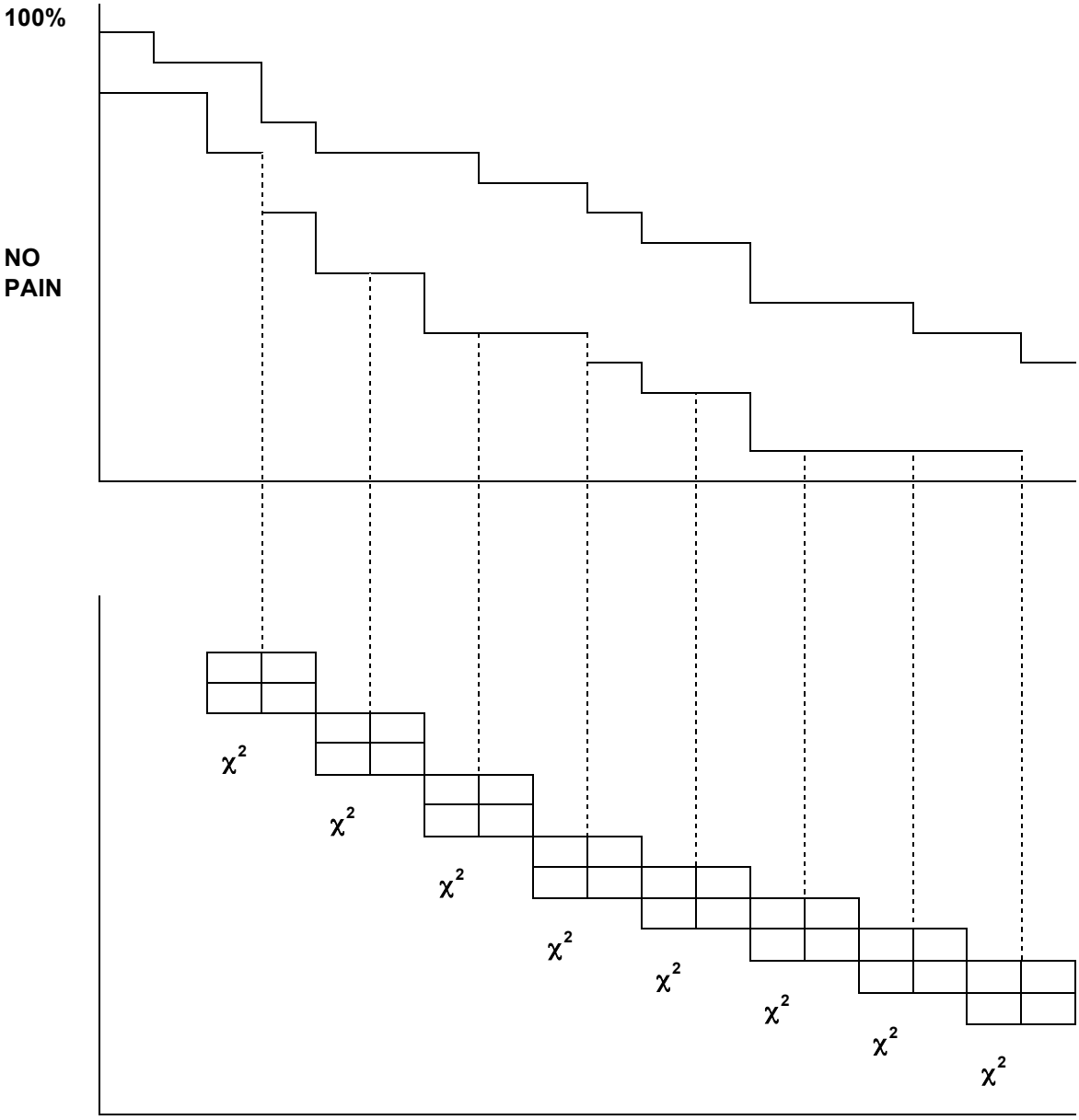


Figure 11. The statistical difference between two survival curves can be estimated by performing χ^2 tests along the entire length of the curves.

A comprehensive analysis can be achieved by applying χ^2 tests along the entire length of two curves (Figure 11). Such analysis is the basis of the Mantel-Haenszel test and the log-rank test [2]. The virtue of such an analysis is that, in cases where at any one point there may not be a statistically significant difference between treatment and control, overall there may nevertheless be a difference over time. The significant difference is revealed essentially by integrating the results of the continuous χ^2 tests.

Another virtue of survival curves is that it protects readers from being seduced by slight of hand, when investigators do not show all of their data but reveal only one snapshot of the only portion of their data that happens to be significantly different.

The case depicted in figure 12 is one illustration of this phenomenon. Here, there is a significant difference between treatment and control in the third week after treatment but not thereafter. An unscrupulous investigator might present their results "at three weeks", and quite legitimately report a statistically significant difference. The risk, however, is that you as a consumer might infer that this proves that the treatment is worthwhile, because it is significantly better than control. But that inference applies only at three weeks. What the investigator has hidden from you is that there is no difference in the first week, and none after the first month. Is this the sort of treatment that you want to buy?

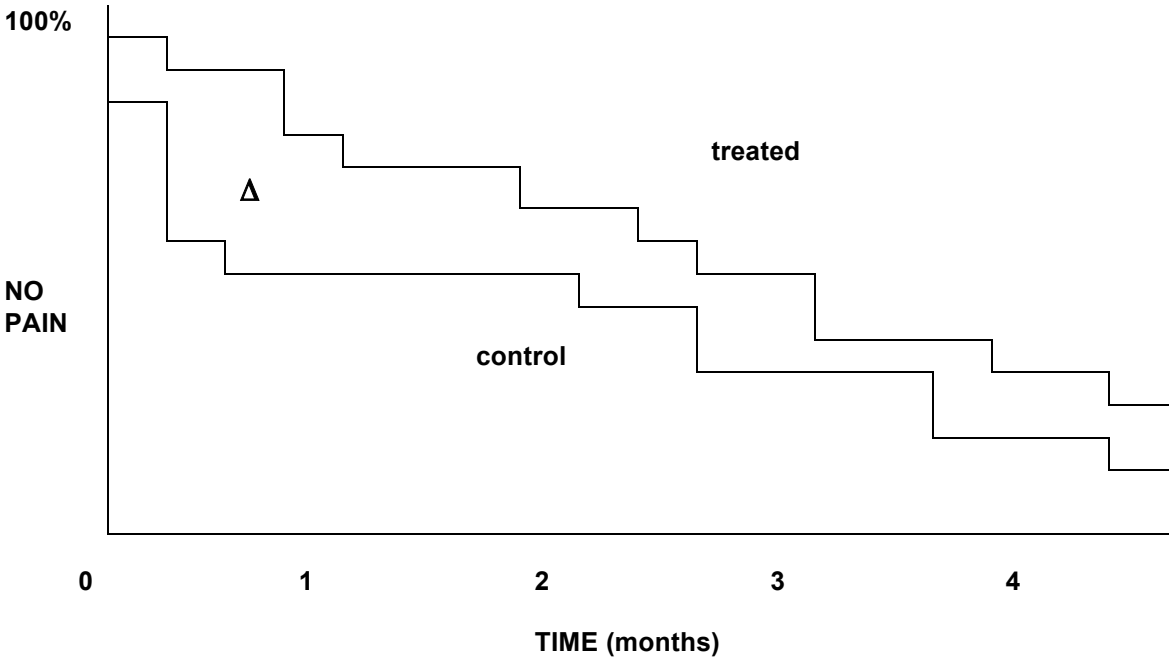
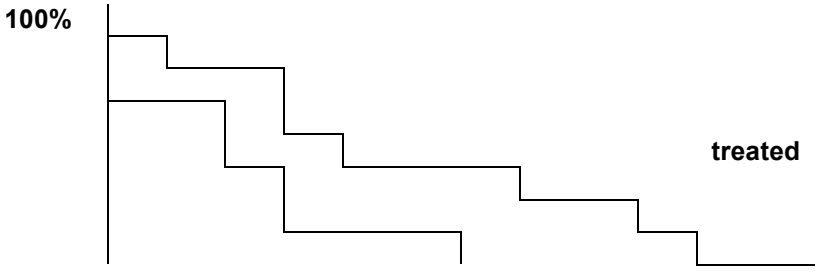


Figure 12. A pair of survival curves between which there is an apparent and statistically significant difference between treated and control groups during the first month (Δ), but after which there is no difference between the curves.

A similar phenomenon is illustrated in figure 13. Here, there appears to be a significant difference at two months. An unscrupulous investigator might report a genuine statistical difference, but show you only the χ^2 test on the data at two months. However, the survival curves are more revealing. There is no difference for the first six weeks, and no difference after the third month. Overall the survival curves indicate no significant difference. There is no sensible biological explanation for the difference at two months that is consistent with a treatment effect. Rather, the difference that arose is more likely to be an aberration. In any case, you are not likely to want to use a treatment that has an unpredictable effect only for the 8th to 12th week after therapy.



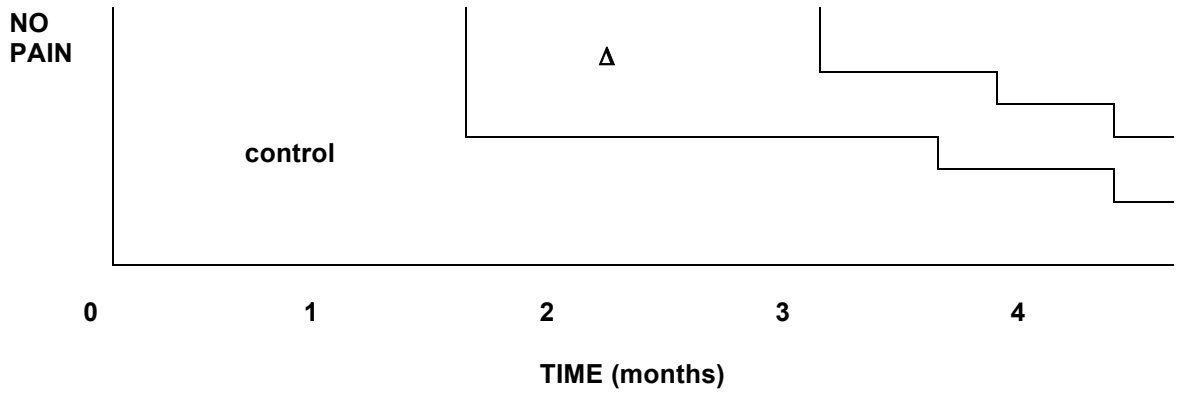


Figure 13. A pair of survival curves between which there is no difference initially or in the fourth month, but a difference is present (Δ) during the third month only.

CONCLUSIONS

As with diagnostic tests, for the assessment of therapy

all truth comes in a 2 x 2 table.

An individual, intent upon promulgating a new therapy, cannot know the truth if he does not have 2 x 2 data. You, as a consumer, cannot know the truth unless and until you are shown the 2 x 2 data. Furthermore, the complete picture of a treatment is not evident until the survival curves are shown.

As a consumer, without the truth tables, you cannot determine if you are being lied to, or if the investigator is irresponsible. In order to protect yourself, learn to demand the truth tables, and in the case of treatments, come to demand the χ^2 tests or the Fisher's exact tests. Without such data you may adopt and perpetuate mythology. The choice is yours.

REFERENCES

1. Armitage P, Berry G. Statistical Methods in Medical Research. Oxford: Blackwell, 1994. pp 137-141.

2. Armitage P, Berry G. Statistical Methods in Medical Research. Oxford: Blackwell, 1994. p 135.

3. Armitage P, Berry G. Statistical Methods in Medical Research. Oxford: Blackwell, 1994. pp 477-481.