Risedronate, the BBC and me

Risedronate is a drug used to reduce fractures in women with osteoporosis. The story of the Sheffield Risedronate trial made national news when one of the clinical researchers, Aubrey Blumsohn, dissociated himself from the trial because he was refused access to the data he had helped to collect. Later, some of that data turned up. Here **Martin Bland** describes how he came to analyse the data on behalf of the BBC, what he found and what happened next.

Doing it on the radio

Compared with chefs, statisticians do not have a large media presence. I think we need to raise our profile, and when I have been approached by broadcasters I have usually said yes. I have always found it an interesting experience. This story began in September 2005, when I was approached by journalist Vivienne Parry about a programme for BBC Radio 4. The programme was to be part of the *File On Four* series and would concern a clinical trial of the drug Risedronate, carried out in Sheffield. She wanted me to analyse some data which had originated in the trial.

Risedronate is a drug manufactured by Procter and Gamble; it is one of the bisphosphonate class of medicines, which reduce the risk of bone fractures in high-risk groups—in this trial women with osteoporosis. They work by reducing the turnover of bone material.

The Sheffield trial was a double-blind randomised controlled trial comparing Risedronate with a placebo. Women were followed for 3 years, with periodic X-rays, and with any fractures they suffered being recorded. The trial was sponsored by Procter and Gamble, who also managed and analysed the data. One of the clinicians involved, Aubrey Blumsohn, had been in dispute with Procter and Gamble. I will not attempt to describe the details of the dispute, but the essence is that Procter and Gamble wanted him to put his name to presentations and papers relating to the Sheffield trial. He was sceptical about their analysis and asked for the data so that he could try it for himself. Procter and Gamble, however, refused him access to unblinded data. He, therefore, refused to sign his name to the research. However, one day, Aubrey Blumsohn discovered that, unknown to Procter and Gamble, he did in fact have some of the data, buried deep within a graphics file with which they had supplied him. I was asked whether I would carry out an independent analysis of these data.

The point at issue was this: if we relate fracture incidence to a marker of bone material turnover, is there a plateau after which fracture rate does not reduce no matter how low the percentage change in the marker is? This was what was claimed in the paper that was published and from which Blumsohn removed his name. [The marker goes by the name of Cross-linked N-telepeptide, or NTX for short.]

I was asked two questions:

1. Do the data generated by Aubrey Blumsohn's work support the conclusion that there is a plateau effect in the relationship of bone fractures with changes in the marker NTX when using Risedronate? The plateau is said to occur at and below a 30% fall in NTX.
2. Do the data support the conclusion that "The relationships between vertebral fracture risk and changes from baseline and NTX were not linear ($P < 0.05$). There was little further improvement in fracture benefit below a decrease of 35–40% for NTX."

Proctor and Gamble refused their researcher access to data that he himself had helped to collect

The data

The data were supplied to me by Vivienne Parry. There were data from two studies: an American study, known as HIP, and the Sheffield study, known as VERT. I used only four variables in the analysis:

- The study concerned, HIP or VERT.
- The percentage change in NTX from baseline at 3–6 months.
- Whether a fracture was observed within 3 years.
- The time to fracture or time at end of follow-up if earlier.

There were 224 women in the HIP study, of whom 11 experienced fractures, and 326 in the VERT study, 36 of whom had fractures.

The distribution of percentage change in NTX is shown in Figure 1. The histogram shows a highly positively skew distribution, with only a few positive increases in NTX and some extremely high outliers. Table 1 shows the numbers of fractures and times to the event or to the end of follow-up. Many fractures could be observed only on X-ray and there are clear peaks at 12, 24 and 36 months, both in fractures and in people who were no longer followed up.

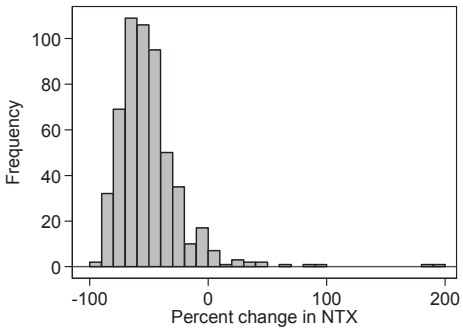

Figure 1. Distribution of percentage change in NTX

Table 1. Fractures and time to event in both studies combined

Time to event (months)	*Fracture*		
	No	*Yes*	*Total*
3	1	0	1
6	4	1	5
9	3	0	3
12	43	11	54
15	6	3	9
18	8	1	9
21	5	0	5
24	51	13	64
27	2	1	3
30	2	0	2
33	6	1	7
36	362	16	378
Total	493	47	540

I thought that this was clearly a job for Kaplan–Meier survival curves and Cox proportional hazards regression. Kaplan–Meier survival curves are plots of the estimated proportion of subjects who have not yet experienced a fracture, against the time of follow-up. This takes into account that some subjects were not followed for the full 36 months. Cox proportional hazards regression uses a model where anything that alters the risk of a fracture does so in a constant ratio. For example, if the HIP trial patients had twice the risk of a fracture as the VERT trial patients, we would assume that this doubling of risk continued throughout the period of follow-up, this ratio of risk being called the hazard ratio. For a quantitative variable such as percentage change in NTX, we estimate the hazard ratio per unit difference in the variable, e.g. per percentage point difference. In this trial, we do not know the exact times of fractures. The event refers to the detection of a vertebral fracture rather than the occurrence of the fracture.

Analysis strategy

I took two approaches to the analysis. The first was to test for the presence of a "plateau" effect, whereby reductions in NTX greater than 30% had the same effect as a reduction of 30%. The problem with doing this was that the plateau hypothesis may have been generated in part from the data with which we test it, giving it a greater chance than we would normally expect of producing a spurious plateau. For this reason, my second approach was to work from the standpoint of ignorance and attempt to model the relationship between time to fracture and percentage change in NTX.

Preliminary analyses and survival plots

The Kaplan–Meier survival curves for the two trials are shown in Figure 2. The survival without fracture appeared to be better for the HIP patients than for the VERT patients. This was confirmed by Cox regression, which estimated the hazard ratio as 2.33. This means that the risk of fracture at any given time for the VERT group was 2.33 times the risk for the HIP group. This was statistically significant (P = 0.01) and we have good evidence that the trial populations were different. The

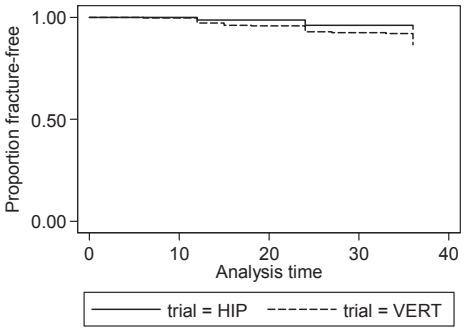

Figure 2. Kaplan–Meier survival curves for fractures in the two trials

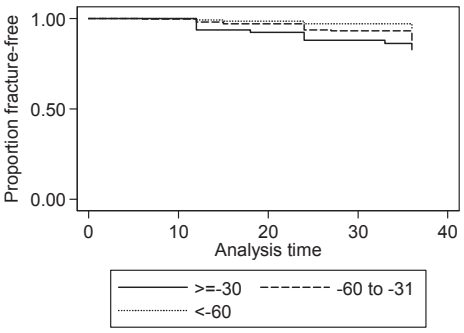

Figure 3. Kaplan–Meier survival curves for fractures by NTX percentage change, in the three groups indicated

95% confidence interval was 1.18–4.57, so we estimate the hazard ratio between the two trial populations, as opposed to the trial samples, to be between 1.18 and 4.57.

Figure 3 shows graphically the relationship between fracture-free survival and percentage change in NTX, with the percentage change classified into three groups: (i) greater than −30%, i.e. above the putative plateau; (ii) between −31% and −60%; (iii) below −60%. If there is a plateau at values more negative than −30%, the second and third groups should not differ in their fracture-free survival. This is not the case and survival appears to be best in the below −60% change group, i.e. those with the largest fall in NTX.

As a simple linear effect, percentage change in NTX was estimated to have a hazard ratio of 1.0070, P = 0.02. The hazard ratio seems very close to 1 because it is the hazard ratio for a difference of one percentage point in percentage change in NTX. It means that the risk of fracture at any moment is multiplied by 1.007 for each increase of one percentage point in the percentage change. An increase here means becoming more positive, moving from −100%, the maximum possible fall in NTX, upwards toward 0% and on into the positive area, increases in NTX. The P value indicates that there is evidence that this is a real effect, with higher (more

positive) percentage changes being linked to increased risk of fracture. The question to ask now is whether the linear model is a good fit to the data.

Results of the test for a plateau effect

To test the plateau at changes more negative than –30%, the percentage change in NTX was split into two variables. The first, above the plateau, was equal to the percentage change unless the percentage change was less than –30% (i.e. more negative), when it was set equal to –30%. The second variable, on the plateau, was equal to the percentage change + 30 unless the percentage change was greater than –30% (i.e. in the positive direction), when it was set equal to 0. The effect of this is that the sum of these two variables is the percentage change. I included both of them in the Cox regression model, along with the trial variable. If there were a plateau, we would expect changes more negative than –30% to have no effect. We therefore test whether the second variable, on the plateau, has any effect upon fracture-free survival. In fact, the hazard ratio was 1.019, $P = 0.05$, for the percentage change less than –30%, i.e. on the putative plateau, giving some rather weak evidence that there is an effect in the region of the plateau. For the percentage change on the positive side of the plateau the hazard ratio was smaller, at 1.0032, and was not statistically significant, $P = 0.5$, so there was no evidence for any effect on the positive side of –30% in this analysis. This does *not* mean that there was no effect. It means that this analysis failed to demonstrate one. However, these results are the opposite of what we would expect if there were a plateau at percentage change in NTX more negative than –30%.

> About all that we can say about these curves is that they do not suggest a plateau

We can test the hypothesis that the plateau model fits the data better than does a simple linear model. We do this with a chi-squared test, using the difference between the chi-squared statistics for the two models. The difference is 12.29 – 10.54 = 1.75 with 1 degree of freedom. This gives $P = 0.2$, so there is no evidence that a plateau model fits better than a linear model.

I also repeated the analysis excluding subjects with positive changes, using a change point at –40% rather than –30%, and for the VERT trial subjects only, with very similar results.

Results of the open analysis

My first step was to test whether a linear fit was adequate. To do this I included a non-linear term, by squaring the percentage change. I then used a chi-squared test to see whether the model including the squared term fits the model better than the model with only the linear term. This gave chi-squared = 15.73 – 10.54 = 5.19 with 1 degree of freedom, with $P = 0.02$, so there is evidence that a non-linear model fits better than a linear model.

Now the question is: what non-linear model best predicts fracture-free survival? The numbers of fractures are too small to answer this question properly, but I plotted the estimated hazard ratio against the value of percentage change in NTX anyway, as shown in Figure 4. (The ratio is set to 1.00 when the NTX change is zero.) This looks very odd, as the hazard rises then plunges down as the percentage change in NTX becomes large and positive. This is clearly nonsense and arises because there are just two outlying observations with very high positive percentage changes, where the subjects did not have fractures. For this analysis it may well be better to exclude them. Repeating the analysis with only those subjects with negative changes gives a chi-squared test for non-linearity = 12.42–11.06 = 1.36, $P = 0.2$. There is no evidence of non-linearity. Figure 5 shows a plot of hazard ratio against change in NTX. Figure 5 also looks very unsatisfactory. I think that general curve fitting is not possible with so few fractures in the data. I also attempted a more powerful curve-fitting technique using fractional polynomials, but the curves produced were very similar to these. About all we can say about these curves is that they do not suggest a plateau.

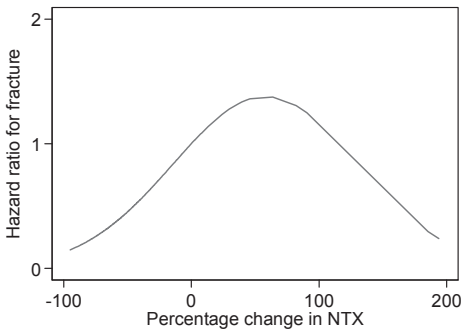

Figure 4. Estimated hazard ratio against percentage change in NTX

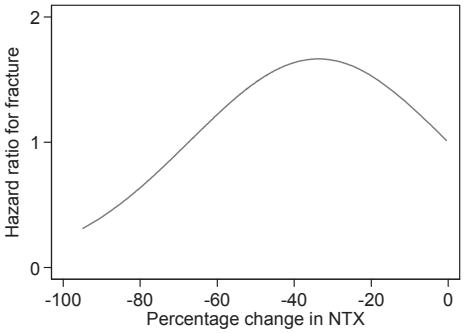

Figure 5. Estimated hazard ratio against percentage change in NTX for negative NTX changes only

Analyses of these data are unreliable because of the small number of fractures. These studies were designed as placebo-controlled trials and, as there is little doubt that bisphosphonates reduce the risk of bone fracture in susceptible people, I suspect that they were of adequate size to compare the active treatment to no treatment. I do not think that they are of adequate size to give a reliable estimate of the shape of the NTX change response curve.

What did the original authors do?

Having sent off my report to the BBC, I then began to wonder just what Eastell *et al.*[1] did to produce their plateau. The paper includes a figure similar to Figure 6. Only the solid lines concern us as these are for the Risedronate group. All the other analyses in the paper used survival analysis but this did not. What did they do? The methods section of the paper included:

> "To visualize the association between fracture incidence and early changes in bone turnover makers [sic], the probability of sustaining a fracture was plotted against

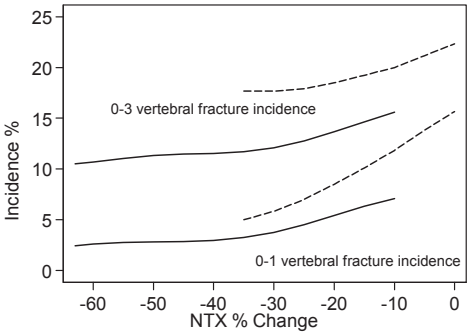

Figure 6. Incidence of vertebral fractures against NTX change, as shown by Eastell *et al.*[1] (solid lines are for the Risedronate group, dashed lines are for the placebo group)

> the 3- to 6-month bone turnover maker [sic] data. Empirical displays of the incidence were constructed using a smoothing curve. Because these displays were not model-dependent, no confidence intervals were constructed.[1]"

As I found this fairly opaque, I decided to play around with the data and see whether I could reproduce Figure 6 by reverse engineering. I started with a plot of fracture (1 = fracture, 0 = no fracture) against percentage change in NTX, as shown in Figure 7. This completely ignores the time of follow-up, of course. I then fitted a curve to this using a smoothing algorithm. I had never done anything like this before, but Stata offered locally weighted smoothing for scatter plots (LOWESS), so I tried it, and it gave me Figure 8.

This is not attractive, as the extreme outliers have produced an unconvincing plunge to zero incidence at percentage change in NTX = 200. Therefore, I omitted these two outliers to give Figure 9, which looked much better. Next I got rid of the fracture variable points at 0 and 1, because they looked very odd. At the same time, I cut off the curve at the top and bottom. I did not like the upward sweep at the extreme left, but if I just cut that off I would have a downturn at about −70, so I cut it off above that, to give Figure 10.

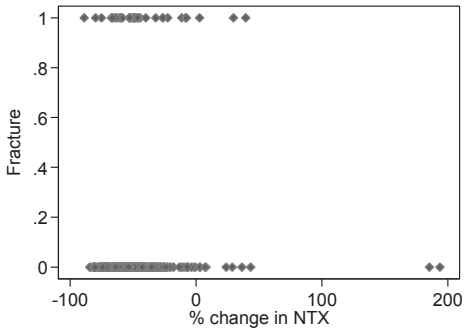

Figure 7. Presence or absence of a fracture (1 = fracture, 0 = no fracture) against percentage change in NTX

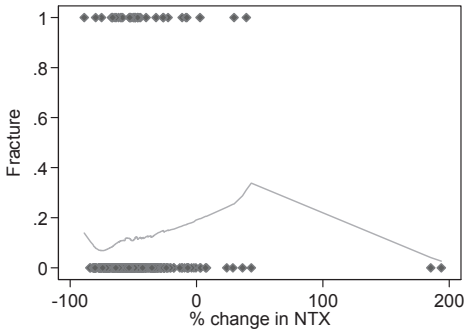

Figure 8. LOWESS smoothed curve for presence or absence of a fracture against percentage change in NTX

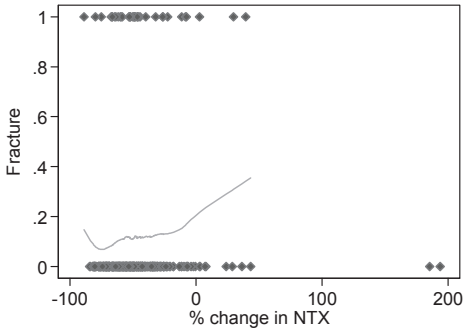

Figure 9. LOWESS smoothed curve for presence or absence of a fracture against percentage change in NTX, omitting the high percentage change outliers

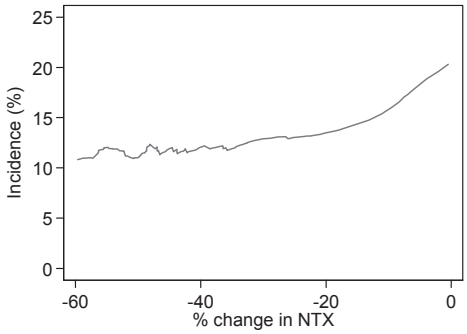

Figure 10. LOWESS smoothed curve for presence or absence of a fracture against percentage change in NTX, truncated at −60% and 0%

I relabelled the vertical axis as "Incidence", because it represents an estimate of the proportion of subjects who had fracture. Now I had a flattish part to the curve which might be described as a plateau. (I thought you went up to a plateau, not down, but I am no geographer!) I did it for 12 months, too, to give Figure 11. If we were to smooth these out a bit more, we would get something quite like Figure 6. As far as I am aware, all requests to describe how Figure 6 was actually produced have been refused.

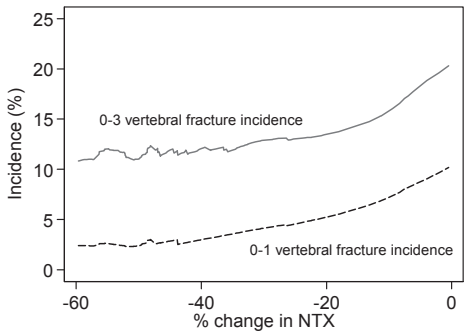

Figure 11. LOWESS smoothed curve for presence or absence of a fracture against percentage change in NTX, truncated at −60% and 0%

The above is pure speculation on my part. If this is what they did, it has no analytical value in my opinion. Cox regression is the correct analysis, but the number of fractures is too small for any but the simplest modelling. The authors themselves used Cox regression for their other analyses of fractures.

What happened next?

I checked my analysis and sent it to the BBC. Vivienne Parry was happy with it and a date for an interview was set. Then the story broke in *The Observer*. I eagerly bought a copy and there it was on the front page. However, I was not there. My analysis was pulled by the lawyers at the last minute.

In *The Observer*, it was reported that Procter and Gamble had promised openness for researchers. *File on Four* was cancelled and I thought my prospect of undying fame had vanished. Procter and Gamble threatened to sue the University of Sheffield anyway. Then, in January 2006, the BBC came back to me. This time the story was to be in the consumer programme *You and Yours*. I was interviewed by Vivienne Parry and waited for the broadcast, which finally happened in February 2006. I did not know when it would be on and was out of the country at the time. The first I knew of it was an email from the University asking whether I could provide a photograph for *The Times Higher Educational Supplement*, which carried the story. Aubrey Blumsohn left the University of Sheffield. Procter and Gamble decided not to sue and handed over the data to Aubrey Blumsohn, who is now preparing for publication with a different statistician.

By the way, you would be amazed at how many of your family, friends and colleagues will tell you that they heard you on Radio 4.

Reference

1. Eastell, R., Barton, I., Hannon, R., Chines, A., Garnero, P. and Delmas, P. (2003) Relationship of early changes in bone resorption to the reduction in fracture risk with risedronate. *Journal of Bone and Mineral Research*, **18**, 1051–1056.

Martin Bland has been Professor of Health Statistics at the University of York since 2003. Before this, he spent 31 years at St Thomas's and St George's Medical Schools, London, and 3 years in agrochemicals with ICI. With others, he is the author of two textbooks on medical statistics, 179 refereed papers and many educational articles and chapters, including the Statistics Notes series in the *British Medical Journal*. He has been secretary and chair of the Medical Section of the Royal Statistical Society.