Neutral Citation Number: [2010] EWCA Civ 346

Case No: C1/2009/0805

IN THE SUPREME COURT OF JUDICATURE
COURT OF APPEAL (CIVIL DIVISION)
ON APPEAL FROM QBD, ADMINISTRATIVE COURT
MR JUSTICE HOLMAN
CO24692008

Royal Courts of Justice
Strand, London, WC2A 2LL

Date: 31/03/2010

Before :

LORD JUSTICE PILL
LADY JUSTICE SMITH
and
LORD JUSTICE WILSON

- -

Between :

Servier Laboratories Limited	**Appellant**
- and -	
National Institute for Health and Clinical Excellence & Anr	**Respondent**

- -
- -

Clive Lewis QC & Martin Chamberlain (instructed by **Bristows**) for the **Appellant**
Michael Beloff QC & Daniel Stilitz (instructed by **Messrs Beachcroft**) for the **Respondent**

Hearing date : 17 December 2009

- -

Approved Judgment

Lady Justice Smith:

Introduction

1. The National Institute for Health and Clinical Excellence (NICE) is a special health authority within the National Health Service (NHS). Its function is to develop guidance on all aspects of healthcare within the NHS. One of the important ways in which it provides such guidance is by appraising the clinical benefits and costs of health care interventions including drugs. It makes recommendations as to which of the drugs available for a particular purpose provide the best value for money. A drug which is recommended as the treatment of choice will be prescribed by doctors in the confident expectation that the NHS will provide the necessary funding. A drug which is not so recommended will be prescribed far less frequently as funding may be refused.

2. Such a method of appraisal and recommendation is necessary in the public interest because it assists in the distribution of the limited resources of the NHS in a way which is cost effective, fair and consistent throughout the country. Plainly the decisions of NICE are of great importance to the commercial interests of drug manufacturers. If a drug is not recommended as the treatment of choice in its particular field, the sales in this country are likely to be modest.

3. This is an appeal from the order of Holman J dated 19 February 2009 when he rejected in part the application of Servier Laboratories Limited (Servier) for judicial review of the refusal of NICE to recommend Servier's drug strontium ranelate (brand named Protelos) as treatment for the prevention of osteoporotic fractures in post-menopausal women. NICE recommended as the treatment of choice alendronate, one of several drugs in a group of bisphosphonates under consideration. It recommended strontium ranelate (Protelos) only for a limited group of patients who could not tolerate alendronate. In effect, NICE was of the view that Protelos was insufficiently effective and too expensive to justify its wider use.

4. Servier sought judicial review of NICE's decision with the support of the National Osteoporosis Society contending that the decision should be reconsidered. Three grounds were advanced before the judge. He granted a review on one ground and rejected the other two. However, Servier is not content with the limited basis of the reconsideration which followed that decision and wishes it to take place on the wider basis that would result if another of its grounds of challenge were upheld. Accordingly, Servier now appeals the judge's decision in respect of one of the grounds of challenge that he rejected. This is that, during its appraisal of Protelos, NICE failed properly to take into account data derived from a post-hoc subgroup analysis of the results of a study entitled Treatment of Peripheral Osteoporosis (TROPOS). This study was published in 2005 by the Journal of Clinical Endocrinology & Metabolism (JCEM) under the heading 'Strontium Ranelate Reduces the Risk of Non-vertebral Fractures in Postmenopausal Women with Osteoporosis'. Servier claims that the post hoc analysis in this study demonstrates that Protelos is as efficacious as alendronate in the prevention of hip fractures and that NICE's refusal to accept this data was irrational and contravened its own settled procedures. Alternatively, Servier contends that NICE failed to give adequate reasons for its rejection of this data.

Osteoporosis

5. I take this description from the judge's summary of the condition. Osteoporosis is a skeletal disorder characterised by low bone mass and deterioration of bone tissue. Its consequence is an increased susceptibility to fracture. It is most commonly found in post-menopausal women. It is estimated that more than 2 million women in England and Wales currently suffer from the condition and more than one in four women will suffer from it in their lifetime. Fragility fractures occur most commonly in the vertebrae, hip and wrist. It is estimated that every year in England and Wales, there are about 180,000 osteoporosis-related fractures of which 70,000 are to the hip, 25,000 to the vertebrae and 41,000 to the wrist. Hip fractures are associated with high rates of morbidity and with increased mortality.

6. The condition is incurable and the aim of drug treatment is to reduce the risk of fracture. It will be apparent from the numbers quoted above that the market for drugs which will effectively reduce the risk of fracture will be substantial.

NICE's methodology

7. NICE responds to requests from the Secretary of State for Health to appraise a particular type of health intervention or drug. Once a topic has been chosen, NICE identifies organisations with an interest in that topic and devises an appraisal plan. An independent academic assessment is commissioned to review and evaluate the existing evidence. This body produces a Technology Assessment Report (TAR) which presents an analysis of the cost effectiveness of the technology or drug in question.

8. The interested bodies are invited to comment on the TAR. Their comments and the TAR are combined into an Evaluation Report. The appraisal is allocated to a committee comprising clinicians, health administrators, academics, representatives of the pharmaceutical industry and lay members. That appraisal committee hears evidence from a wide variety of witnesses before producing its initial recommendations in an Appraisal Consultation Document (ACD). Interested parties and the public are able to comment on this document. The committee takes those comments into account before making its final recommendations in the Final Appraisal Determination (FAD). It is something of a misnomer to call that determination 'final' at that stage, as it is still open to amendment.

9. The FAD is submitted to the NICE Guidance Executive for approval. When approved, it is circulated to the interested parties some of whom (those classed as 'consultees') have a right to appeal against the FAD. If there are no appeals or if the appeals are dismissed, the FAD is issued as NICE guidance. If there is a successful appeal, the FAD will be reconsidered and may be amended.

10. The hallmarks of NICE's methodology are that it should be thorough, scientifically reliable and transparent. Virtually all the evidence is open to public scrutiny. Occasionally commercial confidentiality prevents this but the aim is complete openness.

The factual background and the appraisal process in this case

11. NICE's appraisal of drugs for the prevention of osteoporosis in post-menopausal women began in 2002 and it produced an early report on the prevention of fractures in women who had already suffered at least one fracture. By 2004, new research was about to be published and NICE decided to recommence the exercise dealing separately with treatments recommended for women who had not yet suffered a fracture and those who already had. By this time, Protelos had been licensed for use and joined the list of drugs to be considered.

12. NICE commissioned an independent assessment of the various drugs by Sheffield University's School of Health and Related Research (ScHARR) which produced three TARs relating to different drugs, including the bisphosphonates and strontium ranelate. The TAR which related to strontium ranelate considered all the relevant published evidence which comprised three studies including the report of the TROPOS trial.

13. The TROPOS trial was a randomised control trial (RCT) conducted in 11 European countries and in Australia between 1996 and 2003. It was designed to assess the effectiveness of strontium ranelate in preventing non-vertebral fractures in post-menopausal women and to assess its tolerability. It operated by comparing the number of fractures suffered by women taking strontium ranelate with the number sustained by women taking a placebo. The study was confined to women over the age of 74 and those over 70 who had an additional risk factor for fractures. Servier's main concern at the time when the TROPOS study was begun was to obtain a licence to market Protelos in Europe and the study was designed largely to meet the requirements of the licensing body, the European Medicines Agency (EMA).

14. In 2001, during the licensing process, the EMA decided that they required evidence of the efficacy of strontium ranelate specifically in the prevention of hip fractures, as opposed to fractures at other non-vertebral sites. Analysis of the results of the first three years of the study revealed that the study was not sufficiently powered to produce statistically significant results in respect of the prevention of hip fractures. This was not a surprise as it had not been designed for that purpose. EMA suggested that Servier should identify a subgroup of women who would be at an enhanced risk of suffering a hip fracture. Some of the women in the subgroup would have been taking the drug, others the placebo. It was hoped that the sufficient women taking the placebo would have suffered a fracture as to provide a statistically significant assessment of efficacy. EMA advised Servier on and approved the selection of the subgroup. These were to be women over the age of 74 who had a low bone mineral density (BMD) measured at the neck of the femur. The number included in the subgroup was 1876 out of the 4932 included in the original study.

15. The subgroup analysis demonstrated that, within this group, women on the drug had sustained 36% fewer hip fractures than women on the placebo. That degree of efficacy compared closely with the efficacy of aledronate according to published data. The EMA accepted the TROPOS data including the post- hoc subgroup analysis and concluded that:

> "…from the efficacy viewpoint, the submitted documentation is considered sufficiently robust to support an indication for

> treatment of postmenopausal osteoporosis, to reduce the risk of vertebral and hip fractures… For this indication, the demonstrated effect of strontium ranelate 2g/day appears comparable with of bisphosphonates, and the strategy to accept a therapeutic indication based on post hoc analysis of a revised target population of particular medical interest has regulatory precedent in the European licensing of bisphosphonates."

16. Following that assessment, the EMA granted market authorisation for Protelos for the reduction of vertebral and hip fractures.

17. In January 2005, at a fairly early stage of the NICE assessment, Servier put in submissions relating to the TROPOS trial. These contended that the main TROPOS study confirmed earlier work which showed that strontium ranelate is effective at reducing the risk of vertebral fractures and also that it reduced the risk of all non-vertebral fractures by 16%. A reduction of 16% is expressed as being a relative risk (RR) of 0.84. It also relied on the post hoc subgroup analysis of the TROPOS data showing a 36% reduction in hip fractures and the EMA's acceptance of that data.

18. It appears that a draft report must have been made available to Servier in about mid-2005 from which it was apparent that the ScHARR assessment group was unimpressed by the post hoc subgroup data. In the TAR which is available to us, the subgroup data was dealt with quite briefly within paragraph 3.2.1.5.1.2 of the TAR which was headed "Assessment of effectiveness of strontium ranelate". This is a long paragraph but the relevant passage is to be found just above Table 13 where it said:

> "None of the studies were powered to identify a statistically significant difference in the incidence of fracture at any specific peripheral fracture site, and none reported a significant reduction in hip or wrist fracture in relation to its full intention-to-treat population (see Table 13 and Table 14). Although in the TROPOS study, a significant reduction in hip fracture was seen in the subgroup of women who were aged over 74 and were osteoporotic at study entry (see Table 13), it should again be born *(sic)* in mind that this is not a true randomised comparison."

19. Table 13 recorded the result of the 36% reduction in hip fractures derived from the post hoc subgroup analysis. The table recorded the result as demonstrating a relative risk of 0.64 (with a confidence interval of 0.41 to 0.98). The confidence interval shows that, if the results are reliable at all, they are statistically significant. However, the writers were clearly unimpressed with the reliability of this result because they did not consider that the subgroup analysis provided a 'true randomised comparison'. On two previous occasions in the same long paragraph, when discussing subgroup analyses taken from other studies under consideration, they had expressed their reservations about the results of such subgroup results and the reasons for them as follows:

> "…the study publications did not describe the method of randomisation: as there is therefore no reason to believe that randomisation was stratified taking any of the characteristics

> into account, none of the subgroup data are known to represent true randomised comparisons".

20. Thus, it appears that ScHARR regarded all the subgroup analyses in the studies under consideration (including the post hoc subgroup analysis in TROPOS) as essentially unreliable because the comparisons were not being made between randomly selected groups.

21. In September 2005, Servier put in further detailed submissions, seeking to deal with ScHARR's adverse comments. These stressed that the subgroup analysis had been carried out at the request and with the approval of EMA and was not a 'data mining exercise'. I will refer to these submissions in greater detail later in this judgment. However, these submissions did not apparently find favour with NICE.

22. In June 2007, the Appraisal Committee produced two FADs; one made recommendations for primary prevention (that is for patients who have not yet suffered a fragility fracture) and the other related to secondary prevention (for patients who have already had a fracture). Both FADs recommended alendronate (a bisphosphonate) as the treatment of choice and neither recommended Protelos. Broadly, the committee was of the view that Protelos was not as effective as the bisphosphonates, as well as being more expensive. In particular, in assessing the efficacy of Protelos, the FAD said at paragraph 4.1.10.2 of the FAD on primary prevention:

> "The Assessment Group reported the results of a published meta-analysis *(that is an analysis combining the results of more than one study)* that resulted in a RR for vertebral fracture of 0.60 (95% CI 0.53 to 0.69, two RCTs, n=6551) *(the confidence interval or CI showing that the results were statistically significant)* and an RR for all non-vertebral fractures including wrist fracture in the whole study population was 0.85 (95% CI 0.61 to 1.19, one RCT, n=4932) *(a statistically non-significant result)*. A post-hoc subgroup analysis in women over 74 years of age with a T-score of -2.4 SD resulted in an RR for hip fracture of 0.64 (95% CI 0.41 to 0.98, one RCT, n=1977)*(a statistically significant result)*." (My explanations in italics)

At paragraph 4.3.23, the FAD concluded:

> "The Committee did not accept the estimate of efficacy for strontium ranelate in preventing hip fracture from the post-hoc subgroup analysis, but accepted the statistically non-significant RR of 0.85 for hip fracture to acknowledge an effect on this important type of fracture. The Committee noted that strontium ranelate was dominated by alendronate (based on the price of £95.03 per year for alendronate); that is strontium ranelate has a greater acquisition cost and is not more efficacious. Therefore, the Committee did not consider strontium ranelate to be cost-effective for the initiation of therapy for the primary prevention of osteoporotic fragility fractures in menopausal women."

23. In the second FAD, the Committee used similar words to reject strontium ranelate as a cost effective therapy for secondary prevention.

24. Servier appealed both decisions on a number of grounds, which included the allegedly wrongful rejection of the post hoc subgroup TROPOS data. The Appeal Panel recorded the response of the Chairman of the Appraisal Committee to this ground of appeal as follows:

> "Professor Stevens explained that the reduction in hip fracture rate described with strontium ranelate was only found in a post-hoc analysis of a group of high risk patients that had not been pre-specified. The Appraisal Committee had allowed for this weak evidence by setting the hip fracture rate with strontium ranelate to 0.85."

25. The Appeal Panel accepted that Servier had been put into difficulty by EMA's change of stance but considered that the Appraisal Committee had evaluated the post hoc subgroup data appropriately. They rejected the appeal on that ground. However, other grounds of appeal advanced by Servier succeeded to a limited extent and, in due course, the Appraisal Committee issued revised FADs. The position of Protelos was improved to some extent in that it was recognised for use for some patients who could not tolerate alendronate.

26. Servier was still dissatisfied and judicial review proceedings were commenced in an attempt to have the FADs set aside. Although as I have said, several points were argued before Holman J, only one, the approach to the post hoc subgroup analysis, remains alive for the purposes of this appeal. In pursuing this appeal, Servier accepts that, even if the efficacy of Protelos is reassessed as comparable with alendronate, it cannot hope that its product will be recommended as the treatment of first choice for the prevention of fractures. That is because alendronate, being available in generic form, is considerably cheaper. Protelos is still covered by a patent. Nonetheless, Servier contends that the appeal is important because, if Protelos is accepted as being as effective as alendronate, the terms in which the various drugs are recommended is bound to be more favourable to Protelos than under the present determination. In short, the appeal is not academic.

The decision of Holman J

27. Before the judge, Servier argued that the NICE Appraisal Committee ought not to have taken a different view of the efficacy of strontium ranelate from that taken by EMA. It accepted that the functions of EMA and NICE were different; NICE was assessing cost effectiveness whereas EMA was assessing only efficacy and tolerability. But both bodies had to assess efficacy and it was unsatisfactory if two equally eminent bodies came to different conclusions on so important a point. The Scottish Medicines Consortium (the Scottish equivalent of NICE) had reached the same conclusion on efficacy as EMA. Servier did not suggest that NICE was not entitled to come to a different conclusion from EMA, only that it should not differ without good reason and without clearly explaining its reasons.

28. Holman J expressed the view that NICE was in no way bound by the decision of EMA and said that it was entirely a matter for NICE how much weight it attached to any particular piece of evidence.

29. Servier also contended that the Appraisal Committee had not provided adequate reasons for its decision to be understood. It had not explained why it did not accept the post hoc subgroup analysis or why it had taken a different view of that evidence from EMA. Further, NICE's approach was irrational in that its reasoning did not make sense. It had rejected a statistically significant result but had accepted one which was not statistically significant. The judge rejected the arguments, saying that NICE had given a reason for regarding the hip fracture data as 'weak' evidence; the reason was that the data was from a post hoc subgroup analysis and there were inherent weaknesses in such evidence. Also, there was nothing irrational about NICE's reasoning.

The appeal to this Court

30. Permission to appeal was granted by Jacob and Patten LJJ on limited grounds. Although the grounds were expressed in different ways, the issues are of narrow compass. The questions which arise on the appeal are, first, whether NICE adequately explained its reasons for rejecting the post hoc subgroup data, particularly in the light of the fact that the same data had been accepted by the EMA, an equally distinguished and authoritative body. Second, if the reasons were adequately expressed, was the rejection of that data and NICE's assessment of the efficacy at RR 0.85 rational?

Adequacy of the Reasons

31. Servier submitted that the Appraisal Committee had given no reason at all for rejecting the post hoc subgroup data. At paragraph 22 above, I have quoted paragraph 4.2.23 of the FAD. It stated simply that the Appraisal Committee did not accept the post hoc subgroup data. I accept that no reason was given in the FAD. However, I have also set out the observations of the ScHARR group (see paragraphs 18 and 19 above) from which it is apparent that ScHARR was unimpressed by the data because it considered that the sub-group analysis was not based upon 'a true randomised comparison'. It appeared to me at first sight that it might be assumed that the Appraisal Committee had adopted the opinion of the ScHARR group and that the unspoken reason why it did not accept the subgroup data was that it was not based on a true randomised comparison. However, it is to be noted that the committee did not expressly adopt ScHARR's opinion or even refer to it.

32. Before the judge, NICE put in evidence explaining its thinking. Holman J accepted this evidence as a permissible explanation of the Committee's thinking at the time of the decision as opposed to an impermissible ex post facto rationalisation of its stance.

33. Starting at paragraph 62 of her witness statement, Dr Elizabeth George, the Associate Director of NICE's appraisal programme, explained that there are inherent weaknesses in results derived from a subgroup which had not been identified at the outset of the study but only after the results were known. She said that that in itself meant that there was an increased risk of bias in the results. She then said at paragraph 62:

> "From all discussions amongst experts that I have been part of, …. there is no biological plausibility that any of the drugs under appraisal should be more efficacious in older women (in this case women over 74) than in younger women. Furthermore we have never been presented by Servier with data for other age groups. The logical conclusion of the acceptance that strontium ranelate is more effective in the 74+ age group slot is that it is less effective in women who are younger than 74."

34. If that passage from Dr George's evidence was intended to suggest that the reason or one of the reasons for the Committee's refusal to accept the post hoc results was that they were biologically implausible, I feel bound to observe that, although Holman J accepted that evidence as an explanation of what had been in the minds of the Committee, there is no reference to this argument in either the TAR or the FAD. In any event, Servier contends that if that was part of the reasoning, it was irrational and wrong.

35. At her paragraph 64, Dr George emphasised the inherent weakness of post hoc results while responding to an allegation by Servier that, by refusing to accept its post hoc results, NICE had not complied with its own guidance. NICE publishes guidance to parties who are to submit evidence as to the value and reliability which is likely to be placed on different forms of scientific evidence. At the top of the hierarchy is the randomised controlled trial (RCT). As I have said, the TROPOS trial was an RCT. But Dr George pointed out that "non-pre-specified subgroup analyses are not normally considered as high level evidence by the Appraisal Committee (or anyone else for that matter)". She said that they were regarded as exploratory analyses which should be followed up with a pre-specified analysis. So, Dr George was saying that it was a mistake for Servier to claim that it was relying on an RCT, which was at the top of the hierarchy of evidence because it was not doing so; it was relying on a post hoc subgroup analysis derived from the RCT which everyone knows is not the same thing as the RCT itself. The hierarchy table in the Methods Guide draws a distinction between RCTs with a very low risk of bias which are at the top of the list, RCTs with a low risk of bias which are acceptable and RCTs with a high risk of bias which, according to a footnote to the table, should not be used as a basis for making a recommendation. I observe that nowhere in the papers I have seen has NICE specified whether it regards this post hoc subgroup analysis to be low risk or high risk. However, it must be inferred from the fact that it described the results as 'weak evidence' that it regarded the analysis as carrying considerable risk of bias. But if that was its view, no reason for that view was given.

36. Dr George then went on to quote a passage from the Methods Guide which sets out requirements to be complied with if subgroup analysis is to be offered as evidence. This said:

> "There should be a clear clinical justification and, where appropriate, biological plausibility for the definition of the patient subgroup and the expectation of a differential impact. Ad hoc data mining in search of significant subgroup effects should be avoided. Care should be taken to specify how subgroup analyses were undertaken, including the choice of

> scale on which effect modification is defined. The precision of all subgroup estimates should be reflected in the analysis of parameter uncertainty. The characteristics of the patients associated with the subgroups presented should be clearly specified to allow the Appraisal Committee to judge the appropriateness of the analysis with regard to the decision problem."

37. It seems to me that Dr George's evidence shows clearly that post hoc sub group data carries a risk of bias and that NICE is entitled to expect particular explanations and justifications for the submission of and reliance on such data. However, Dr George does not say in what respects Servier has failed to comply with those requirements or why NICE rejected this particular subgroup analysis, which it knew had been accepted as 'sufficiently robust' by EMA. The evidence only explains that NICE rejected the evidence because it was derived from a post hoc subgroup analysis.

38. The second item of evidence relied on by NICE came from Professor Andrew Stevens the chairman of the Appraisal Committee. He first made the point that it was clear from the FAD that the committee had considered the post hoc subgroup evidence. However, it had not accepted it. At paragraph 23, he explained how the committee had estimated the efficacy of strontium ranelate, which had been, in effect, to take a broad brush to the totality of the available evidence. I will consider the rationality of that approach later. However, in paragraph 24, Professor Stevens explained why the committee did not accept Servier's submission. He acknowledged that the results of RCTs are the preferred form of evidence, as compared for example with observational data. But, he continued:

> "But what the Claimant fails to notice is that results for the analysis of post-hoc subgroups of randomised trial are not the same as randomised evidence. The point is a very elementary one. A randomised controlled trial may generate high quality data. It does not follow that any subsequent selective manipulation of that data must be of equivalently high quality. The Methods Guide is clear on this even if the Claimant's selective quoting of it is not."

39. Here again, as with Dr George's evidence, Professor Stevens has explained that a post hoc subgroup analysis does not necessarily provide high quality evidence. Indeed, there are reasons why it will not do so. Servier does not dispute that. But, as Mr Clive Lewis QC for Servier pointed out, Professor Stevens does not go so far as to say that such an analysis never provides good quality evidence and Servier's complaint is that the Committee has not explained why it would not accept this particular subgroup analysis. It has not explained what was wrong with this subgroup selection, what Servier had failed to explain, what information it had failed to provide or in what way it was suggested that the selection of the subgroup amounted to data mining.

40. As I have said, Servier accepted that there are reasons why a post hoc subgroup analysis might be less reliable than the results of a pre-designed randomised controlled study. One of these is that the subgroup has been chosen in the hope of obtaining a particular result. This practice is known as 'data mining' and is recognised as giving rise to unreliable results. But, submitted Servier, NICE has not suggested

that this subgroup selection amounted to data mining. Indeed, the subgroup had been selected in consultation with EMA and it would be remarkable if so eminent and responsible a body had suggested and then accepted a group selection that would be open to so obvious an objection as data mining. In any event, Servier had gone to great lengths in its submissions (in particular in September 2005) to explain how and why it had selected its subgroup. If NICE rejected those explanations, it could and should have said so and explained why. Servier accepted that NICE should not be expected to deal with every argument raised by every participant; that would be too onerous. But, it submitted that a short explanation for the rejection of a piece of evidence which lay at the heart of the assessment of a particular drug was not too much to ask; indeed it was obligatory.

41. Mr Michael Beloff QC for NICE submitted that, in truth, Servier well understood why this data had been rejected. He pointed to the submissions of September 2005 as showing that Servier was aware at that time that ScHARR was unimpressed by the evidence because patients within the subgroup had not been properly randomised. I accept that it is clear from Servier's submissions that at that time they understood that that was ScHARR's stance. However, the September 2005 submissions dealt with that point in some detail. I do not wish to burden this judgment with a prolonged exposition of these submissions. Suffice it to say that, in addition to reminding NICE that the EMA had been satisfied with the robustness of the data and that the results of the post hoc study had been published in a reputable peer-reviewed journal, Servier explained its contention that the subgroups had in fact been randomised. They had been well balanced for baseline characteristics. In short, it was taking on ScHARR's objection and dealing with it by specific reference to a table analysing the baseline characteristics in the whole study population and in the subgroup.

42. That direct rebuttal of ScHARR's reason for non-acceptance was followed by a detailed explanation of the thinking behind the selection of the subgroup, which seems to have been designed to answer any implied suggestion that the selection had been a data mining exercise.

43. I noted earlier that the FAD did not expressly adopt ScHARR's reason for rejecting the subgroup data (that the subgroup was not randomised). It did not refer to Servier's attempt to rebut that reason, which one might have expected if it considered the rebuttal to be invalid. Nor did the Committee give any reason of its own in the FAD. I find it strange and unsatisfactory that, when ScHARR had advanced one reason for rejecting the subgroup data, the Appraisal Committee should neither adopt that reason nor reject it and give another. It rather looks as though the Committee felt that it could not justify the adoption of ScHARR's reason but did not have another reason to rely on. So no reason was given. It is to my mind significant also that neither Dr George nor Professor Stevens deals with the issue of inadequate randomisation. As I have pointed out, their evidence goes only to explain why post hoc subgroup studies may be unreliable and does not deal with the reasons why this particular study was unreliable.

44. Holman J was content to accept that the generic reason was enough – post hoc results have inherent weaknesses and it is entirely a matter for the Appraisal Committee how much weight they give to a particular piece of evidence. The judge adopted an analogy with the approach that a judge might take to hearsay evidence. There are reasons why hearsay evidence may be unreliable; for one thing it cannot be subjected

to cross-examination. It may however, be sufficiently reliable for the judge to accept it and act upon it even though it is hearsay. It is entirely a matter for the judge how much reliance he places on it. I accept that the judge's analogy is a good one. Post hoc analyses may be unreliable just as hearsay evidence may be. But, in an individual case, both may provide good reliable evidence. A judge may not simply reject a piece of hearsay evidence as worthless without explaining why he has done so. That is particularly so where the hearsay evidence is central to a party's case. Nor in my view, should a decision maker reject a post hoc subgroup analysis without explaining where its weakness lies.

45. In my view, this post hoc subgroup analysis was central to Servier's case in support of strontium ranelate. As such, it might be unreliable but is not necessarily so. ScHARR gave a reason for rejecting it which Servier sought to rebut. The Appraisal Committee did not adopt ScHARR's reason; nor did it mention Servier's rebuttal. Nor in the FAD did it offer any reason of its own for non-acceptance. In these proceedings, the main reason advanced has been the alleged 'inherent weakness' of all post hoc subgroup data. Yet, on its own evidence, such data may be reliable. The other reason advanced was that the results of the analysis were biologically implausible. Yet, as I have said, that argument was not mentioned in either the TAR or the FAD. I have come to the conclusion that the rejection by NICE of the Servier post hoc subgroup analysis is inadequately explained. I cannot tell what its reasons are and I accept Servier's claim that it cannot either.

46. For that reason, I would allow this appeal. I have considered whether NICE should be permitted to advance further reasons for its original decision or should be required to take a fresh decision. I have concluded that the latter course is preferable. Quite apart from the inherent danger that the decision maker will use that as an opportunity to justify what may be a flawed decision, it seems to me that there is a good reason why, in this individual case, it would be preferable for NICE to make a fresh decision. The reason is that, even though I do not fully understand the reasons for NICE's decision, I have grave doubts about its rationality.

Rationality

47. I am reluctant to embark on a detailed consideration of the rationality of the various reasons which have been canvassed before us. I am however, prepared to make some observations (necessarily *obiter*) which I hope might be of assistance to the parties and in particular to NICE in the future conduct of this dispute.

48. On the face of the decision as explained in the FAD, the only reason given for the rejection of the data is that it came from a post hoc subgroup. The implication is that that class of scientific evidence is inherently unreliable. In my view, that reason simpliciter is not rational. The evidence of NICE itself is to the effect that such data may in some circumstances be acceptable and reliable. Therefore, if such data is to be rejected, the reason for rejection must relate to the particular study. If the evidence is considered to be weak, NICE must explain why it is of that view. Various potential reasons for taking that view have been canvassed.

49. It could be rational to reject this data on the ground that selection of the subgroup amounted to data mining. Of course, that is denied and a full explanation has been offered as to why the selection was made as it was. That explanation must be

considered and rejected if data mining selection is to be relied on. It is fair for Servier to comment that it would be surprising if EMA had proposed a selection which amounted to a data mining exercise.

50. It could be rational to reject this data on the ground that the patients in the subgroup were not randomised. That is denied and the information required for the assessment of that denial was provided in September 2005 and never commented upon. That information must be considered and rejected if that reason is to be relied on. Here again, it is fair for Servier to comment that it would be surprising if EMA had accepted results flawed in that way.

51. It could be rational to reject this data on the ground that the results are biologically implausible. That is denied and the arguments have been aired in the proceedings for judicial review and on this appeal. Servier never had the opportunity to deal with NICE's argument on this point before the final decision was taken as the issue was raised for the first time during the proceedings for judicial review. I do not propose to discuss the detail of the arguments on both sides. Suffice it to say that, it seems to me that there is a serious issue between the parties on this and I am by no means convinced at the present time that reliance on biological implausibility would be a rational reason for rejecting the subgroup data. Once again, if biological implausibility this is a valid concern, one might have expected EMA to think of it.

52. By making these references to EMA I am not, of course, suggesting that NICE is bound by a prior decision of that body. However, I would expect to see some reason given for NICE reaching a different view from a body of similar standing.

53. In all the circumstances, I consider that the right conclusion is that NICE should be required to make a fresh decision.

54. Finally, I wish to mention Servier's argument that NICE's overall assessment of the efficacy of Protelos was irrational because it relied on results which were not statistically significant (those of the whole TROPOS trial) in preference to the results of the subgroup analysis which were statistically significant. NICE's answer to this contention depends upon its entitlement to regard the results of the subgroup analysis as unreliable. If they were so entitled, then it seems to me that the rationality of their overall assessment could not be challenged. In my view, it is not irrational to rely on a statistically non-significant result of a high quality RCT in preference to the statistically significant result of a fundamentally unreliable study. So, if NICE was entitled to reject the post hoc subgroup data as unreliable, (even though statistically significant) its overall assessment was rational. Indeed one might even say that it was generous to Servier because although Servier had demonstrated efficacy in non-vertebral fractures to the standard of statistical significance in the full RCT population, it had not proved done so for hip fractures in the same population.

55. That said, I would allow the appeal.

Lord Justice Wilson :

56. I agree.

Lord Justice Pill:

57. In her judgment, Smith LJ has set out the background to the present claim and the issues raised and I gratefully adopt her narrative. Smith LJ has referred to the contemporaneous documents dealing with strontium ranelate and to the post-decision statements submitted in justification of the decision by Professor Stevens and Dr. George.

58. In seeking to uphold the decision of NICE, Mr Beloff QC submitted that the fairness of the procedure is ensured by the elaborate processes through which an application passes. Judgments on medical and scientific matters are required and the court should be more deferential than in other contexts when considering decisions taken.

59. I agree with the conclusions of Smith LJ and with her reasons. The court must be prepared to analyse the decision making process and the reasoning involved. The central flaw in the decision making progress is the absence of a satisfactory explanation as to why the post-hoc subgroup analysis on which the appellants relied has been rejected. The criteria adopted are set out in paragraph 14 of the judgment of Smith LJ.

60. I refer to the approach of EMA. In making its decisions, EMA is assisted by its scientific Committee for Medicinal Products for Human Use (CHMP). Assessment of the product by CHMP started in July 2003. The results of the trial were published in the Journal of Clinical Endocrinology & Metabolism in February 2005:

> "In the subgroup analysis of women at higher risk of hip fracture, those aged 74 yr or over and with femoral neck BMD T-score of -3 SD or less, the risk of hip fractures was reduced by 36%. These high-risk patients were defined according to risk factors for hip fracture. Seventy-four years, which was the main age criterion for inclusion in the study was reported to be the age starting from which incidence of hip fracture rises exponentially. This has been confirmed in the placebo group in the pooled data from SOTI and TROPOS studies, which showed that the incidence of hip fracture was 1.1% over 3 yr in patients less than 74 yr old and 4.4% in patients with age 74 yr or older."

61. There followed scientific discussion of the testing in the European Public Assessment Report. The post-hoc subset analysis was considered. Under the heading "overall conclusions benefit/risk assessments and recommendation", it was accepted that an indication partly based on post-hoc analysis of revised target population of particular medical interest has regulatory precedent in the European licensing of bisphosphonates. It was stated:

> "The comprehensive clinical programme, and especially the contribution of data in the elderly and very elderly is acknowledged. From the efficacy viewpoint, the submitted documentation is considered sufficiently robust to support an indication for treatment of postmenopausal osteoporosis, to reduce the risk of vertebral and hip fractures . . ."

62. It is not suggested that NICE are bound by EMA's decision or its reasoning but the appellants are entitled to expect any decision against them to be properly reasoned, especially when it is contrary to the reasoned decision of an equally eminent body.

63. On the face of it, it is not unreasonable to choose a subgroup of those at higher risk, as EMA advised, when considering the risk of fracture which is the aim of drug treatment. If there is a reason for not doing so, it has not been explained by the respondents. The number of women included in the subgroup was 1876, about 40% of these in the original study, the data from which was retained. Over half the fractures sought to be prevented are fractures to the hip.

64. The NICE Committee in its appraisal of June 2008, at paragraph 4.3.27, "did not accept the estimate of efficacy for strontium ranelate in preventing hip fracture from the post-hoc subgroup analysis" but no reason is given. Analysis of the high risk group is not on the face of it unreasonable when considering the efficacy of a drug. To take an extreme case, a study of the clinical effectiveness of a drug in combating a condition which affects only the 40s, is unlikely to be assisted by a study of people under 40.

65. NICE's guide to the Methods of Technological Appraisal (April 2004) does not involve a blanket ban on post-hoc subgroup studies. In section 5.9.5 the value of subgroup analysis is accepted though it is right to say that the stated context is the capacity for patients with different characteristics to benefit from treatment. The guide provides, at paragraph 5.9.5.2:

> "There should be a clear clinical justification and, where appropriate, biological plausibility for the definition of the patient subgroup and the expectation of a differential effect. Ad hoc data mining in search of significant subgroup effects should be avoided. Care should be taken to specify how subgroup analyses were undertaken, including the choice of scale on which effect modification is defined. The precision of all subgroup estimates should be reflected in the analysis of parameter uncertainty. The characteristics of the patients associated with the subgroups presented should be clearly specified to allow the Appraisal Committee to judge the appropriateness of the analysis with regard to the decision problem."

66. While recognising the dangers inherent in subgroup analysis, the guide acknowledges that, subject to stringent safeguards, such an analysis may be of value. In this case, a reason, which on the face of it is a good one, has been given for the definition of the patient subgroup and the characteristics of the patients associated with the subgroup presented have been clearly specified. NICE cannot in my judgment argue, as they sought to do before this court, that any subgroup analysis is suspect and that is the end of it.

67. Late in the argument before the court, there was a suggestion that the post hoc analysis was unreliable because it did not take into account factors such as smoking and drinking. That was not mentioned during the decision making process and no

thought appears to have been given to the relevance of such factors in any part of the research process.

68. In her statement, Dr George states, at paragraph 62:

> "From all discussions amongst experts that I have been part of, including discussion of the Committee and the GDG, there is no biological plausibility that any of the drugs under appraisal should be more efficacious in older women (in this case women over 74) than in younger women."

A bare assertion, in a post-decision statement, about biological plausibility does not assist in saving the decision. Moreover, and with respect, it appears to miss the point made by the appellants and by EMA. The subgroup study is said to be valuable because it deals with the efficacy of the product in the age group which is shown, statistically, to be at much greater risk. Because of the dramatic increase in the number of hip fractures in women aged 74 or more, results obtained from testing the product on these women are more relevant to the efficacy, it is submitted, and also more reliable statistically than tests on a comparable number of women for whom the risk of fracture is low.

69. I find it impossible to comment on the rationality of a decision reached as this one has been reached. I mention only, because of Smith LJ's reference to it at paragraph 54, the value of the evidence, stated not to be statistically significant, on which NICE did purport to rely. Comment in advance on what reliance can be placed on that evidence is in my view inappropriate. On a reappraisal, NICE will, I am confident, consider the efficacy of this drug in a comprehensive and open-minded way.

70. I have added these few comments in support of the comprehensive reasoning of Smith LJ. I too would allow the appeal and quash the decision.