COMMENTS ON APPRAISAL OF STRONTIUM RANELATE FOR THE PREVENTION OF OSTEOPOROTIC FRAGILITY FRACTURES IN POSTMENOPAUSAL WOMEN

REPORT BY THE DECISION SUPPORT UNIT

September, 2010

Professor Keith Abrams[1], Professor Sheila Bird[2], Professor Stephen Evans[3] & Professor Gordon Murray[4]

[1] Department of Health Sciences, University of Leicester, Leicester, UK & Decision Support Unit, UK
[2] MRC Biostatistics Unit, Cambridge, UK
[3] Department of Medical Statistics, London School of Hygiene & Tropical Medicine, UK
[4] Centre for Population Health Sciences, University of Edinburgh, Edinburgh, UK

Decision Support Unit
ScHARR
University of Sheffield
Regent Court
30 Regent Street
Sheffield
S1 4DA

EXECUTIVE SUMMARY

The methodology used by Servier to identify their specific subgroup is inappropriate, meaning that there is a very high risk that their estimated subgroup RR exaggerates the benefit of Strontium Ranelate (SR) in reducing hip fracture risk. This is because factor-specific cut-points were used to define the Servier subgroup in such a way as to maximize hip fracture risk in placebo-treated patients. The use of a subgroup by the EMA for its purposes may be valid, though the one defined by Servier differed to that pre-specified by the EMA, but for the purposes of NICE- that is estimating the magnitude of effect - it is clearly incorrect.

Even the fact that Servier was exploring subgroup analyses is based on a power fallacy and is a wrong *post hoc* argument, i.e. that a subgroup analysis would be expected to yield a more powerful test of the impact of SR on hip fracture risk than an analysis of the entire TROPOS trial sample.

In the absence of statistical evidence of differential efficacy, coming from a test of interaction, the most appropriate estimate of the impact of SR on hip fracture is the Relative Risk (RR) estimated from the *entire* TROPOS trial sample (or from a meta-analysis of TROPOS and any other relevant unbiased data from trials).

INTRODUCTION

BACKGROUND

The National Institute for Health and Clinical Excellence (NICE) has issued guidance on the use of a number of pharmaceutical products (alendronate, etidronate, risedronate, raloxifene and strontium ranelate) for the primary (TA 160) and secondary (TA 161) prevention of osteoporotic fragility fractures in post menopausal women.

The technical issue to be addressed in this report is a reconsideration of the relative effectiveness of strontium ranelate in preventing hip fractures in the population covered by its marketing authorisation and which is considered by NICE to be its *de facto* target population. Strontium ranelate has an EU (and therefore UK) marketing authorisation for the treatment of postmenopausal osteoporosis to reduce the risk of vertebral and hip fractures. It is not recommended in patients with severe renal impairment and should be used with caution in patients at increased risk of venous thromboembolism (see Section 4.4 of SPC).

Servier, the manufacturer of strontium ranelate, which is one of the drugs appraised in TA160/161, applied for a judicial review following publication of the guidance. One of the points raised was related to the approach taken by the Appraisal Committee in their consideration of a particular subgroup analysis. The point was not upheld at the Judicial Review stage. Servier then applied to the Court of Appeal to challenge the ruling on the subgroup analysis and a hearing was held in December 2009.

In developing the recommendations for TA160/1 the Appraisal Committee used - in respect of hip fracture - a relative risk (RR) of 0.85 with a 95% confidence interval (CI) of 0.61 to 1.19 for strontium ranelate from the overall study population in the TROPOS trial. Servier contends that NICE should have used a RR for hip fracture from a subgroup analysis in women whose eligibility was determined by i) being aged 74 years and over, and ii) having femoral BMD T-score ≤ -3 (≤-2.4 NHANES III) but iii) irrespective of prior fragility fracture. This subgroup, which differs from that originally pre-specified by the European Medicines Agency (EMA);

> *"The CPMP asked the applicant to present data also for the subset with established osteoporosis (i.e. BMD T-score <-2.5 and prevalent fragility fracture)",*

was nonetheless accepted by EMA as a *post hoc* subgroup analysis which demonstrated efficacy in the reduction of hip fracture in a different target population than had been recruited to Servier's overall TROPOS trial. EMA had requested additional analyses for a) the Servier subgroup at four years – RR = 0.69 (95% confidence interval: 0.467 to 1.032) (in addition to the original three years for which, as above, RR was 0.64 (95% CI: 0.41 to 1.00)) and b) the subgroup defined by ii) only, for which RR over three years was 0.70 (95% CI: 0.47 to 1.04) before EMA accepted Servier's *post hoc* subgroup analyses as support of the granted indication for prevention of hip fracture as part of the of the marketing authorisation for strontium ranelate, see EPAR, Scientific Discussion, top of page 18;

> *"To this end, the applicant presented post hoc subset analyses at three years for a revised target population aged ≥74 years and with femoral neck BMD T-score ≤-3 SD (≤-2.4 SD NHANES III), for which efficacy of the same order of magnitude as shown for bisphosphonates is indicated. This has now been further supported by consistent*

> *risk reduction estimates from four-year follow-up and from the whole TROPOS population meeting the specified BMD criteria."*

Servier claimed that;

- NICE had not adequately explained its reasons for not adopting the RR based on Servier's *post hoc* subgroup, particularly as the same data had been accepted by the EMA *in respect of a revised target population* (the 'reasons' ground).
- the Appraisal Committee's decision to reject the *post hoc* subgroup data was not rational (the 'rationality' ground).

Servier's arguments are closely linked/make reference to the European Public Assessment Report (EPAR).

The EPAR states (Scientific Discussion, top of page 18) that - because the hip fracture efficacy for strontium ranelate was not demonstrated in the overall data, that is, strontium ranelate did not show a statistically significant effect (RR = 0.85 [0.61 to 1.19]) in the TROPOS trial - the EMA requested a specific subgroup analysis;

> *"The CPMP asked the applicant to present data also for the subset with established osteoporosis (i.e. BMD T-score <-2.5 and prevalent fragility fracture)",*

Servier provided a different subgroup analysis (RR = 0.64, [0.412 to 0.997]), and this was used by the EMA, after requesting further subgroup analyses (EPAR, Scientific Discussion, bottom of page 16), to determine that strontium ranelate had a positive benefit-risk profile for the indication as eventually specified in the SPC (Section 4.1). As a result of the EMA's acceptance of the Servier *post hoc* subgroup analysis as demonstrating hip fracture efficacy in a different target population (see above), Servier argues that NICE should use the RR from the Servier subgroup analysis as a basis for the decision making on cost-effectiveness for the entire population for which there is UK/EU market authorisation.

The Court of Appeal has ruled in favour of Servier on the 'reasons' ground, and the judges ordered NICE to reach a fresh decision and issue fresh guidance in respect of strontium ranelate. The Appraisal Committee will therefore reconsider the relative effectiveness of strontium ranelate and, if they consider appropriate, review their consideration of its cost effectiveness.

The key issue is whether the use of the Servier subgroup estimate of hip fracture RR at 3 years in the TROPOS trial was a more appropriate estimate with which to populate the economic decision model used by NICE. This model was developed in order to arrive at the guidance contained in TA 160/1 for the treatment of osteoporosis in postmenopausal women to reduce the risk of vertebral and hip fractures, i.e. the UK/EU marketing authorisation.

Servier had originally argued on the basis that the European Medicines Agency (EMA) had accepted the Servier *post hoc* subgroup estimate as evidence of efficacy in a licensing application. However, in their response to the 'Statement of Reasons' by NICE, Servier acknowledge that extrapolation from their defined subgroup to a broader population is challenging (see page 5), and that the consideration of the subgroup, rather than the broader population, might be a basis for cost-effective administration of strontium ranelate because different aspects of information/evidence are provided in the two cases. Whether the difference justifies limited guidance is a more complex issue (requiring an appropriately formulated and populated economic decision model), but Servier's recognition that separate cost-effectiveness considerations would, in general, be needed for a subgroup versus broader population is welcomed. These issues will be discussed by us in addressing the three questions raised by NICE.

We preface our remarks by correcting a potentially mis-leading statement made by NICE in the Court of Appeal documents and based on the Assessment Report produced by ScHARR (see page 44 of Stevenson et al. 2005) to the effect that the use of the Servier subgroup was inappropriate because the subgroup was "not properly randomized". The subgroup was properly randomized as regards an Intention-to-Treat (ITT) analysis, though not when a Per Protocol (PP) analysis using levels of strontium ranelate to define eligible patients is adopted. We agree with Servier on this point. It is the notion that a post hoc sub-group's result can be applied directly to a whole population that we do not accept and we discuss our reasons for this below.

QUESTIONS TO BE ADDRESSED

The questions to be addressed in this report are;

1. How scientifically valid is the proposition put forward by Servier related to the use of data derived from the TROPOS subgroup analysis?
2. From a statistical viewpoint, what is the most appropriate approach to the use of data from the whole data set of the TROPOS study and the subgroup data set in relation to determining the relative effectiveness of strontium ranelate?
3. Given the data reviewed what, in their expert view, is the most plausible relative risk for strontium ranelate to use in making recommendations for the population covered by the marketing authorisation for strontium ranelate?

QUESTION 1

> *How scientifically valid is the proposition put forward by Servier related to the use of data derived from the TROPOS study subgroup analysis?*

In the context of accruing evidence of efficacy with respect to the reduction of vertebral (or hip) fractures, it could be argued that it is reasonable to accept evidence of *efficacy* in a population subgroup if merely support for *an* effect is deemed sufficient. Much less clear is whether the magnitude of effect from such a sub-group is well determined, and transferable between target populations.

Subgroup analyses are inherently difficult to interpret. There are problems because multiple examinations of data distort the statistical significance tests commonly used to assess reliability of conclusions; their *post hoc* nature allows for judicious choices to obtain a desired result; biological plausibility for true differences in effect between subgroups rarely exists and has only been proven for subgroups based on specific biochemical or genetic markers; and their appropriateness for making decisions is very dependent on the context. These issues are acknowledged by the EMA itself in a recent document calling for Guidance on sub-group analysis (EMA/CHMP/EWP/117211/2010). They state specifically that "estimated treatment effects in subgroups can be unreliable". They also say;

> *"A common mis-use of subgroup analysis is to rescue a trial which, formally fails based on the pre-specified primary analysis in the full analysis set. Concerns with this strategy and factors which determine the limited scenarios where exceptions might be made will be explained. Subgroup analyses are also used in positive trials to identify groups where benefit-risk is improved compared to the full analysis set, in particular where benefit is estimated to be higher, or to make additional label claims in addition to those made on the full analysis set."*

We would accept that there is evidence that strontium ranelate does have some efficacy in reduction of hip fracture. This is partly a biological plausibility argument: that there is good evidence on vertebral fracture and some suggestion from the full analysis of the trial that it has efficacy in hips. However, the key question for a NICE appraisal is **the magnitude** of that effect. This is in contrast to the EMA assessment which requires positive evidence of efficacy**,** but actual magnitude is not the key point.

A crucial aspect for the analysis of subgroups is the question as to whether the effect size for one sub-population actually differs from another. The appropriate statistical method used to assess this is called a 'test for interaction' (Assmann *et al*, 2000). Estimation and hypothesis testing of the effect in separate subgroups can be mis-leading – a simulation study for example found that when there is no overall effect, between 7% and 21% of subgroup-specific effects were false positives (Brookes *et al*, 2004). No test for interaction was performed by Servier to ascertain if there was evidence that the RR for hip fracture in its sub-group differed from the RR for patients who were excluded from the subgroup. Nor was a test for interaction performed across all possible subgroups defined either using Servier's two risk factors (age and BMD with associated cut-points) or that prescribed by the EMA. In fact, using the three factors that Servier initially investigated (age, BMD and prior fragility fracture) there were 7 potentially defined subgroups, and evidence for the majority of these has not been presented by Servier (see below). Servier did not supply the data which we requested in the clarification questions to enable us to do such a test. However, it is fairly clear that the number of fractures in the counterpart sub-group was so small that there would be no statistical evidence for genuine differences in effect. In fact, in Servier's response to the clarification document, Servier did not supply an estimate of the RR and 95% CI (or any summary level data) for those women in TROPOS who were < 74 years of age or had a BMD T-score of > -3 (> -2.4 NHANES III), citing that;

> *"due to the low frequency of fractures in the requested subgroups, such information would be unable to provide reliable or robust estimates for the efficacy of strontium ranelate".*

The Servier response makes it clear that they chose the cut-off boundaries for i) age and ii) femoral BMD T-score because, for the placebo patients in TROPOS, the chosen cut-offs were "the most discriminating" in terms of placebo risk for hip fracture. [Clarification response 15th September 2010 page 6, line 2]. The clarification response also suggests that in fact the pooled data from the TROPOS and SOTI trials were used to identify this cut-point, which was then applied solely to TROPOS trial data. The consequence of this is that when you choose a point with the highest rate of hip fracture in the placebo group, it will automatically make it likely that the difference between the placebo and treated groups will be biased (Yusuf *et al*, 1991). The result is an artificially inflated estimate of the true RR of hip fracture between strontium ranelate and placebo patients in the so-chosen high-placebo-risk subgroup. The table in Servier's clarification response (page 6) illustrates that a cut-point of ≥ 74 produces a local maximum – the RR being 3.94 if the cut-point was ≥ 73, 4.27 ≥ 74 and 3.39 if the cut-point was ≥ 75. (Note that the RRs here refer to comparisons between subgroups of placebo patients defined by the cut-point and not between Strontium Ranelate and placebo). The clarification response also clearly states that the analysis was *restricted* to identifying a cut-point between 70 and 80 years old – cut-points beyond 80 in fact produce maximal discrimination between subgroups, though this nevertheless suffers from the same problems identified above. Indeed epidemiological evidence cited by Servier (in their Response to 'Statement of Reasons' by NICE) (Donaldson *et al*, 1990) provides evidence of a 'threshold effect' at ≥ 75 and not 74 years of age. It is important to note the difference between the definitions of the age groups as reported by Donaldson and cited by Servier as justification for their choice of threshold, and the actual threshold employed.

In addition, when asked by us (via NICE) to supply RRs or summary data for 6 further subgroups (defined by combinations of age, BMD or prior fragility fracture), Servier presented RRs for only two of the three marginal component factors, i.e. women aged ≥ 74 years (RR 0.73, 95% CI: 0.498 to 1.059) and for women with a BMD ≤ -3 (≤ -2.4 NHANES III) (RR 0.70, 95% CI: 0.473 to 1.041), citing that;

> *"the very small numbers of patients in the other subgroups (defined by combinations of age, BMD or prior fragility fracture) made it highly unlikely that reliable or robust estimates would be obtained from such analyses."*

Servier have argued that, since EMA accepted its *post hoc* subgroup analysis for licensing purposes, NICE at least needs good reasons not to use it also. However, the roles of the EMA and NICE differ. Whilst EMA was interested in addressing the harm-benefit profile of strontium ranelate, NICE is concerned with making population-based decisions based on cost effectiveness. The subgroup analysis presented by Servier together with the further analyses that EMA required (which evidenced some shrinkage of the estimated RR, for example the Servier subgroup at 4 years produced a RR of 0.69, as well as Servier's response to our questions of clarification above) may indeed address the former - by at least showing a statistically significant benefit in terms of hip fracture for a sub-population (referred to as 'revised target population' by EMA). What concerns NICE, on the other hand, is most appropriate RR estimate for the population of women covered by the SPC and for whom NICE made its original decision.

Secondly, the argument of statistical power/significance for a RR analysis is fallacious. Power is dependent on the number of events (hip fractures) and choosing a sub-group will tend to reduce power unless the effect size – the relative risk - is markedly greater in that sub-group. Because Servier's subgroup was, in effect, chosen to maximise the placebo group's hip fracture risk, and thereby exaggerated the RR between strontium ranelate and placebo, Servier was faced with the conundrum of appearing to have gained more power from fewer events. Indeed, a subgroup with 130-140 hip fractures has approximately 80% power to detect a 40% relative hip fracture risk reduction (i.e. a RR of 0,6). However, the Servier defined subgroup of the TROPOS trial had only 83 hip fractures, and so it certainly was not "fully powered" (as stated in the Servier response to the 'Statement of Reasons' document by NICE, page 4) even with respect to a 40% relative reduction.

What is required for a cost-effectiveness analysis is the most appropriate estimate of average effect and associated uncertainty which can then be propagated through the economic decision model and displayed in a Cost-Effectiveness Acceptability Curve (NICE Methods

Guide, 2008). Choosing a sub-group without any knowledge of the data but some biological plausibility may in some circumstances be sensible. There is no general evidence that anti-osteoporosis drugs differ in their *relative* effect across sub-groups. There will inevitably be variation in the absolute *magnitude* of effect because the absolute risk of fracture varies across sub-groups: a one fifth reduction on a 1% annual risk brings the absolute risk down to 0.8% (an absolute difference of 2 in 1,000) whereas a one fifth reduction on a 10% annual risk brings the absolute risk down to 8% (an absolute difference of 20 in 1,000). This variation of absolute effect is taken into account in the economic decision model, but it would generally be best to use a single estimate of relative risk unless there is good evidence, supported by test for interaction, or prior biological or empirical plausibility, to suggest that the relative risk differs by subgroup as opposed to the absolute risk differing by subgroup. Any set of data will show some variation by sub-group simply by chance. However, Servier in effect chose a sub-group in such a manner that it was logically likely to yield an unduly large relative effect, which leads to a biased estimate of RR.

QUESTION 2

> *From a statistical viewpoint, what is the most appropriate approach to the use of data from the whole data set of the TROPOS study and the subgroup data set in relation to determining the relative effectiveness of strontium ranelate?*

There are two issues here: (i) for which population is the relative effectiveness required, and (ii) dependent upon the relevant population, what is the most appropriate statistical analysis for estimating the pertinent RR.

If the relative effect is required in a general population (taking into account the fact that even using the whole TROPOS population there are still questions of generalisability from RCTs to a broader NHS population) then an “Intention-to-Treat” (ITT) analysis of all patients randomised in the trial will yield the most appropriate estimate of effectiveness (using purely trial data – these could be adjusted further if the selection of patients into the trial was deemed to be too stringent). In, in fact there are a number of trials available, as there are here, then a pooled analysis of the ITT data using meta-analytic methods to provide an Integrated Analysis of Efficacy (IAE) is the most appropriate overall summary of effectiveness (EPAR, Scientific Discussion, page 17).

If the relative effect *is* required in a sub-population, then the issues of the subgroup’s selection/definition, plausibility and estimation need to be addressed (as well as feasibility from NICEs perspective). In terms of selection, the choice of which risk factors and which cut-points (if they are continuous risk factors) to use in order to define subgroups is clearly an important one, and is prone to the methodological issues we describe above in our response to Question 1. Biological and clinical plausibility obviously are potentially more subjective issues, though the extent to which plausible differential effects across subgroups can be substantiated using other external data sources, or expert judgment, is usually a crucial step to establishing such a principle. Indeed, Servier’s other trial (SOTI) also involved 1,649 postmenopausal women with established osteoporosis (low lumbar BMD and prevalent vertebral fracture) and assessed the efficacy of strontium ranelate. As with other outcomes considered by the EMA in EPAR, a meta-analytic approach (i.e. pooling data across trials) should be adopted to estimating the RR for hip fracture whether that be in the whole population or indeed a specific subgroup. In their response to clarification document, Servier declined to present the results for the subgroup of women in SOTI who were 74 years and over and had a femoral BMD T-score ≤ -3 (≤ -2.4 NHANES III), stating that only 9 hip fractures were observed over 3 years in total (the trial was, however, due to be followed up for 5 years). Servier maintained that so few hip fractures rendered implied that the requested analysis would be unable to provide reliable or robust estimates for the efficacy of strontium ranelate, and that the SOTI trial had been powered with respect to vertebral fractures. However, it is precisely in such situations that a meta-analytic approach (pooling data from a number of potentially under powered trials with respect to the endpoint in question) has an important role to play.

Whether considering the TROPOS trial in isolation, or adopting a meta-analytic approach, a key issue in estimating the magnitude (and associated uncertainty) in one of many potential subgroups is that of subgroup-to-subgroup variability. This needs to be appropriately accounted for, and a Bayesian approach (Spiegelhalter *et al*, 2003 – pages 91-100 & 269-278), in our opinion, should be adopted. In this situation, such an approach would almost inevitably lead to an estimate of the RR for women who were 74 years and over, and had a femoral BMD T-score ≤ -3 (≤-2.4 NHANES III) that was intermediate between 0.64 and 0.85 (that observed in the overall population).

QUESTION 3

> *Given the data reviewed what, in their expert view, is the most plausible relative risk for strontium ranelate to use in making recommendations for the population covered by the marketing authorisation for strontium ranelate?*

For the reasons given above in addressing questions 1 and 2, the RR (and associated uncertainty) from the whole trial population is the most appropriate if the target population is that covered by the SPC and adopted by NICE in its original guidance (TA 160/1).

If the indication were to be restricted to the sub-group used to obtain the RR of 0.64, even this estimate would be likely to be too extreme. This is both because of the method of selection of the sub-group, but also because *any* choice of sub-group should have its effect-size moderated*, or shrunk,* to be closer to the overall result. This is a general principle and is true whether one is looking at benefit or at harm (Spiegelhalter *et al*, 2003 – pages 91-100 & 269-278),.

If NICE were to consider whether potentially different decisions were to be made (as alluded to by Servier in their most recent response) then the approaches outlined in answering question 2 need to be adopted, as well as addressing issues of populating the rest of the economic decision model with evidence appropriate to that particular subgroup.

REFERENCES

A response by Servier to the Statement of Reasons provided by NICE; 25 August 2010.

A response by Servier to the clarification document in response to the questions posed by the DSU; 15 September 2010.

Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000; 355:1064 –1069.

Brookes ST, Whitely E, Egger M, Smith GD, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. J Clin Epidemiol. 2004;57(3):229-236.

Court of Appeal approved document, Royal Courts of Justice, London, case no: C1/2009/0805, 31 March 2010.

EPAR – Scientific Discussion; November 2005. http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Scientific_Discussion/human/000560/WC500045522.pdf (Accessed 17th September 2010)

National Institute for Health and Clinical Excellence Guidance TA160: Alendronate, etidronate, risedronate, raloxifene and strontium ranelate for the primary prevention of osteoporotic fragility fractures in postmenopausal women (amended); October 2008, (amended January 2010).

National Institute for Health and Clinical Excellence Guidance TA161: Alendronate, etidronate, risedronate, raloxifene, strontium ranelate and teriparatide for the secondary prevention of osteoporotic fragility fractures in postmenopausal women; October 2008.

Summary of product characteristics – Protelos; December 2009, (updated February 2010). http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Product_Information/human/000560/WC500045525.pdf (Accessed 17th September 2010)

Spiegelhalter D.J, Abrams K.R, Myles J.P. Bayesian Approaches to Clinical Trials and Health-care Evaluation. Chichester: Wiley and Sons. December 2003.

Stevenson M, Davis S, Lloyd Jones M, Beverley C. The clinical effectiveness and cost effectiveness of strontium ranelate for the prevention of osteoporotic fragility fractures in postmenopausal women, Assessment Report, available at http://www.nice.org.uk/nicemedia/live/11680/36630/36630.pdf

Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA 1991;266:93-8.