AlixPartners

BritNed v ABB: the probative value of statistical evidence in cartel damages cases[1]

Introduction

Anyone with an interest in antitrust litigation in the EU will have watched closely the recent judgment of Mr Justice Marcus Smith in *BritNed v ABB* – the first cartel damages claim to reach final judgment in the English courts.[2]

By way of background, BritNed owns and operates the BritNed interconnector, which is a submarine cable system connecting the Dutch and UK electricity grids. BritNed claimed that it had suffered damages due to its purchases of power cables from ABB, because ABB was found by the European Commission to have been party to a global cartel relating to high voltage submarine and underground power cables.[3] BritNed sought damages in excess of €180 million under three heads:

- For cartel overcharge associated with higher prices for the cable element of the BritNed interconnector;
- For lost profit, as BritNed contended that absent the cartel it would have acquired a high capacity cable system, which would have generated higher profits that were consequently foregone; and
- Compound interest due to the higher capital costs that BritNed consequently faced due to the cartel overcharge.

ABB disputed all these heads and added that any claim for damages needed to be assessed in light of a regulatory cap on BritNed's earnings.

This article focuses on the judge's evaluation of the statistical evidence deployed to assess the cartel overcharge. There is much to digest. After careful consideration the judge considered that the claimant's statistical model was unreliable and disregarded it.[4]

Ultimately, the judge ruled out any *direct effect* of the cartel on pricing based on two key pieces of factual evidence. First, he scrutinised closely the evidence on how bids had

---

1 By Colley L., Vincent D., Darbaz B. of AlixPartners, forthcoming in *Journal of Competition Law and Economics* (OUP)

2 *BritNed v ABB* (EWHC 2616, 2018). Judgment available at: https://www.judiciary.uk/judgments/britned-v-abb-another/.

3 Decision of the European Commission dated 2 April 2014 in Case AT.39610 – Power Cables.

4 See *BritNed v ABB*, paragraphs 417 and 419.

been compiled and negotiated in relation to the BritNed project.[5] Second, he considered what happened to ABB's accounting profit margins during and after the cartel, and it was striking that ABB's profit margins on the BritNed project were lower than the post-cartel average.[6]

The judge did however award damages on two novel heads of loss comprising:

- 'baked-in inefficiencies' that were deemed to have been passed-on to BritNed;[7] and
- a share of 'cartel-related cost savings' for ABB, resulting from not having to compete with the other cartelists.

These damages amounted to €13 million, which is equivalent to 4.9% of the final price in the contract between BritNed and ABB.[8] The rationale for, and the approach to quantifying, these losses will attract much commentary[9] (and both parties have appealed).

To assess the merits of the court's scrutiny of the statistical evidence on cartel overcharge,[10] we consider those factors that affect the probative value of statistical evidence – which can be defined as how well the models avoid false positives and false negatives (i.e. finding an overcharge where none exists or failing to find an overcharge when one exists). Scientific scrutiny of the probative value of such models requires consideration of the concepts of *statistical significance* and *statistical power*.

In rejecting the claimant's statistical evidence, the judge notably discussed how he thought about the concepts of *statistical significance* and the balance of probabilities test

---

5 The judge placed considerable weight on the fact that the key ABB personnel who negotiated the BritNed contract and formulated the bid had no knowledge of the cartel. This was reinforced by the direct evidence that ABB's adjustments to bid offers, when those in the know could have influenced things, were only in one direction (down), and ABB's final price was materially affected by a major concession during final negotiations by someone unaware of the cartel (see *BritNed v ABB*, paragraphs 439-444).

6 This showed that, although ABB's profit margins on the other cartelised projects were materially higher (average 26.7%) than post-cartel projects (average 21.1%), ABB's margin for the BritNed project was actually lower than average post-cartel margins at 18.6%, albeit this was 1% higher than those projects deemed most comparable (see *BritNed v ABB*, paragraphs 385 and 443).

7 Such that prices were higher, notwithstanding that ABB's profit margins on the BritNed project were lower than the post-cartel average.

8 The judge reduced the damages award from €13 million (plus simple interest) to €11 million (plus simple interest) in his supplemental judgment of 1 November, by reason of uncertainty over the future operation of the regulatory return on capital cap.

9 The 'cartel savings' element would not appear to work in a compensation framework (there is no mention of exemplary damages) as the claimant would not have been harmed by these savings. There is clearly a tension between the two heads regarding pass-on: the judgment must assume that inefficiencies were passed-on, but savings were not, else the 'cartel savings' would have to be offset against, not added to, the inefficiency damages. The empirical underpinning of the inefficiency damages will also attract much scrutiny in particular why there is no offset for the benefits of ABB's higher copper content reducing BritNed's energy waste and operating expenses as recognised in paragraph 448 (1).

10 Both experts deployed statistical evidence, but the defendant used it as a cross-check against the accounting evidence of margin that found no *a priori* evidence of overcharge. Coupled with the requirement for the claimant to prove its loss, this led to the scrutiny focusing on the claimant's statistical evidence.

associated with the court's standard of proof. We find that the judge's treatment of *statistical significance* was entirely correct and demonstrate this using the related concept of *positive predictive value*. We also find that, although he did not refer to *statistical power* directly, the judge's assessment of overall reliability amounted to a reasonable examination of that concept too, at least in relation to the claimant's statistical model.

We conclude that the judgment shows the English court's keen ability to deal with complex statistical evidence, and the reasons for rejecting the claimant's model outright on this occasion are entirely case-specific. Accordingly, we expect statistical evidence to continue to feature heavily in cartel damages cases. We also consider that the framework laid out here, which gives more clarity on how various factors influence *statistical power* and *positive predictive value*, is useful for assessing the ultimate probative value of statistical evidence. In support of this, we refer to the latest academic literature which raises serious concerns about the quality of social science research due to a lack of focus on these concepts.

Finally, we note that understanding these concepts will be particularly important in making up-front assessments of the likely probative value of methodologies at the certification stage in class action procedures (i.e. prior to analysis being carried out), which are becoming more common in the English court.

Summary of the judge's findings on the statistical evidence

The claimant's statistical evidence estimated an average overcharge of 25.4% across all the defendant's cartel period projects and a project-specific overcharge of 21.8% on the BritNed submarine cable project.[11] The claimant expert deemed these estimates to be 'statistically significant' using the statistical convention of a 95% 'confidence level'.[12]

However, the judge concluded that the claimant's statistical analysis was "*insufficiently reliable to be used in any way at all*".[13] He anchored this finding in a detailed assessment of whether the model was fit for purpose given the direct evidence and assessing the *"fragility"* of the model given the results of sensitivity tests for reasonable adjustments.

To demonstrate the model's unreliability, the judge pointed to:

---

[11] A key issue was that the claimant's model was only capable of identifying an overcharge on BritNed with a model that estimated overcharge on all fifteen projects that ABB won in the cartel period. If the claimant's model was re-run excluding the other fourteen projects, the BritNed overcharge became negative and not statistically significant. The judge preferred a model that sought to measure the actual harm suffered by BritNed, rather than an overcharge across different projects to which BritNed was not party, reflecting the fact that submarine projects are "unique and bespoke" and circumstances may differ greatly, including whether other projects had BritNed's "option of simply not proceeding with the project" (see *BritNed v ABB*, paragraphs 420-1). However, the judge also recognised that the relevant counterfactual may not necessarily include ABB winning the BritNed project, and that absent the cartel BritNed may have accessed a lower cost third-party provider (see *BritNed v ABB*, paragraph 18). As data on third-party prices and costs was not available, a finding of no overcharge in ABB's pricing does not necessarily prove that BritNed was denied lower prices.

[12] Although see later discussion on 'one-tailed' versus 'two-tailed' tests.

[13] Paragraph 417.

- the removal of underground projects leading to the overcharge becoming statistically insignificant i.e. failing to find an overcharge at the conventional 95% confidence level;
- the removal of subsidiary variables (the order backlog and time trend) having a disproportionate effect on results;
- the wide variance in estimated effects on the other cartelised projects (including negative overcharges for five projects and "*massive*" overcharges of more than 40% for six others); and
- the *"scarcely impressive"* range of outputs associated with using the 95% 'confidence interval'.[14] Although the claimant's central estimate was 21.8%, the range at the 95% level of confidence was 0.3% to 38.7%.

The judge went on to consider arguments for findings at levels of confidence lower than 95% given the court's use of a balance of probabilities test for standard of proof (i.e. 51%). However, he concluded that *"the analogy to the balance of probabilities test used by lawyers is entirely spurious."*[15]

Assessing the probative value of statistical models

Statistical models seek to isolate the cartel's effect on prices by filtering out the influence of other drivers of price (such as cost, capacity, or demand) that may have also changed between the cartel period and the cartel-free period. The probative value of this evidence ultimately depends on its ability to avoid the risk of false positives (i.e. a type 1 error: finding an overcharge when none really exists) and false negatives (i.e. a type 2 error: failing to find an overcharge when there really is one).

Statistical analysis attempts to deal with such risks of error by applying the concepts of *statistical significance and statistical power:*

- **Statistical significance** involves setting a high bar before a positive result is deemed valid. The convention is to set a 5% risk of a type 1 error (denoted by α) using a 95% 'confidence level' (explained below).
- **Statistical power** involves aiming for a model that can find an overcharge if there really is one, thereby minimising the risk of a type 2 error (denoted by β). The convention is for a risk of a type 2 error of 20% which implies a *statistical power* of 80%.[16] As we shall see, such levels are very aspirational, but what is important to understand is what drives *statistical power*.

---

[14] Paragraph 418 (3).

[15] Ibid.

[16] 'The Power of Bias in Economics Research', Ioannidis JPA, Stanley TD, Doucouliagos H., (2017), *Economic Journal*, 127, p. 239*: "Since Cohen (1965), adequate power in most social sciences has been conventionally defined as 80%. That is, the probability of type 2 error should be no larger than four times the probability of the conventional type 1 error (0.05). A 20% type 2 error can still be considered to be rather high (Schmidt and Hunter 2015). Others have argued that the optimal pair of type 1 and 2 errors vary according to the circumstances and aim of the study".* The authors note that applications for research grants are routinely required to estimate the *statistical power* of the study (and the sensitivity to sample size in particular) with target levels set at 80%.

Table 1 summarises the four possible outcomes from statistical testing comparing two possibilities for underlying truth – whether there really is an overcharge or not – with whether the statistical test finds one or not.

Table 1: outcome versus truth – balancing error risks

		Test finding	
		No overcharge	*Overcharge*
Truth	*No overcharge*	**True negative**	**False positive** type 1 risk: α = 5%
	Overcharge	**False negative** type 2 risk: β = 20%	**True positive** statistical power: (1 – β) = 80%

The following discusses in more detail the concepts of *statistical significance* and *statistical power*, and how they were addressed in the judgment.

Statistical significance and confidence levels

The first thing to note from Table 1 is that setting the threshold for *statistical significance* at a 5% risk of type 1 error does not imply there is a 95% chance that the model would correctly identify an overcharge if one exists (or that we could be 95% confident our positive is really true). In a nutshell, that is why it cannot be equated with the balance of probabilities. The prospect of finding an overcharge if one exists is defined as the *statistical power* (here, ideally, 80%) which is the inverse not of the 5% risk of a false positive, but of the 20% risk of a false negative.[17]

So, what does the 95% relate to? To understand this, we have to review the way the amount of type 1 risk gets set using what are called 'confidence levels'.

Statistical analysis involves assessing how sure we can be that an effect is sufficiently far from zero that we can rule out it being zero. To assess this the model produces a point estimate of the effect and a range around that estimate that reflects the statistical certainty in that estimate. If that range includes the number zero, then the estimate is deemed not statistically significant because it is not sufficiently far from zero for us to be sure (given our tolerance of risk of type 1 error) that it is not a chance result emanating from imprecision in the statistical analysis. Setting a higher level of confidence results in a wider range for a given estimate and so increases the chance of that range including the number zero. Accordingly, a higher confidence level translates into a higher bar for a finding of *statistical significance*.

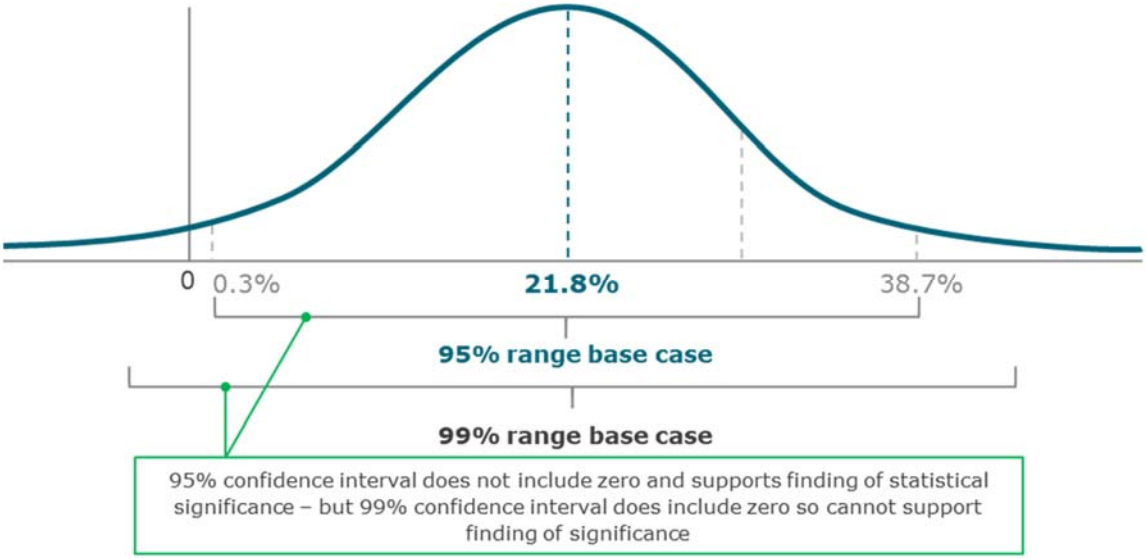
To illustrate, in the BritNed case the claimant produced a central estimate of the BritNed project-specific overcharge of 21.8% with a range that, at the 95% confidence level,[18] only just avoided including zero (Diagram 1). This allowed the claimant expert to deem

[17] Although, as explained later, increasing the risk of type 1 does increase the *statistical power* (because it increases the chance of any positive result whether true or not). The relationship however is not one for one because *statistical power* is determined by a range of factors, including the size of the effect and the sample size.

[18] See *BritNed v ABB*, paragraph 418 (3).

the result statistically significant. Imposing a higher confidence level (such as 99% confidence) would have turned the result insignificant because at that higher bar, we could not be confident that the range does not include zero.

Diagram 1: statistical significance and confidence level



The statistical convention is to choose the confidence level that minimises type 1 error to under 5% — or 1 in 20. However, the 95% level is not set in stone and may vary.[19] The judgment notes this with reference to the European Commission antitrust damages guidelines,[20] as follows:

> *"it is a convention in economic science for both the notion of 'confidence interval' and 'statistical significance' to use a 95% threshold of probability. It should be stressed that this represents a pure convention and that more, as well as less, stringent thresholds (for instance: 99%, or 90% probability) may likewise provide useful information."* [21]

As Diagram 1 shows the claimant's base case was statistically significant using the 95% convention. However, some of the sensitivity tests flipped the finding from significance to insignificance at this level. We can see this in Figure 2 which shows the result of the sensitivity when only the submarine cable projects were included in the model. The higher estimate of 27.7% had a much larger range around it that now included the

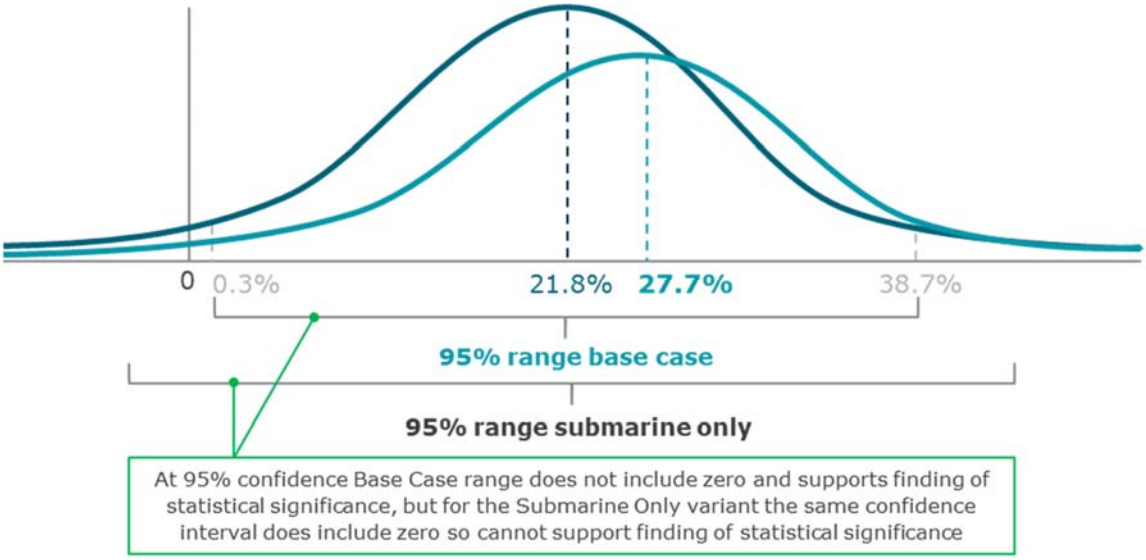
[19] The convention emanates from the work of Sir Ronald Fisher, the godfather of modern statistics, who applied this working rule after years of conducting experiments in the early twentieth century – many of which involved testing for any impact of different fertilisers on crop growth in randomised control trials

[20] European Commission: Practical Guide to Quantifying Harm in Actions for Damages Based on Breaches of Article 101 or 102 of the Treaty of the Functioning of the European Union. http://ec.europa.eu/competition/antitrust/actionsdamages/quantification_guide_en.pdf

[21] See *BritNed v ABB*, paragraph 309.

number zero, resulting in a finding of statistical insignificance. That may partly explain why the claimant expert was effectively calling for a lower 90% threshold to be used.[22]

Diagram 2: turning from statistical significance to insignificance



Confidence levels and the balance of probabilities – the concept of positive predictive value (PPV)

The judge explicitly considered the relationship between confidence levels and the balance of probabilities used as the court's standard of proof. Having found that the 95% confidence level produced a 'scarcely impressive' range of 0.3% to 38.7%, he considered how that range would vary under lower levels of confidence and found that even at 51% level the range of estimates is still very wide (15% to 28%).

However, although this exposition of the results for lower levels of confidence suggests that the judge did not rule out relaxing the confidence level, he ultimately concluded there can be no suggestion of confusing statistical confidence levels with the balance of probabilities test: *"it would be unconventional to use a 51% confidence interval for the analogy to the balance of probabilities test used by lawyers is entirely spurious."*[23]

This is entirely correct and can be shown easily by using the concept of *positive predictive value.* This concept shows how the probative value of statistical analysis depends not only on the confidence level used but also on the *statistical power*, and, critically, on how the thresholds for type 1 and type 2 error risk actually translate into the probability of a true positive, given how prevalent the searched for effect actually is.

---

[22] See later discussion on one-tailed versus two-tailed tests.

[23] In doing this he also rejects the claimant expert's notion that the bottom end of the range (whatever the level of confidence chosen) could represent the minimum level of overcharge. He emphasises that the range is a *"measure of the degree of uncertainty relating to an estimate of overcharge, but it does not provide a measure of certainty"*. See *BritNed v ABB,* paragraph 418 (3).

To illustrate, consider a drug test that has a *statistical power* (i.e. chance of correctly detecting a cheat) of 80%,[24] and in which we are 95% confident will not make a type 1 error (i.e. it will only wrongly flag 5% of clean athletes as cheats).[25]

If someone fails the test, what is the probability of them being guilty?

The answer depends on the proportion of competitors who actually do cheat, i.e. the prevalence of cheating.

Let's suppose that 10% of 1000 athletes in a race are cheating. Then the *positive predictive value* of this drug test will be as follows.[26]

- The total accused will be a combination of true positives (those who cheated and failed the test) and false positives (those who did not cheat, but still failed the test).
- The number of true positives will be the total number of actual cheats (100) times the probability of identifying them as cheats (80%), which equals 80.
- The number of false positives will be the total number of clean athletes (900) times the probability of an incorrect positive test (5%), which equals 45.
- The PPV is the proportion of all who fail the test (125) who are in fact cheats: PPV = 80/125 = 64%.

For a test with a lower confidence of avoiding type 1 errors of 90% but the same *statistical power*, the PPV would be 80/170 = 47% (i.e. it would add no additional information to a balance of probabilities test). For a test with only 51% confidence the PPV would be 15% (i.e. useless). What's going on here is that at low prevalence of cheating, relaxing the confidence level means that false positives dominate. If prevalence increases, then the PPV – and therefore the probative value of the test – increases.

Hence, the PPV is much closer to the concept of probability used in the civil litigation standard of proof which asks how likely it is, based on the evidence, that the accused is really guilty or that the cartel has really had an effect. The judge was therefore entirely correct in dismissing the notion that the 51% balance of probabilities is comparable with the 95% confidence level, and that a lower confidence level should be contemplated in view of this. That is because the probative value of statistical evidence (as given by the PPV) depends on not just the confidence level but also the *statistical power* and the prevalence of the effect we are searching for.

---

[24] In the medical literature this is known as the *sensitivity* of the test.

[25] In the medical literature this is known as the *specificity* of the test.

[26] This can be formally presented as PPV = ([1 – β] x R)/ ([1 – β] x R + α) where α = risk of type 1 error, β = risk of type 2 error and R = prevalence (defined as affected divided by unaffected). See *Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience*, by Button KS et al (2013), *Nature Reviews Neuroscience 14*, p366. This means that PPV is increasing in *statistical power* (i.e. inverse of type 2 error), in prevalence and confidence level (i.e. inverse of type 1 error).

So, what is the prevalence of cartels having an effect? As all antitrust litigation practitioners know, the EC antitrust damages guidelines draw on a major study (the Oxera Study) that includes a meta-study of cartel effects.[27] Famously, the Oxera Study estimates average cartel effects at around 20% and finds that cartels lead to a positive overcharge 93% of the time.[28] Below we set out how seminal academic research on the implications of properly accounting for statistical power in such meta-studies suggests that these figures may be grossly inflated. However, taking such a high prevalence rate at face value would imply that the relationship between confidence levels and balance of probabilities is likely to be much closer. That said, the judge is still right to reject the analogy as spurious in principle, because that relationship is not only a function of prevalence but also of statistical power.

We now consider what drives *statistical power*.

Statistical power

Assessing *statistical power* is critical because the results of the estimate of cartel effect alone have no probative value:

- A defendant model that shows no statistically significant overcharge has not proven there was no overcharge. To assume so would be an example of the Argument from Ignorance Fallacy. In fact, in statistics you never actually prove that the null hypothesis (here, absence of overcharge) is right. Statistical analysis proceeds by the scientific method of falsification i.e. seeking to falsify the null hypothesis. The defendant expert must show that their model has the statistical power to find an overcharge if there really was one.[29]
- Similarly, a claimant model that shows an overcharge that is statistically significant does not prove the overcharge. To argue that it does would suffer from the Prosecutor's Fallacy – sometimes colourfully referred to as the "Bigfoot fallacy": just because you have found a very large footprint in the snow does not mean you have proven the existence of Bigfoot. The claimant expert must demonstrate that their model has the statistical power to find (an unbiased) estimate of the overcharge if there really is one (i.e. that the model is robust enough to correctly filter out relevant confounding factors and that the finding of overcharge is not a result of bias).

This gets us much closer to gauging directly the probative value of the model and much of the debate in BritNed was relevant to assessing the statistical power of the models.

---

[27] Oxera (2009) Quantifying Antitrust Damages: Towards Non-Binding Guidance for the Courts, available at http://ec.europa.eu/competition/antitrust/actionsdamages/quantification_study.pdf. See paragraphs 139 et seq. of the EC antitrust damages guidelines. This includes an adjusted meta-study (see pp 90-92) based on original work by Connor and Lande (2008). A meta-study is a review of all underlying research in a specific research area. The Connor and Lande study contained outputs of more than 200 social science studies of cartel effects. The Oxera Study revised this to a sub-set of 114 for studies that were peer reviewed, discussed the methodology used, were published after 1960, and presented an overcharge for the entire cartel period. Studies prepared in the context of litigation were explicitly removed.

[28] Ibid., Figure 4.1.

[29] As discussed, the defendant's statistical model received far less scrutiny in the BritNed judgment. When the defendant asserts that their model finds no overcharge, one technique would be to run a type of "placebo analysis" where an artificial overcharge is super-imposed on the raw data and we then assess whether the defendant's model can pick it up. This can be tested at different levels of (artificial) overcharge.

However, it's important to understand the drivers of statistical power before making an overall assessment. There are four main elements:

1. Confidence level – reducing the risk of type 1 error reduces the ability of the model to find an overcharge at all (if it does exist) and therefore also reduces *statistical power*;

2. Sample size – larger samples increase *statistical power* because they make the model more certain of its estimates, thereby leading to narrower confidence intervals that are more likely to lead to *statistical significance*;

3. Effect size – the likely size of the effect that is being tested for if it exists. Are we searching the haystack for a (potential) needle or a sword? The larger the likely effect the lower the power needed to detect it; and

4. Bias – the ability to identify the true estimate of any overcharge will be frustrated by any bias that is introduced – whether by error, design or mere difference of opinion regarding the implications of factual evidence.

Assessment of bias is essentially a qualitative exercise. However, the three other variables can be combined in a formula to compute the level of *statistical power* of a model to identify a true effect if one exists and to produce an *unbiased* estimate of that effect.[30]

We now consider each in turn.

1. Confidence levels

Higher confidence levels that minimise the prospect of type 1 error also reduce the *statistical power* of a test because they restrict the ability of the test to find an overcharge at all.

Assessing the relevance of the balance of probabilities threshold in determining the right confidence level, was only one aspect of the question the judge had to consider. The claimant expert was also implicitly urging the judge to depart from the convention of 95% by promoting the use of a 'one-tailed test' instead of the standard 'two-tailed test'.[31] The basis for this, the claimant expert argued, was that it would be safe to assume the cartel could only increase prices and never reduce them. This is important because the threshold between a finding of *statistical significance* at the 95% confidence level in a one-tailed test is equivalent to relaxing the level to 90% in a two-tailed test.[32]

The difference between 90% and 95% would not alter the significance of the claimant's base case results. However, as illustrated in Diagram 2, interpretation of the sensitivity tests will also be affected by the threshold set, and therefore much could turn on this seemingly modest adjustment.

---

[30] Computation can be done in a standard statistical software like Stata.

[31] The claimant expert does rightly point out that the choice of number of tails and the choice of risk tolerance are separate (*BritNed v ABB*, paragraph 311).

[32] A two-tailed test splits the 5% risk of type 1 error between a very positive tail and a very negative one (i.e. 2.5% in each). The one-tailed test puts all of the 5% into one tail, thereby making it an easier threshold to cross. So, a 5% risk of type 1 error in a one-tail test is equivalent to a 10% risk in a two-tailed test.

In the written-up cross-examination we can see that the defence vigorously resists the notion of relaxing the confidence level. The judge merely notes that the prior assessment of whether one could rule out a cartel reducing prices might be valid but that if one remained neutral, one would stick with a two-tailed test. It seems that although the judge recognises that economic theory strongly indicates that cartels will raise market prices, he appears to attach weight to the fact that we are here considering a specific project and is therefore reluctant to rule out any prospect of a lower price.[33]

The defendants may well have also referred to US precedents where the courts have emphasised application of the 95% convention.[34] However, as the above discussion shows, context is important. Those US authorities refer to the use of statistical evidence to help identify liability for example in breaches of wage discrimination legislation. Some have argued that in a follow-on context it may be appropriate to use lower levels of confidence.[35]

In the Copper Tubes litigation in the UK High Court[36] the claimants also argued that revisiting the statistical evidence submitted by one of the defendants to the European Commission,[37] and applying a lower threshold for confidence of 80% to that evidence turns a finding of no overcharge into finding a statistically significant 4% overcharge. However, this was based on the crude argument that 80% was still a long way north of the 51% balance of probabilities, as opposed to a proper evaluation of the implications for *statistical power* and ultimately the *positive predictive value* of the model.

What we may see therefore are arguments based on the likely prevalence of cartels having an effect. A high-profile example of the use of discretion in setting confidence levels was the hunt for the Higgs boson particle, where the chief scientist at CERN opined that "*extra-ordinary claims require extra-ordinary evidence*" and set the risk of type 1 error at 1 in 3.5 million (i.e. requiring a confidence level of 99.99997%). After all, before you announce to the world you have found the Higgs boson particle, you want to be very sure.

Claimants may well therefore point to the high prevalence of cartel effects in the Oxera Study (which "found" that 93% of cartels have an effect) to argue that lower confidence levels could be appropriate. This would be on the basis that with such high prevalence, there may be ample *positive predictive value* to play with, such that we could live with

---

[33] When some of the claimant's results start to show negative overcharges the judge remarks that *"a negative figure is, of course, entirely at variance with [the claimant expert's] assumption that a cartel would produce a nil or positive effect, and not a negative effect"* (*BritNed v ABB*, footnote 435).

[34] *In Castenda v Partida*, the Supreme Court emphasised that least 95% confidence levels are needed. In *Lopez v Laborers*, the Supreme Court stated that 99% confidence levels are needed to safely reject chance.

[35] See for example Nera Economic Consulting's response to EC consultation on antitrust damages: http://ec.europa.eu/competition/antitrust/actionsdamages/hofer.pdf

[36] The case settled before trial. AlixPartners represented the lead defendant.

[37] Referred to in the EC decision on industrial tubes (available at http://ec.europa.eu/competition/antitrust/cases/dec_docs/38240/38240_29_1.pdf) as part of an unusually long discussion of the likely effect of the cartel that included a detailed evaluation of statistical evidence (*COMP/E-1/38.240*, paragraphs 295 to 314).

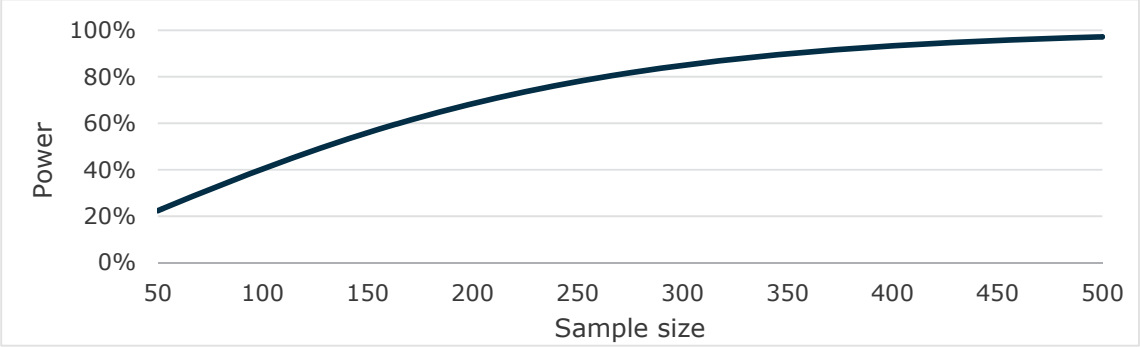
more type 1 error and less type 2 error.[38] This would be a far more sensible debate than the vague references to the balance of probabilities some experts have made. However, as we discuss later, recent academic research into the accuracy of meta-studies suggests that great caution should be attached to any prevalence findings in the Oxera Study, and much more detailed work on this is required.

2. Sample size

Sample size is a key driver of *statistical power*. If sample sizes are small, then there is less variation in the data to control properly for other factors that affect prices in order to isolate the overcharge.[39] As a result, the range of estimates around the overcharge at given levels of confidence will be wider, thereby reducing the prospect of a finding of *statistical significance*.

Diagram 3 plots the *statistical power* for different sample sizes when the probability of a type 1 error is set at 5% and assuming the true effect is 25.4% (the claimant's base case estimate of average overcharge). This shows that using the claimant's model of 92 observations the claimant's model would have a *statistical power* of 38%.[40] The defendant's model with 67 observations has an estimated *statistical power* of 29%.[41] Relaxing confidence level to 90% changes these powers to 50% and 41% respectively.[42]

Diagram 3: statistical power at different sample sizes when the true effect is 25.4%



[38] A related consideration is that the convention of four times the risk of type 1 than type 2 in social sciences reflects in part the dynamic nature of research – that a type 1 error as well as being inherently damaging, may stymie the pursuit of the true relationships. It may be argued that this concern does not apply in a follow-on damages case which is a one-shot game. Similarly, how the criminal justice system balances these risks is likely motivated by more philosophical concerns regarding the inherent awfulness of a false positive (e.g. Dreyfus) compared to false negatives (OJ Simpson?) who may after all get their comeuppance in the next life (or the civil courts!). But, also here too is the practical point that a false positive prevents the further pursuit of the real culprit.

[39] The appropriate minimum sample size is also a function of the number of variables that have to be controlled for to retain sufficient "degrees of freedom" to enable the estimation process to work.

[40] Assuming that the true overcharge is equal to the estimate of 25.4% with a standard error of 0.18 (*BritNed v ABB*, paragraph 320).

[41] See *BritNed v ABB*, paragraph 331.

[42] As discussed in the next section, if the true effect is in fact lower than 25.4%, *statistical power* would be lower.

In *BritNed v ABB*, the impact of a small sample size features starkly, in particular when the judge questions the relevance of including underground cables which accounted for 43 of the total of 92 projects in the sample. There are several differences between underground and submarine cable projects, but the claimant expert argued that including them and introducing other control variables to allow those differences to be filtered out, was worth it to raise the "*accuracy and reliability*" of the model.[43] Essentially, the claimant expert was prepared to introduce a potential source of bias with a view to increasing the *statistical power* of the model by using a larger sample size.

The judge considered that the direct evidence showed that underground cables were very different products (under half the price of submarine cables), and subject to very different market dynamics. As a result, the judge found that such fundamental differences could not be controlled for properly, and the *"consequence was the introduction of significant unreliability into the model's output."*[44]

The impact of removing the underground projects reduced the sample from 92 to 49, and as Diagram 3 shows, this would considerably reduce the power of an already low-powered model. Although the estimate of overcharge increased to 27.7%, crucially the range for the 95% confidence interval increased dramatically, resulting in the estimate becoming statistically insignificant (see Diagram 2).[45]

What perhaps escaped attention was that at these relatively modest sample sizes, the defendant's model would also have had low *statistical power*. The defendant expert was careful to deploy their own statistical evidence as more of a cross-check against the accounting margin evidence. However, for it to serve any purpose as a cross-check, it would still need to have had some probative value and that does not appear to have been addressed robustly.

Indeed, the judge appeared attracted to the comparative simplicity of the defendant's model: "*given the far more limited number of parameters involved in [the defendant's] model, when compared to [the claimant's] model, this was a much more straightforward exercise*".[46] The judge focused on the model's low and insignificant overcharge estimates without really questioning whether such a simple model, given the even smaller sample size, would have had the *statistical power* to find an overcharge if there really was one.

3. Likely effect size

The third key driver of *statistical power* is the likely size of the effect that is being searched for, if indeed there is an effect. A model with relatively low power may be sufficient if the effect is likely to stick out like a sore thumb.

This means that whilst prior expectations should play no role in determining the outcome in a specific damages case, they can feature in an assessment of whether the modelling approach is likely to be fit for purpose. This is particularly relevant in the context of class

[43] See *BritNed v ABB*, paragraph 397.

[44] *BritNed v ABB*, paragraph 397.

[45] *BritNed v ABB*, paragraphs 395 and 396(1)(b).

[46] *BritNed v ABB*, paragraph 334.

actions, where such assessments are made at the certification stage, in advance of the analysis being carried out.[47]

As discussed, the Oxera Study estimates average cartel effects at around 20%. This figure has attracted a lot of attention and there was even flirtation with the idea of attaching such a figure to the rebuttable presumption of harm in cartels now included in the EC's Damages Directive.[48]

Unsurprisingly the judge does not address this and if any of the experts' reports or testimony refer to this context, it is not reflected in the judgment.[49] However, in the EU at least, it would be naïve to suggest that the Oxera Study does not provide at least some mood music for these cases. Interestingly, the judge does come close on occasion to allowing expectations to feature in the assessment of the models. For example:

- he observes that: "*there is the fact that removal from the model of variables that might be said to be subsidiary has a disproportionate effect on the model's outcome. Neither the time trend variable nor the order backlog variable ought to be fundamental to the operation of the model.* ***If there is an overcharge, then it ought to be capable of being demonstrated in a statistically significant manner without these variables.***" (emphasis added);[50] and
- he also describes 5% overcharges as "*small*" and 40% as "*massive*".[51]

It is also interesting that the judge did not appear to overtly challenge the plausibility of the claimant model results in light of the margin evidence that showed the claimant would have a mountain to climb to show material overcharges. This margin evidence was therefore not deployed by the judge as a basic 'sniff test' to the statistical evidence. Rather, the judge evaluated the statistical evidence on its merits and then considered which body of evidence he could ultimately rely on.

The role of likely effect in determining *statistical power* means that it is perfectly respectable to have some notion of likely effect size – if indeed there is an effect – and to include that notion in the assessment of likely model robustness. When likely effect size is large, it means that relatively low powered models can be good enough.

A key question, then, is whether any weight can be attached to the Oxera Study and recent academic literature on meta-studies and *statistical power* may be very illuminating here. In their seminal work, Ioannidis et al. (2017) find that the *statistical power* of most of the underlying research used in meta-studies across a range of social science fields is woefully short of the desired 80% convention, and in economic research it tends to be

---

47 Certification must assess not only the likely *statistical power* of models to find overcharges but also to reasonably estimate the way impact may vary across different types of customers.

48 Department for Business, Innovation & Skills (2013), 'Private Actions in Competition Law: A Consultation on Options for Reform – Government Response', 29 January 2013.

49 Although there was a brief discussion about the relevance of any presumption of overcharge. See *BritNed v ABB*, paragraphs 21-23.

50 *BritNed v ABB*, paragraph 418.

51 *BritNed v ABB*, paragraph 418(2)(d).

around the 20% level.[52] They find that the key problem is that meta-studies present average effects giving equal weight to underlying individual studies irrespective of their power. They also find that underpowered individual studies tend to overestimate the true effect, and put this down to a mix of chance, error and publication bias.[53] Crucially, their research shows that weighting the reported average by the *statistical power* of the studies leads to reported effects that are 50% smaller and in a third of cases 75% smaller. That is, meta-studies across the range of economic research areas tend to produce effect estimates that are twice as large and in many cases four times as large.[54]

All this adds up to a need to be very wary of the average effects reported by meta-studies. It is certainly the case that the Oxera Study gives equal weight to studies irrespective of *statistical power*, and so could suffer from similar levels of inaccuracy.[55] To put that in context, if the Oxera Study does suffer from a similar low power bias, and true average overcharges were therefore more in the region of 5% to 10%, it would be less likely for a judge to describe a 5% overcharge as "small". More importantly, it can be easily shown that with much lower expected effects a given sample size has far lower *statistical power*. For example, returning to Figure 3, if the true effect is in fact 6% – the average effect the defendant model derives across the other cartel projects excluding BritNed (where it found no overcharge) – the *statistical power* of the claimant's model falls to 6.4% and the defendant's model falls to 6%.

The upshot of all of this is that the evaluation of the whether there is enough sample size to generate appropriate levels of *statistical power* can be highly sensitive to the size of the expected effect being searched for, and both factors may therefore need to be closely considered. If the above estimates of *statistical power* are correct, the probative value of the defendant model's finding of zero overcharge would be extremely low as the model would anyway struggle to find an overcharge if there really was one. Similarly, the finding of positive overcharge in the claimant's model would be much more likely the result of bias than a true positive, given the model's poor inherent ability to find a true overcharge.

---

[52] In Ioannidis et al. (2017) p. 253, the authors summarise their findings as follows: "*Our survey of 159 meta-analyses of economics reveals that empirical economics research is often greatly underpowered. Regardless of how 'true' effect is estimated, typical statistical power is no more than 18% and nearly half of the areas surveyed have 90% or more of their reported results stemming from underpowered studies. This survey also identifies widespread bias. The majority of the average effects in the empirical economics literature are exaggerated by a factor of at least 2 and at least one third are exaggerated by a factor of 4 or more.*" The areas covered include research into international economics, labour economics, growth and development, microeconomics, macroeconomics, finance and public economics. Note that microeconomics does not include damages analyses and the authors find that "*microeconomics research has more power on average than other research areas*" (footnote 13).

[53] In p. 240, Ioannidis et al. (2017) report: "*When power is low, reported statistically significant findings are quite likely to be artefacts from chance and bias.*" Hence, a low powered study with a high effect should attract particular scrutiny of the *statistical power* of the study and of whether any bias is present.

[54] See footnote 52.

[55] The adjustments made to the Connor and Lande database by Oxera (see footnote 26 above) do not include any adjustment for *statistical power*.

4. Bias

Bias can arise in any of the steps that underpin the results of the models. These steps include data gathering, variable selection, model specification selection, and the judgments regarding weights to attach to the outputs of different models. Even models with high statistical power can produce misleading results if any of these biases are material.

In fact, most of the BritNed judgment focuses on the extent to which the experts' methodologies appropriately controlled for confounding factors given the direct evidence before the court. Primarily this involved examining the factual evidence behind the inclusion or exclusion of explanatory factors in the model. Much of the discussion of bias went under the heading of assessing the overall 'reliability' of the models.

As discussed, a key example was whether the evidence supported the inclusion of underground projects. The assessment of the order backlog variable (a measure of capacity constraints facing a supplier that may influence willingness to price keenly) was also interesting. Here the judge considered whether the order backlog had any material influence on prices by examining this relationship outside the cartel period. The counterargument that post-cartel relationships were not a good guide to during-cartel ones (due to the financial crisis) did not carry weight. Ultimately, the judge clearly felt that the evidence did not support the claimant's position, and that including the variable carried too great a risk of introducing bias or as he put it, "*dangerous levels of uncertainty*".[56]

The overall reliability of the model was also tested through appropriate sensitivity tests to examine the extent to which results changed when alternative, but similar models were used. A finding of low reliability would support the notion that the claimed for effects were more chance results and flowed from inappropriate model selection. Rightly, the judge recognised that the fact that these variant models led to different outcomes was not itself problematic – if they were important variables to include then one would expect results to change.[57] However, sensitivity analysis was nonetheless an important part of understanding overall reliability.[58]

Bias can result from both unnecessary inclusion and exclusion of key explanatory variables.[59] Interestingly, the judge seemed comfortable with the technique of refining the model using a 'general to specific' approach where you drop variables that are revealed to be statistically insignificant to improve the accuracy of the estimates for those variables that are statistically significant.[60] What is going on with this process is a trade-off of precision against the risk of introducing bias by omitting a key variable. It is

---

[56] *BritNed v ABB*, paragraph 413.

[57] See *BritNed v ABB*, paragraph 379.

[58] In examining this the judge considered not only the impact on estimates of overcharge but also on whether the coefficients of the other explanatory variables continued to make sense — for example their magnitude and whether they continued to indicate positive or negative associations.

[59] Strictly speaking, including redundant variables does not bias the estimates but it does undermine *statistical power* by inflating the standard errors which reduces the prospect of finding a statistically significant overcharge if there is one.

[60] See *BritNed v ABB*, paragraph 337 et seq. for discussion of the 'slimmed down' model 2.

a standard procedure but has to be handled with care. A related concept here is the R-squared statistic, which is a measure of the extent to which the proportion of variation in the dependent variable (here, prices) is explained by variation in the included explanatory variables.[61] In US litigation, which has a much longer record of statistical evidence in the courts, this statistic is often scrutinised as a short-hand way of assessing model robustness. As our colleagues have set out in another article such reliance is misplaced and not in keeping with the precise role it plays in scientific research, and it is good to see that it did not feature in the judge's assessment.[62]

A key test that clearly did carry a lot of weight was the de-averaging of the overcharge into estimates for the fourteen other cartelised projects. This found a wide variance in overcharges including negative overcharges for five projects and "massive" overcharges of more than 40% for six others. The fragility of the model in this respect was further disguised by the fact that the individual estimate for the BritNed project overcharge was similar to the average estimate of the model (across all cartelised projects). This was not enough to save the claimant's analysis. The sense was that the model produced such a range of estimates that it could not possibly be reliable, and in essence it was very likely to produce a (highly) biased estimate of the BritNed overcharge.

Finally, we note that the impact of sample size on power also depends on the number of explanatory variables required, and if there are many, a larger sample size is needed. Hence, when the claimant ran the de-averaged model (i.e. with fourteen additional variables to indicate each cartelised project) this was bound to further hamper the *statistical power* of an already fragile model that had only 92 observations. This very likely meant that at least some of the overcharge estimates for the other cartelised projects were not statistically significant. Although the BritNed overcharge was found to be (just) statistically significant, the confidence levels associated with the fourteen other project overcharges were not presented in the judgment. If these confidence levels were not close to the 95% level, whether the estimate is -1% or +40% is irrelevant because the model has found that number to be statistically insignificant and no real weight should be attached to it at all. The likely correct finding is that the model simply does not have the *statistical power* to estimate all the individual project-specific overcharges. Accordingly, the judge may have put too much weight on these values in his finding of general unreliability.

Conclusions

What is clear from the judgment is that the statistical analysis was scrutinised thoroughly. This paper has considered that scrutiny through the lens of the way statistical experts would scientifically examine the probative value of a statistical analysis. We conclude that the judge's approach is largely in sync. In particular, the judge's correct treatment of *statistical significance* in relation to the balance of probabilities standard can be demonstrated using the concept of *positive predictive value*. Also, the evaluation of overall reliability of the models amounted to a reasonable assessment of whether the statistical estimates could be the result of bias.

---

[61] It is not sufficient to identify candidate variables to be dropped on the basis that their coefficient is not statistically significant. One must also test for joint significance which involves assessing whether the model's overall fit is improved or not. This relies on an F-test, which essentially involves comparing the R-squared of the model with and without the candidate variable to be dropped.

[62] *No Minimum Level for R-Squared in Regression Analysis, Law360* (April 22, 2016), by William Choi and Pablo Florian of AlixPartners.

However, the evaluation could have benefitted from a more explicit consideration of the concept of statistical power and its key drivers. Ultimately, it is difficult to attach weight to a finding of positive overcharge if one does not understand the balance of risk of that result being a false positive or a true positive (which is a function of statistical power). Similarly, it is difficult to attach weight to a finding of no overcharge if one does not understand the balance of risk of that result being a true negative or, because the model simply does not have the *statistical power* to find an overcharge, a false negative.

This is not to say that all evidence should be assessed on a purely quantitative and probabilistic basis. As the judge said: "*an assessment or quantification of damages involves the taking into account of all manner of risks and possibilities. (…) Fundamentally, the process is evidence driven, and it is difficult to be very prescriptive.*"[63] However, an understanding of how the concepts of *statistical significance*, *statistical power*, and *positive predictive value* fit together formally is very important to understanding the probative value of the statistical evidence.

Our views chime with academic literature regarding the credibility of statistical research in a range of social science fields, which has found that serious problems can result from a lack of attention to *statistical power* in particular. For the reasons set out here we think these problems are also likely to bedevil the EC's own meta-study of likely cartel effects, and that this would be a fruitful area for further research. Finally, we are likely to see more explicit scrutiny of the concept of *statistical power* in class action situations where consideration of proposed modelling techniques takes place at the certification stage in advance of the actual analysis being conducted.

[63] *BritNed v ABB*, paragraph 12(6).