How severity analysis can help courts better assess econometric evidence

Rethinking statistical significance when quantifying cartel damages

Jorge Padilla, Thilo Klein, Peter Bönisch, Rahel Krauskopf, and Milo Minella[1]
18 June 2026

When determining cartel damages, courts typically draw on multiple sources: legal presumptions, prior case law, and empirical estimates. Each is uncertain, but not in the same way, nor to the same extent. Often, they point in different directions. How should the court decide?

In this article, **Jorge Padilla**, **Thilo Klein, Peter Bönisch**, **Rahel Krauskopf**, and **Milo Minella** argue that a framework is needed that enables courts to better assess how much confidence can responsibly be placed on empirical estimates, which in turn helps weigh empirical evidence against other sources of evidence. The severity principle provides that framework. It reveals how capable an empirical analysis is of testing each specific claim and, therefore, how strongly it contradicts or supports it in practice.

The views expressed in this article are the views of the authors only and do not necessarily represent the views of Compass Lexecon, its management, its subsidiaries, its affiliates, its employees, or its clients.

1 The challenge for courts: weighing empirical evidence alongside other sources

More than a decade after the adoption of the EU Damages Directive, private enforcement of antitrust law has become an important complement to public antitrust enforcement in the European Union. Courts now routinely determine and quantify damages arising from infringements such as cartels. In doing so, they often weigh econometric estimates against other sources, including legal presumptions and related judgments.

1.1 Competing sources of evidence

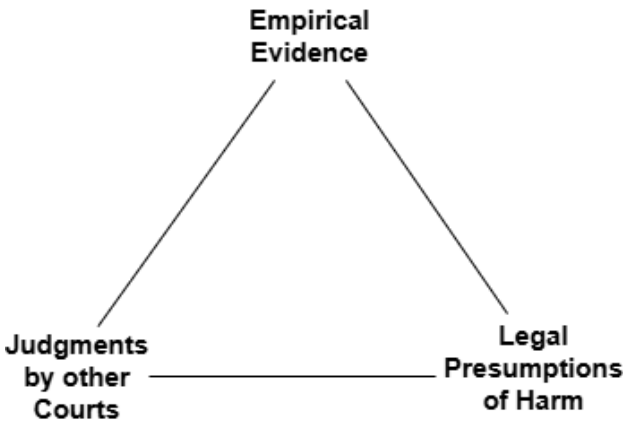
When deciding on damages, courts frequently face a difficult evidential triangle. Three sources of evidence – using the term in its broadest sense – may be available, each with different evidential value.

- **Empirical evidence.** Economic experts for both claimants and defendants often submit estimates of the effects that antitrust law infringements, such as cartels, have on market parameters such as prices or sales volumes. These estimates often require sophisticated economic and statistical analysis. However, no empirical estimate is perfectly precise; it is subject to a degree of statistical uncertainty.[2] The regression results presented to the court, therefore, include both (a) a point estimate, which is the model's central and, therefore, "best" estimate given the data, model specification and assumptions used; and (b) metrics that help assess the degree of statistical uncertainty, such as standard errors, confidence intervals, and statistical significance tests.

- **Judgments by other courts**. Courts may also consider damages determinations made in related proceedings. In the EU, courts in different Member States may consider claims relating to the same antitrust law infringement. The Trucks cartel is an obvious example: damages actions arising from the same Commission decision have been brought before courts in several Member States. The parties and circumstances may vary from claim to claim, which affects the impact of the infringement. But there may also be many similarities. Understandably, courts consider the determinations that other courts have made on related claims. The extent to which these determinations truly apply to the circumstances of the specific case is an empirical question and, therefore, uncertain.

- **Legal presumptions of harm**. The EU Damages Directive contains a rebuttable presumption that cartel infringements cause harm. This presumption concerns the existence of harm, not its precise amount. Depending on the jurisdiction, there may or may not be a quantified presumptive benchmark, for example a rebuttable presumption that a cartel caused an overcharge of 10 percent.[3] However, even where the presumption is not quantified, there is often scepticism that cartel harm was merely trivial or negligible. How much weight the presumption should carry in the specific circumstances is likewise uncertain.

Figure 1: Different sources of evidence: the evidential triangle courts face



Notes: The figure illustrates three broad categories that may inform the assessment of cartel harm: case-specific empirical evidence (e.g. regression analyses), findings in related proceedings, and legal presumptions of harm. The categories are complementary, not hierarchical.
Source: Illustration by the authors of this article.

If the three sources corroborate each other, there is little issue. But that is often not the case. Where the sources disagree, a court must assess how much confidence can responsibly be placed on each, both in absolute terms and relative to the others. Only then can it determine which magnitude is best supported by the overall evidential record.

1.2 The difficulty of weighing competing sources of evidence

Assessing how much confidence to place on a particular source is harder than it may seem.

An empirical estimate must be interpreted in light of two distinct considerations: the statistical uncertainty surrounding the estimate and the methodological choices and assumptions underlying the analysis:

- **Empirical evidence is statistically uncertain.** Even if the model producing the empirical estimate is correctly specified and its maintained assumptions are satisfied, the observed data represent only one realisation of the underlying economic process. A different realisation could have produced a different estimate. Therefore, even if the true price effect of an infringement

was exactly 8 percent, the observed data would yield estimates that fluctuate around 8 percent but would almost never equal 8 percent exactly. The same reasoning applies if the true price effect was zero; the estimated effect will typically be close to zero but almost never exactly zero. The court requires tools that allow it to meaningfully appreciate and evaluate that degree of statistical uncertainty.

- **Empirical evidence relies on methodological choices that may be contested**. In damages cases, it is common for experts to present materially divergent estimates. While both are statistically uncertain, typically that alone will not explain the gap between them. The wide disparity reflects differences in the empirical approaches that produced each estimate and in particular the assumptions underlying them. That may include the explanatory variables each expert considers relevant, the data they analysed, the counterfactual scenario each considered, or other differences. Every empirical analysis therefore rests on assumptions and choices that must themselves be tested and justified. The adequacy of the model, and therefore the weight that can responsibly be placed on its results, remains uncertain.

By comparison, courts may place substantial weight on legal presumptions and related judgments, regarding them as informative, and potentially more informative than empirical estimates. However, like empirical estimates, these alternative sources are not necessarily decisive. Parallel judgments may be informative without being binding. Legal presumptions may be useful, but they remain legal constructs rather than direct measurements of harm. Each has limitations, and potentially important ones.

Courts therefore often struggle with assessments of this kind: they may give too little weight to informative but statistically uncertain empirical evidence, or too much weight to sources whose uncertainty is less visible or quantifiable. This may be at least in part due to the fact that statistical significance tests make the statistical uncertainty associated with empirical estimates explicit (albeit in a limited and binary way), whereas no quantification is available for the (potentially considerable) uncertainty of other sources of evidence.

1.3 The trouble with metrics of statistical uncertainty

The problem is partly that courts ask whether a result is significant in the ordinary sense of being meaningful or informative, while economists answer the narrower question whether it is statistically significant. Statistical significance has a specific meaning that does not align with the meaning of "significance" in natural language.

An example is the 2021 Trucks II judgment of the German Federal Court of Justice ("BGH"). It determined that a regression analysis may constitute a relevant piece of evidence for or against the existence of a damage if "*it has been carried out methodologically correctly, on a sufficiently reliable data basis, and with significant results*".[4] Here, *statistically* "significant" results are interpreted with the natural language meaning: i.e., that they are good, meaningful and informative, whereas insignificant results are not. However, if that is what courts mean when they ask about the significance of the empirical results it is a question that should be answered. The issue is that telling them whether an estimate is *statistically* significant does not answer that question.

For reasons we explain below, the statistical significance test does not convey how certain or informative an estimate is in the general sense. Rather, it asks a narrow question: if there was no effect at all ("the null hypothesis"), how unusual would the observed estimate be? If it is surprising, the null hypothesis is rejected and the estimate is labelled statistically significant; otherwise, it is not.

Conflating this statistical test with natural language meaning of "significance" means that results are often interpreted in specific ways that go beyond what statistical tests are designed to show.

- **It would be incorrect to infer from a statistically significant result that the point estimate is likely to be accurate.** A statistically significant result only says that, if there is truly no effect, then the estimated effect would be unusual. It may warrant rejection of the null hypothesis at the chosen significance level, but it does not establish the probability or accuracy of the specific effect estimated.
- **It would also be incorrect to consider that a statistically insignificant result is uninformative and should be disregarded.** Such interpretations treat insignificant estimates as "just noise" and incapable of providing valid information about the scale of the likely effect. However, if the true effect is small, then an estimate may be both relatively certain and also not significant. In fact, given that significance always tests the plausibility of the null hypothesis, it is impossible for an estimate of zero effect to be statistically significant, no matter how good that estimate is – because, obviously, it could never be unusual enough to rule out the possibility that there is truly no effect.
- **Finally, it would be incorrect to infer from a statistically insignificant result that there is no effect.** Practitioners sometimes 'simplify' a statistically insignificant result in this manner. This misinterpretation assumes that failure to detect an effect proves that there was none to detect. That ignores the possibility that the empirical test simply lacked the capability to detect an effect of a relevant magnitude. Courts have explicitly acknowledged this interpretation as fallacious.[5]

The root problem may be that tests of statistical significance, and metrics of statistical uncertainty more generally, are incapable of providing the answers that courts require.

Firstly, metrics such as statistical significance tests and confidence intervals capture only the degree of statistical uncertainty surrounding an estimate conditional on the data, specification, and assumptions used. They do not, by themselves, answer the broader question of how much confidence the court should place in the empirical analysis or the point estimate derived from it. Confidence intervals have similar problems. They are often misinterpreted as showing that there is a 95 percent probability that the true effect lies within the interval around the estimate, which is not correct.[6] This may lead to a similar misleading binary logic as a misinterpreted significance test: the point estimate is treated as determinative when the interval is considered sufficiently narrow, but the empirical evidence is treated as uninformative once the interval exceeds a particular width. Neither response helps the court evaluate the degree of statistical uncertainty or assess how the empirical evidence compares with competing claims.

Secondly, even properly interpreted, these metrics of statistical uncertainty do not allow alternative sources of evidence to be examined in an informative way. Those sources of evidence have their specific evidential limitations, but these limitations cannot be quantified in the same way as statistical uncertainty can be made transparent. As a result, the court is often left without a practical framework for assessing how these alternative sources of evidence compare with the empirical estimates. Nor can it assess the extent to which the empirical evidence contradicts or supports effects of the magnitude suggested by those sources. Ultimately, this comparison is what a court needs to know.

The relevant question for courts is: how strong is the evidential support for a particular claimed overcharge or determination?

To answer that question courts do not need a better explanation or understanding of statistics. Rather, economists need to offer a framework that helps courts better assess not only statistical uncertainty, but also how thoroughly competing determinations have been tested against the

evidence, and how strongly the evidence supports or contradicts them. The severity principle and statistical severity analysis provide such a framework.

1.4 Using empirical evidence to test legally relevant magnitudes

The crucial point is that every empirical estimate – statistically significant or not – is informative. It carries information that allows us to assess, given the point estimate we observe and the statistical uncertainty surrounding it, how strongly the data speak against or in favour of a claim that the effect has a specific magnitude. The framework that allows us to evaluate the evidential value of an estimate is statistical "severity analysis".[7,8,9]

It is useful to distinguish the severity principle from statistical severity analysis.

The severity principle concerns the relationship between a claim, the evidence offered in its support, and the test that generated that evidence. Evidence deserves greater weight when the test would probably have revealed a relevant error had the claim been wrong. By contrast, evidence deserves less weight when the same result could readily have arisen even if the claim were false. In this sense, severity asks not merely whether the evidence is consistent with a claim, but whether the claim has survived a test capable of detecting relevant contrary evidence.

Statistical severity analysis applies that principle to an econometric estimate by asking which effect sizes the empirical evidence – including both the point estimate and its statistical uncertainty – is capable of contradicting if they were true. It does this by adapting the logic of a statistical significance test. A statistical significance test asks whether the observed estimate would be unusual if the null hypothesis of no effect were true. Statistical severity analysis asks how unusual the estimate would be under other non-zero effect size hypotheses.

That answers a more court-facing question: how capable was the empirical evidence of producing an estimate more favourable to a specific claimed overcharge and, therefore, how strongly does the observed evidence speak against that magnitude of overcharge if no such estimate was observed?

The court can therefore move beyond asking only whether the empirical evidence rules out the possibility of no overcharge at all. It can ask whether the evidence is capable of ruling out, or supporting, overcharge claims concerning specific and more legally relevant magnitudes – e.g., a 5 percent overcharge, or a 10 or 25 percent overcharge. For each, the court can assess how likely the empirical method would have been to produce an estimate more favourable to that hypothesis than the point estimate that was actually observed.

Illustration of statistical severity analysis in a cartel damages case

Consider a court assessing the likely overcharge in a cartel case. For illustrative purposes, let's assume that the court is presented with three different claims.

- A claim that there is no overcharge, i.e. that the true overcharge is 0 percent.
- A claim that the true overcharge is 5 percent.
- A claim that the true overcharge is 10 percent.

Consider also that one party submitted an empirical overcharge analysis, which finds a statistically insignificant effect of 1 percent.

The court now wants to know how consistent each of the three claims is with the empirical evidence underpinning the 1 percent estimate. **Table 1** presents the results of a statistical severity analysis applied to this case.

We present the results in two ways: the first quantifies the capability of the test to detect an overcharge of interest; and the second assesses the compatibility of the estimate with the claimed overcharge. Both ways lead to the same conclusion but offer a different perspective. Neither quantity is the probability that the claimed overcharge is true.

- The **Capability Assessment** shows how capable the empirical test is of estimating an effect larger than the one actually observed *if the claimed overcharge were true*. The more capable the test is of doing so for the claimed overcharge, the more strongly that claim can be rejected if an effect of this size is not observed. For example, under the assumption of a 5 percent overcharge, we would have estimated an overcharge larger than the 1 percent actually estimated with 91 percent probability. That is, the test was very capable of detecting an overcharge of this size, but it did not. On that basis, the court may decide to reject an overcharge of 5 percent or more.
- The **Compatibility Assessment** shows how likely it would be to obtain an estimate no greater than the estimate actually observed *if the claimed overcharge were in fact true*. This is the other side of the coin of the Capability Assessment. Therefore, there is only a 9 percent probability that an overcharge of 1 percent or less would have been estimated if the true overcharge was 5 percent. This is quite unlikely, or in other words the empirical evidence is incompatible (in the sense of difficult to reconcile) with the claimed overcharge. The court may therefore conclude that the statistically insignificant estimate of 1 percent warrants ruling out an overcharge of 5 percent or more.

Table 1: Using statistical severity analysis to assess the evidential support for claimed overcharges

Claimed overcharge	Capability assessment Probability of obtaining an estimate **larger than the estimate actually observed**, assuming that the claimed overcharge is true	Compatibility assessment Probability of obtaining an estimate **no larger than the estimate actually observed**, assuming that the claimed overcharge is true
Claim based on misinterpretation of the statistical test: 0%	37%	63%
The empirical point estimate: 1%	50%	50%
Claim from parallel judgments: 5%	91%	9%
Claim based on legal presumption: 10%	>99%	<1%

Source: Analysis based on simulated data.

In this article, we discuss how statistical severity analysis can help courts assess the extent to which statistical uncertainty contributes to differences between competing sources of empirical evidence. Statistical uncertainty is important, but it may not be the only relevant factor.

Consider the illustrative example above. A second expert, acting for the claimant, produces an alternative estimate of a 12% overcharge. The court must decide which estimate provides the more informative guide. The fact that the larger estimate is statistically significant while the smaller

estimate is not, does not establish that the larger estimate is more accurate. It may be, but it may not be. Statistical significance merely indicates that the larger estimate would be unlikely if the true effect were zero. Conversely, the lack of statistical significance for the 1% estimate means only that such a result would not be unusual if the true effect were zero. Neither statistic tells the court which estimate is more reliable.

Statistical uncertainty may explain some or all of the difference between the two estimates. Statistical severity testing can help determine whether that is so. In many cases, however, substantial differences between estimates reflect differences in empirical approach rather than statistical uncertainty alone. Those differences may arise from the data analysed, the identifying assumptions adopted, the counterfactual chosen, the statistical assumptions imposed, or other modelling decisions.

To choose between competing estimates, those methodological differences must be examined and assessed. The severity principle can assist that broader inquiry, but that is not the focus of this article. Here, we concentrate on the role of statistical uncertainty.

In the next two sections, we set out the underlying principles of hypothesis testing, which explain the limitations of the statistical significance test, and the advantages of statistical severity testing.

We then further develop the illustrative example set out here, to demonstrate practical steps that help courts use empirical evidence to better inform damages proceedings and discriminate between the various sources they must consider.

2 The limits of conventional statistical significance testing

Statistical significance testing is a familiar tool for assessing econometric evidence. But the conventional test answers only a narrow question: whether the observed estimate is sufficiently unusual under the hypothesis that there was no effect at all. It does not assess the evidential support for claims of specific magnitudes.

2.1 What is the statistical significance test?

Statistical significance or insignificance is the outcome of a statistical hypothesis test. Such a test assesses whether the observed evidence – in a cartel damages case, the estimated overcharge – is sufficiently inconsistent with a previously specified hypothesis, the so-called “null hypothesis”. Conventionally in cartel cases, the null hypothesis is that the cartel had no price effect or, in other words, the overcharge is zero. If the observed evidence indicates a large and positive overcharge, the test may justify the rejection of that null hypothesis.

To determine whether that is the case, the statistical significance test evaluates whether the observed overcharge is larger than could plausibly be explained by random variation or statistical noise. It does that by assessing, under the assumption that the null hypothesis is true, the probability of observing an overcharge at least as far away from the null hypothesis as the one observed in the data. This probability is called the p-value.

Depending on whether that probability is sufficiently small or not, the test leads to one of two outcomes:

- If the probability is sufficiently small, the null hypothesis is rejected, and the estimated effect is labelled “statistically significant”.
- Otherwise, the null hypothesis is not rejected and thus the estimated effect is labelled “not statistically significant”.

The threshold for what counts as sufficiently small is chosen to suit the case at hand. Conventionally, that probability is measured against one of three possible "significance levels", often indicated with asterisks: 1 percent (***), 5 percent (**), and 10 percent (*). In economics, we typically label an estimate "statistically significant" if the probability of observing this estimate (or a larger deviation from the null hypothesis of no effect) is smaller than 5 percent under the null hypothesis. That is, we employ the 5 percent significance level.

2.2 How does the hypothesis test work?

Any empirical overcharge analysis provides a point estimate, that is the model's central estimate of the true effect given the data, specification and assumptions used. But the estimate is subject to statistical uncertainty and does not need to be equal to the true effect,[10] because the observed data represent only one possible realisation of the underlying unknown real-world process that generated them. Therefore, even if the true price effect of an infringement was exactly 8 percent, the observed data would yield estimates that fluctuate around 8 percent but would almost never equal 8 percent exactly.

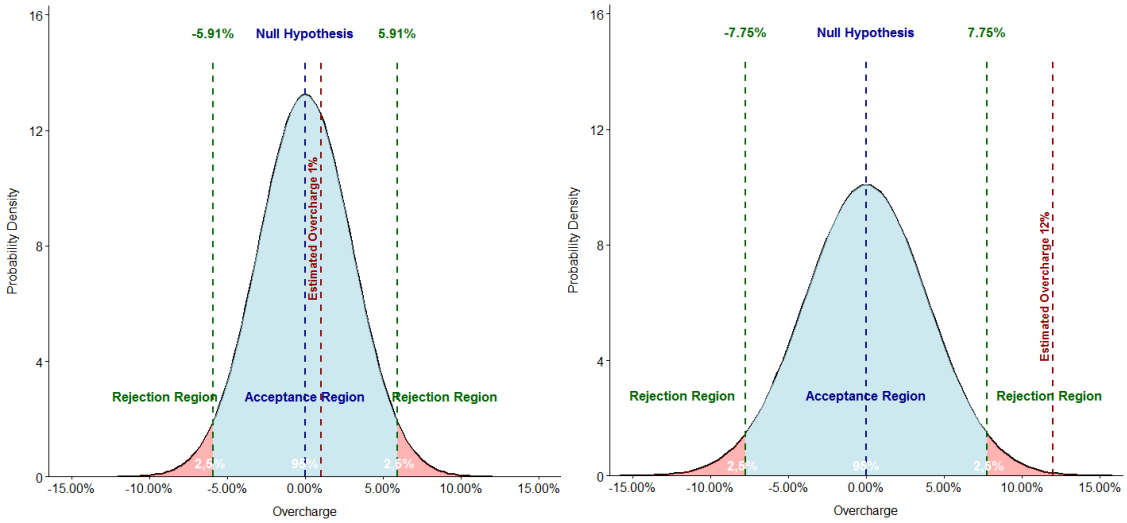
The statistical significance test asks, how unusual would the observed estimate be if the true effect were zero. It does not measure the probability that the true effect is zero. Nor does it establish whether the observed estimate is accurate. If there was in fact no effect, then the same reasoning about fluctuating estimates will arise: the estimated effect will typically be close to zero, but not exactly zero. A hypothesis test therefore asks whether the observed estimate is far enough away from the value stated in the null hypothesis that random variation alone is no longer a plausible explanation for the observed estimate.

To do so, the test compares two rival hypotheses:

- The **null hypothesis** usually states that there is no effect – in this case, that the true price effect is zero.[11]
- The **alternative hypothesis** always states the opposite of the null hypothesis, that is, it conventionally states that the true effect is not zero. Thus, the alternative hypothesis covers a whole range of possible values rather than one specific number: any price effect different from zero falls under the alternative hypothesis. A two-sided test therefore treats both sufficiently positive and sufficiently negative estimates as inconsistent with the null hypothesis.

Figure 2 illustrates this logic using two examples. In both panels, the bell-shaped curve shows the distribution of estimates that would be expected if the null hypothesis were true – here, if there was no overcharge due to the infringement, i.e. the true effect was zero. The bell-curve is therefore centred around the null hypothesis of no overcharge. The width of each bell-curve reflects the degree of statistical uncertainty associated with the respective estimate. The area under each curve represents probability: it shows how likely different estimates would be if the null hypothesis were true.

Figure 2: Statistical hypothesis test against the null hypothesis of no overcharge for different estimates



Source: Analysis based on simulated data.

In the figure, 95 percent of that area lies in the central region around zero, meaning that estimates in this range are treated as sufficiently consistent with the null hypothesis that it is not rejected. The remaining 5 percent lies in the two tails, with 2.5 percent in each tail, adding up to the pre-defined significance level of 5 percent. These tail areas are therefore called the **rejection regions**. They contain estimates that would be so unlikely if the null hypothesis were true that the null hypothesis is rejected. If the estimate falls in one of those regions, the result is labelled statistically significant and treated as supporting the alternative hypothesis. Under the null hypothesis, less extreme estimates fall into the blue area in **Figure 2**, also labelled the **acceptance region**.

It is essential to note what such a significance test is actually designed to measure. The bell-curves, the rejection regions, and the 5% threshold are all defined under the assumption that the null hypothesis is true. **Statistical significance is therefore a statement about how unusual the observed estimate would be if the null hypothesis were true. It is not a statement about the probability that the null hypothesis, or any competing hypothesis, is true given the data**.

The other crucial point is to recognise a common misconception. Statistical significance does not convey accuracy; and insignificance does not convey inaccuracy. In both cases, the point estimate is the best estimate in the sense that it constitutes the model's central estimate. However, the point estimate may be relatively accurate or inaccurate. Its actual accuracy is unknown. The degree of statistical uncertainty and the power of the test indicate how effectively the empirical method can distinguish between relevant hypotheses. Whether that evidence is enough to dismiss the null hypothesis as insufficiently compatible with the observed estimate is a different question. An insignificant estimate can be relatively precise. A significant one can be relatively imprecise.

Illustrating this, the two panels in **Figure 2** differ in terms of the statistical uncertainty surrounding the estimates. This is reflected in the different width of the bell-curves.

- In the left-hand panel, the number of observations is larger, so the distribution is narrower and the estimated overcharge of 1 percent is subject to less statistical uncertainty. Nevertheless, the observed estimate remains within the acceptance region and is not statistically significantly different from zero.

- In the right-hand panel, the number of observations is smaller, so the distribution is wider and the estimate is subject to greater statistical uncertainty. Here, the empirical method is less capable of distinguishing the null hypothesis from smaller deviations from zero than in the left-hand panel. However, the observed estimate of 12 percent is large enough that it still falls in the rejection region and is statistically significantly different from zero.

The figures therefore show that statistical significance depends not only on the size of the estimate – and thereby the deviation from the null hypothesis – but also on how precisely it is measured.

3 How statistical severity analysis improves on conventional significance testing

Statistical severity analysis builds directly on the same error-probability logic as statistical hypothesis testing. Instead of stopping at the binary question whether the null hypothesis of no effect can be rejected, it asks what the evidence reveals about specific magnitudes of the effect of interest.

3.1 How does statistical severity testing work?

The reasoning builds directly on the result of a standard test. When the null is not rejected, the outcome may simply reflect that the test had little capacity to detect a real effect — in which case non-rejection is only weak evidence that the effect is absent. Statistical severity analysis addresses this limitation by asking a different question: *which effect sizes are inconsistent with the observed estimate, given the capacity of the test?*

Consider a hypothetical cartel case in which the regression model yields a statistically insignificant overcharge estimate of 1 percent (left panel of **Figure 2**). This means that we cannot reject the null hypothesis of no effect. At the same time, it does not tell us that there was no effect. Obviously, this response is unsatisfactory to a judge who needs to decide on the magnitude of the harm. This is where statistical severity testing comes in.

A statistical severity test instead asks: if the true overcharge was some level the court regards as relevant – say 5 percent – how likely is it that the empirical method would have produced an estimate larger than the observed estimate of 1 percent if the overcharge was indeed 5 percent?

To answer this, we calculate the probability of obtaining an estimate larger than 1 percent under hypothesised alternative overcharges such as 5 percent, 10 percent, or any other value of interest:

- If that probability is high – i.e., had the true overcharge been 5 percent, the test would very likely have produced an estimate above the observed estimate of 1 percent – then the failure to observe such an estimate is strong evidence that the true overcharge is below 5 percent. The claim that "the overcharge is no greater than 5 percent" passes with high severity, and a true overcharge of 5 percent or more can be ruled out.

- If that probability is low, the test lacked the capacity to distinguish a 5 percent overcharge from the estimate we actually observed. The claim that "the overcharge is no greater than 5 percent" passes only with low severity, and a true overcharge of 5 percent or more cannot be ruled out.

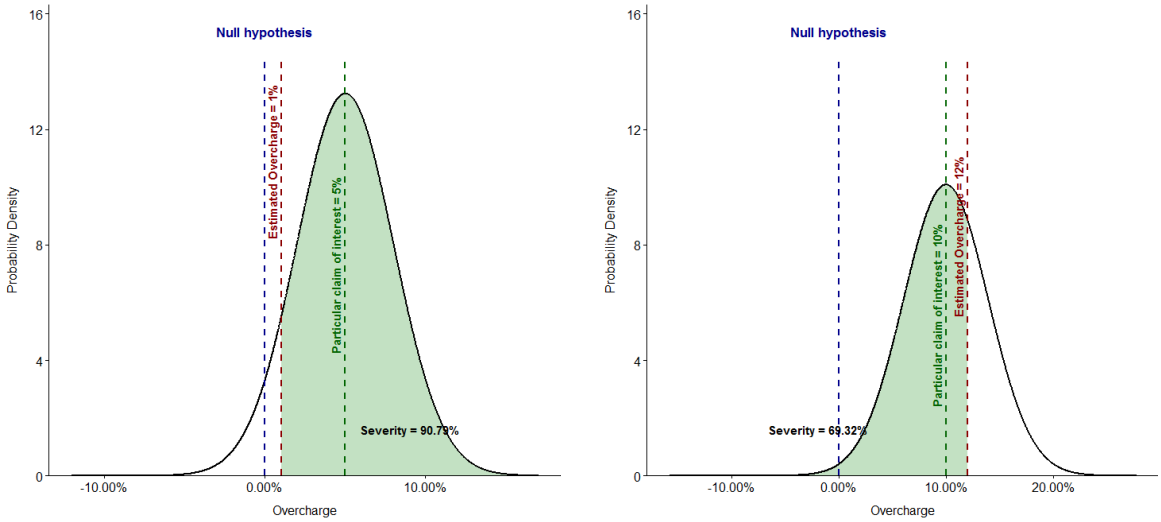
The intuition is simple: a test that had the capacity to produce a larger estimate if a higher overcharge were true, but did not, provides evidence against that higher overcharge. Certain overcharge magnitudes can therefore be ruled out even when the point estimate is statistically insignificant in the conventional sense.

Figure 3 illustrates this graphically using the bell-curves introduced in **Figure 2**:

- The left panel depicts the statistically insignificant estimate of 1 percent. Severity asks: how likely would it have been to obtain an estimate above 1 percent if the true overcharge had been 5 percent? Graphically, this is the area under the bell-curve to the right of 1 percent. In our example, that probability is above 90 percent. The test therefore had a high capacity to produce an estimate above 1 percent had the true overcharge been 5 percent; the fact that it did not warrants the conclusion, with high severity, that the true overcharge is below 5 percent.

- The right panel applies the same logic to a statistically significant result. Here, severity asks how likely it would have been to obtain an estimate smaller than the observed 12 percent if the true overcharge had been only 10 percent. If that probability is high, the fact that the observed estimate is 12 percent provides strong evidence that the true overcharge indeed exceeds 10 percent. If that probability is low, however, the 12 percent estimate provides weak evidence that the true overcharge exceeds 10 percent. In the example on the right panel of **Figure 3**, this probability is not overwhelmingly high that a judge may not want to exclude lower magnitudes on that basis.

Figure 3: Statistical severity test against a particular claim of interest for different test outcomes



Source: Analysis based on simulated data.

Severity therefore allows courts to extract information about the possible magnitude of the effect from *both* statistically insignificant and statistically significant results. In the former case, it does so by ruling out larger effects and, in the latter, by supporting claims that the true effect exceeds a particular magnitude of interest. The analysis identifies magnitudes that can be ruled out or remain unexcluded. It does not assign probabilities to possible true effects.

Statistical severity analysis, therefore, avoids misinterpretations by shifting the focus from whether a null hypothesis can be rejected to assessing how strongly the estimated overcharge speaks against an effect that is claimed to have a specific size. Importantly, the actual estimated overcharge matters for the severity assessment, whereas a binary interpretation of a standard statistical significance test (or confidence intervals) treats *all* estimates within the rejection region, or within the non-rejection region, in the same way.

Therefore, while neither a statistically non-significant nor a statistically significant overcharge estimate can prove the absence of damage (i.e., an overcharge of exactly zero) or the existence of an effect of a particular magnitude, every estimate – statistically significant or not – carries

information about the effect sizes that can be ruled out or remain unexcluded, given the observed estimate and the capacity of the test.

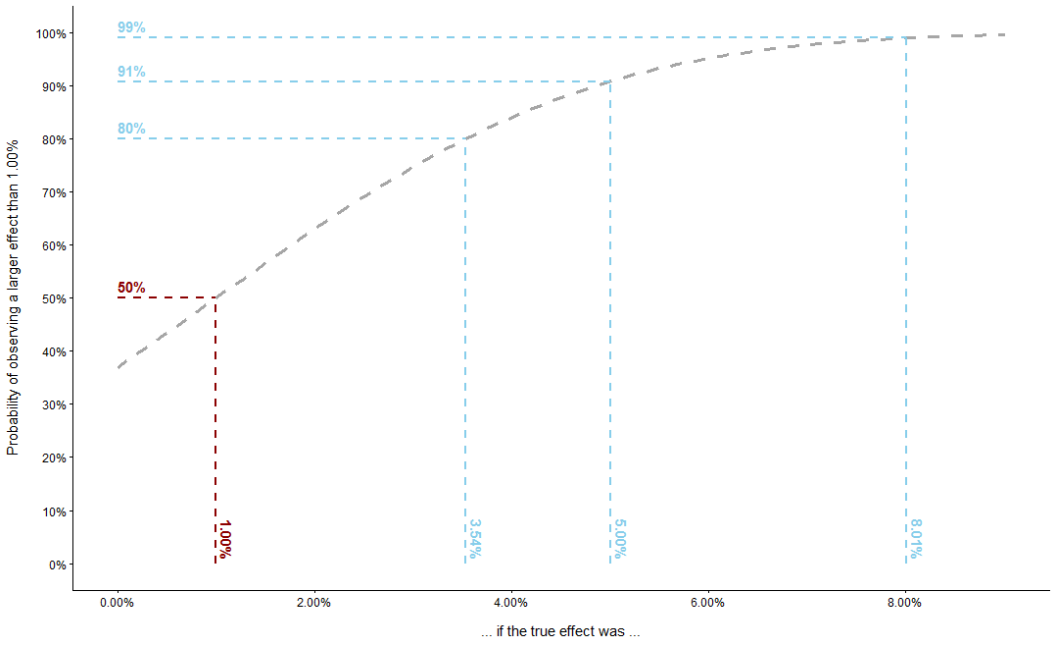
3.2 Assessing the full range of potential effects

In section 1, we showed how capable the empirical evidence was of ruling out specific overcharge claims that the court might consider, and how compatible the evidence was with each claim. Despite the statistically insignificant estimate of 1 percent, the evidence is reasonably capable of detecting overcharges of 5 percent or 10 percent if either were the true effect. The fact that the observed estimate is much smaller makes those claims difficult to reconcile with the evidence.

While useful, statistical severity testing is not limited to showing the evidential support, or lack of it, for specific claims. We can assess the evidential support for the full range of possible effects. **Figure 4** plots the capacity assessment for every candidate alternative overcharge. It produces an upward-sloping severity curve. Two features help read it:

- At a hypothetical true overcharge equal to the estimate itself, severity is always 50 percent (indicated by the red colour). This is intuitive: If the true effect were 1 percent, the probability of estimating an effect above 1 percent is exactly 50 percent.
- The further right we move along the horizontal axis, the higher the probability that we would observe an estimate above the 1 percent that we actually observe – and therefore the more severe the test of that claim is, and more confidently we may choose to rule it out.

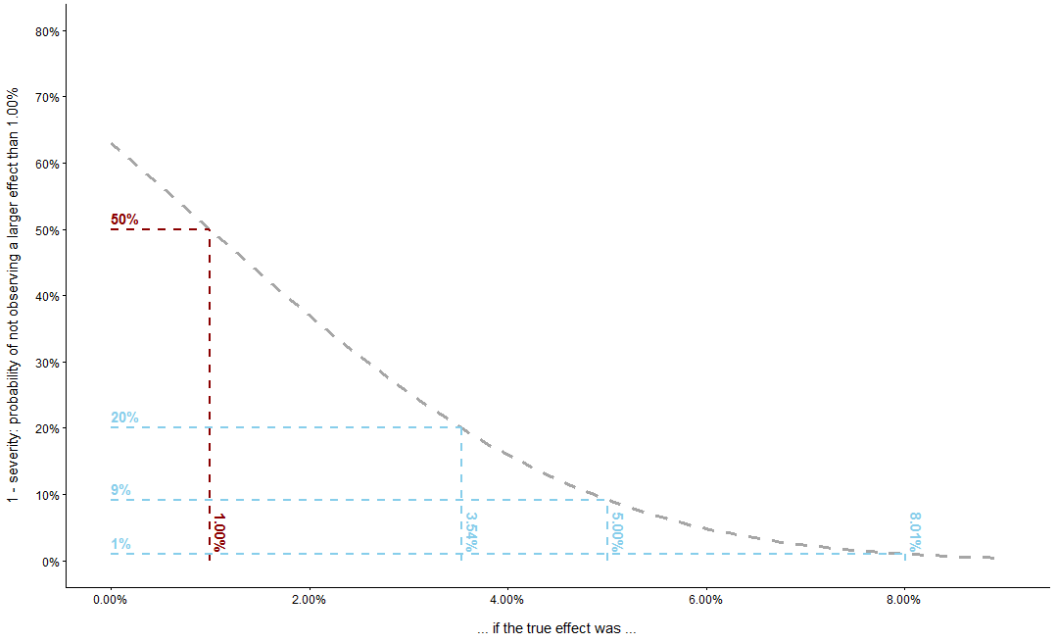
Figure 4: The capacity assessment, showing the evidence's ability to detect all potential claimed overcharges



Source: Analysis based on simulated data.

Figure 5 shows the **compatibility assessment.** It shows the same information for the alternative perspective. The larger the claimed overcharge is, the less compatible that becomes with the empirical evidence, given that we observe an estimate of only 1 percent.

Figure 5: The compatibility assessment between the evidence and all potential claimed overcharges



Source: Analysis based on simulated data.

These charts can be useful, as the court may determine for itself the degree of evidential support for a particular claim it is prepared to tolerate. Unlike the binary presentation of a statistically significant test, or the undifferentiated region of a confidence interval, it can evaluate the amount of evidential support each claim has. It can answer the more practical question: what is the smallest hypothetical true overcharge for which an estimate above 1 percent would have been sufficiently likely that the failure to observe one effectively rules that magnitude out?

For illustrative purposes, in the charts above we show an 80 percent severity threshold (or 20 percent compatibility threshold). On that basis, the analysis shows that any overcharge of about 3.5 percent or above can be ruled out given the observed estimate of 1 percent. This is illustrated in **Figure 4** by the point at which the grey severity curve intersects the vertical line at an overcharge of 3.5 percent.

Statistical severity analysis augments the conventional testing logic. Once the significance test outcome, the estimate, and its associated uncertainty are known, severity uses the same estimate and its statistical uncertainty to map which specific magnitudes the data can rule out. When the null is not rejected, severity bounds the effect from above, identifying which discrepancies from zero can still be excluded. When the null is rejected, severity bounds the effect from below, identifying the lower-bound claims that the data positively support.

Either way, rather than stopping at the binary verdict that the estimate is either statistically significantly or statistically insignificantly different from zero, the court is told what the evidence actually shows about the possible size of the overcharge.

4 Using severity testing in practice

This section illustrates how the severity framework can help courts assess empirical evidence from different sources in damages proceedings. The first example considers how a court may evaluate a small statistically insignificant estimate alongside a legal presumption and related judgments. The

second considers a more difficult situation in which two experts present materially divergent empirical estimates.

Before placing weight on any empirical estimate, the court should assess how thoroughly the underlying empirical model has been tested. Each expert should explain how the quality of the data, the chosen counterfactual, the identifying assumptions, the statistical assumptions, and the specification choices underlying the analysis have been tested for relevant error. The broader severity principle applies to this assessment: the relevant question is whether those tests were genuinely capable of revealing material violations of these assumptions, had such violations been present.

Once the court is provisionally satisfied that the empirical approach has been sufficiently tested, statistical severity analysis can help it assess what the estimate and the statistical uncertainty surrounding it reveal about claims concerning the magnitude of the effect. It can show which magnitudes the evidence strongly contradicts, and which remain unexcluded.

For explanatory purposes, the examples below proceed in stages. The first stage focuses on statistical severity analysis, assuming that the underlying empirical approach has been sufficiently probed. The second stage considers competing empirical approaches and shows how the broader severity principle becomes relevant where differences between their results cannot be explained by statistical uncertainty alone.

4.1 First illustrative example: testing alternative claims against a small statistically insignificant estimate

Consider a case in which an expert estimates an overcharge of 1 percent, but the result is not statistically significantly different from zero. The applicable legal framework contains a presumption of harm or a presumptive overcharge of 10 percent. Meanwhile, other courts dealing with the same infringement but different claimants, periods, jurisdictions, or evidential records, have awarded or accepted an overcharge of 5 percent.

For illustrative purposes, we assume in this example that the court is provisionally satisfied that the empirical approach has been sufficiently probed. The narrower question is then what the estimate of 1 percent and the statistical uncertainty surrounding it reveal about the different magnitudes the court may consider.

Step 1: identify the propositions before choosing the number

The empirical analysis produces a central estimate of 1 percent. The legal framework supplies a presumption of harm or a presumptive benchmark of 10 percent. The related judgments supply another candidate figure of 5 percent. The court may also wish to consider the claim that there was no overcharge at all.

These figures arise from different sources and may play different legal roles. For the purpose of statistical severity analysis, however, the court can initially treat each numerical figure as a candidate claim about the magnitude of the overcharge in the present case. It can then ask how strongly the empirical evidence contradicts or supports that claim.

Step 2: test the candidate magnitudes against the empirical evidence

The court can start with the presumption of 10 percent. It should ask: if the true overcharge were 10 percent, how likely would the empirical method have been to produce an estimate larger than the observed estimate of 1 percent? If that probability is high, the failure to observe a larger estimate provides strong evidence against an overcharge of 10 percent or more.

The court can apply the same analysis to the 5 percent figure from the related judgments. At this stage, the court need not assess whether those judgments were correctly decided on their own evidential records. The narrower question is whether a 5 percent overcharge is compatible with the empirical evidence in the case before it. If the true overcharge were 5 percent, how likely would the empirical method have been to produce an estimate larger than the observed estimate of 1 percent?

In our example in **Table 1**, that probability is 91 percent. The empirical method was therefore highly capable of producing a larger estimate if the true overcharge were 5 percent. The fact that no larger estimate was observed provides strong evidence against applying a 5 percent figure to the present case.

This does not imply that the related judgments were incorrect. The 5 percent figure may have been appropriate in the circumstances of those cases. The point is only that the empirical evidence in the present case speaks against the claim that the same figure applies here.

The court can also consider the claim that there was no overcharge at all. A statistically insignificant estimate of 1 percent does not, by itself, establish that claim. The absence of statistical significance should not be equated with evidence of no harm.

Step 3: identify the magnitudes that remain compatible with the evidence

The court should not respond by automatically adopting either the 1 percent point estimate or an intermediate figure. Instead, it should state how much capability it requires before treating a claimed magnitude as sufficiently contradicted by the empirical evidence.

Applying an illustrative severity threshold of 80 percent, the court can ask which hypothetical overcharges would probably have produced an estimate larger than the observed estimate of 1 percent. In this example, overcharges of approximately 3.5 percent or more can be ruled out at that threshold. Smaller magnitudes remain unexcluded.

The choice of threshold is not dictated mechanically by the method. It reflects the degree of evidential support the court considers sufficient in the circumstances of the case. Statistical severity analysis makes that choice transparent and shows its implications.

Step 4: reach a reasoned award

Statistical severity analysis does not necessarily identify a unique award. The point estimate remains the model's central estimate and therefore an important reference point. Statistical severity analysis adds a further piece of information: it identifies the magnitudes that the empirical evidence is sufficiently capable of contradicting.

The court may still need to rely on legal and factual considerations when selecting an award within the range that the empirical evidence does not rule out. But it can do so transparently. It should not award 5 percent merely because related judgments did so, nor adopt 10 percent merely because it is the presumptive benchmark, if the empirical evidence severely contradicts those figures in the present case. The conclusion remains conditional on the adequacy of the empirical approach. If the underlying data, counterfactual, assumptions, or specification choices have not themselves been sufficiently probed, the court should not place substantial weight on the statistical severity results.

4.2 Second illustrative example: two empirical analyses that appear to contradict each other

The second example is more difficult. One expert estimates a statistically insignificant overcharge of 1 percent. The opposing expert estimates a statistically significant overcharge of 12 percent. The applicable presumption remains 10 percent, while related courts have reached 5 percent.

The court can proceed in two stages.

- First, it uses statistical severity analysis to ask whether the apparent conflict can be explained by statistical uncertainty.
- If the two analyses remain irreconcilable it should move to the broader severity principle and assess the empirical models that produced them.

Step 1: apply statistical severity analysis symmetrically to both estimates, refusing to be impressed by labels alone

The court should first ask what each empirical analysis is capable of showing about legally relevant magnitudes.

For the statistically insignificant estimate of 1 percent, the relevant question is an upper-bound question: which larger overcharges would probably have produced an estimate above the observed estimate and can therefore be ruled out because no such estimate was observed? The fact that the 1 percent estimate is statistically insignificant establishes only that the null hypothesis of no effect cannot be ruled out. But the evidence may be able to rule out other larger overcharges, beyond an upper bound.

For the statistically significant estimate of 12 percent, the relevant question is a lower-bound question: which smaller overcharges would probably have produced an estimate below the observed estimate and can therefore be ruled out because the observed estimate was materially larger? The fact that the 12 percent estimate is statistically significant establishes only that the null hypothesis of no effect can be rejected at the chosen significance level. It does not, by itself, establish that the true overcharge exceeded 5 percent, 10 percent, or any other legally relevant benchmark.

It is important to keep in mind that each of these questions are asked assuming that the respective model that has produced the results is correct.

Step 2: ask whether statistical uncertainty reconciles the estimates

The two empirical analyses may initially appear contradictory without genuinely conflicting. The court should compare the claims that each analysis supports or leaves unexcluded, rather than comparing the two point estimates alone.

Suppose, for example, that the claimant's statistically significant estimate of 12 percent is subject to substantial statistical uncertainty and does not severely rule out overcharges below 5 percent. Suppose also that the defendant's analysis leaves some overcharges in that range unexcluded. The ranges then overlap. Statistical uncertainty explains at least part of the difference, and statistical severity analysis reconciles the apparently divergent estimates by identifying a common range of possible effects.

If, however, the defendant's analysis severely supports the claim that the overcharge was below approximately 3 percent while the claimant's analysis severely supports the claim that the overcharge exceeded 5 percent, the ranges do not overlap. Statistical uncertainty alone cannot explain the difference between the results, because it assumes that the underlying model is correct. However, one or both of the models may be incorrectly specified.

Step 3: apply the broader severity principle to conflicting underlying assumptions

Where statistical uncertainty alone cannot reconcile the estimates, the court must move from statistical severity analysis to the broader severity principle. It is no longer asking only what can be inferred from a particular point estimate and the statistical uncertainty surrounding it. It must assess which empirical approach deserves greater evidential weight.

The broader severity principle provides a disciplined framework for that assessment. The court should ask what errors each expert's approach was capable of revealing and which potentially important errors may have remained undetected.

In that situation, it is not enough to ask whether each expert has performed "robustness checks". The term captures a wide range of exercises, some of which provide meaningful tests of the approach and some of which do not. The more relevant question is whether the checks were genuinely capable of revealing an error that could materially affect the conclusion, had such an error been present. A check that leaves the disputed assumption effectively untested deserves limited weight. A check that meaningfully probes the vulnerability of the result deserves more.

The court should ask, in particular:

- How was the quality and completeness of the underlying data tested? Is there any selection applied to the data that might influence the result? Were the checks capable of revealing errors that could materially affect the estimated overcharge?
- How was the chosen counterfactual scenario justified? Were comparator markets or periods plausibly unaffected by the infringement?
- Which identifying assumptions were examined, and through which empirical sensitivity analyses?
- Has the statistical adequacy of the model been tested?
- Did each expert test weaknesses that could undermine their own preferred conclusion, rather than only the opposing analysis?

Not every assumption can be tested conclusively. Some may remain maintained assumptions. The purpose of the severity principle is not to eliminate judgment, but to make transparent how thoroughly each approach has been subjected to tests capable of revealing relevant error.

The detailed application of the broader severity principle to individual methodological choices is outside the scope of this paper. Here, we focus on the application of statistical severity analysis to statistical uncertainty. The broader principle nevertheless identifies the questions that experts should be required to answer where their empirical approaches produce irreconcilable results.

Step 4: compare the approaches on a common basis

The same discipline should be applied symmetrically to both experts. A claimant cannot rely on the label "statistically significant" without showing that the analysis has survived meaningful attempts to reveal alternative explanations for the estimated effect. A defendant cannot rely on a small statistically insignificant estimate without showing that the method was capable of detecting relevant magnitudes of harm had they existed.

The court can then explain its conclusion in structured terms. It may find that the estimates are compatible once statistical uncertainty is properly considered. Alternatively, it may find that the

approaches remain irreconcilable and prefer one because it has survived more capable tests of the errors most relevant to the dispute.

4.3 Implications

Severity testing is most useful when courts ask experts to present their evidence in judicially intelligible terms. Experts should not merely report estimates and their statistical significance. They should explain the claim tested, the errors the method was designed to detect, the errors it might miss, and the implications for legally relevant benchmarks such as legal presumptions.

A practical expert report should therefore include a plain-language severity table. For each key estimate, the expert should state the proposition tested, the data used, the counterfactual; the principal identifying and statistical assumptions and how they were tested, the magnitude of effect the model could reasonably detect; and the conclusion that follows.

Table 2: What Judges Should Ask from Experts

Item	Question for the expert	Judicial purpose
Point estimate	What is the central estimate?	Retain the estimate as an important reference point without treating it as determinative.
Statistical hypothesis test	Is the observed estimate sufficiently unusual if the true effect were zero?	Assess whether the null hypothesis of no effect can be rejected while controlling the risk of a false rejection at the chosen significance level.
Statistical severity	Which legally relevant magnitudes would the empirical method probably have contradicted if they were true?	Identify effects that can be ruled out or remain unexcluded.
Data	How was the data selected? How were the quality and completeness of the underlying data tested?	Identify errors that may distort the estimate.
Counterfactual	What is the counterfactual, and which identifying assumptions support it? What prices would have been but for the infringement?	Identify the causal benchmark and the assumptions required to interpret the estimate.
Statistical assumptions	Were the assumptions underlying statistical inference tested?	Assess whether standard errors and related metrics can be interpreted as intended.
Robustness	Does the result survive reasonable alternative specifications?	Test whether the conclusion is fragile.
External coherence	Is the result consistent with documents, market facts, and related judgments?	Check economic and legal plausibility.

Source: The authors of this article.

5 Advantages and limitations of statistical severity analysis

Statistical severity analysis offers several advantages.

First, it improves transparency. Rather than treating an estimate as informative merely because it is statistically significant, or disregarding it because it is not, the court asks what the evidence was capable of showing.

Second, it promotes symmetry. Claimants cannot rely on large statistically significant estimates without demonstrating that the underlying approach has survived meaningful attempts to reveal relevant error. Defendants cannot rely on statistically insignificant estimates without showing that their method was capable of detecting relevant harm had it existed.

Third, the framework is compatible with legal presumptions. It does not abolish presumptions, but clarifies how strongly empirical evidence speaks against, or is compatible with, the magnitudes they suggest.

The approach also has limits.

Statistical severity testing cannot create evidence where none exists. The broader severity principle cannot eliminate uncertainty about the quality of the data, the appropriateness of the counterfactual, or the validity of the assumptions underlying an empirical analysis. Nor can statistical severity analysis transform a poorly specified model into reliable evidence. Its quantitative conclusions remain conditional on the adequacy of the empirical approach and on the court's choice of the degree of evidential support required before treating a claim as sufficiently contradicted.

The framework does not dictate a unique award or replace the court's judgment about burdens of proof, presumptions, standards of proof, and the permissible use of approximation. Its contribution is more modest but still important. It makes clear which claims have been subjected to tests capable of revealing relevant error, which magnitudes the empirical evidence strongly contradicts or leaves unexcluded, and where judgment remains necessary.

The merit of the severity framework is therefore not that it supplies a magic formula. It provides a grammar of evidential responsibility. It asks parties to show not only that their preferred number can be calculated, but that it has survived a test capable of revealing relevant error.

---

[1] Jorge Padilla is Chair of International Board, Thilo Klein is an Executive Vice President, Peter Bönisch is a Vice President, and Rahel Krauskopf and Milo Minella are Economists at Compass Lexecon International. The authors thank colleagues for their feedback and support on earlier drafts, including Andrew Tuffin and Research Team.

[2] In this article, *statistical uncertainty* refers to the sampling variation associated with an empirical estimate: even if the model is correctly specified, a different realisation of the underlying economic process could have produced a different point estimate. The standard error estimates the extent of this sampling variation. Statistical uncertainty in this sense does not encompass all existing sources of uncertainty relevant to the assessment of an empirical analysis. In particular, it does not by itself capture uncertainty about the adequacy of the model specification, the appropriateness of the chosen counterfactual, the quality of the underlying data used to estimate the model, identification problems, selection effects, or the sensitivity of the results to reasonable alternative specifications. Any interpretation of the statistical results remains conditional on the adequacy of the data, model, and maintained assumptions.

[3] Article 17.2 of Directive 2014/104/EU. This is a rebuttable presumption of the existence of harm, not a presumption of any specific quantum, as such, once a cartel infringement has been established, the claimant does not need to prove from scratch that the cartel caused some harm at all. There is no figure attached, it only presumes that cartels generally generate harm. Hungary and Latvia have a rebuttable presumption that cartels cause an overcharge of 10 percent. Romania has a rebuttable presumption of 20 percent. See European Commission, Commission Staff Working Document on the implementation of Directive 2014/104/EU of the European Parliament and of the Council of 26 November 2014 on certain rules governing actions for damages under national law for infringements of the competition law provisions of the Member States and of the European Union SWD (2020) 338 final, 14 December 2020, page 9.

[4] BGH KZR 19/20 (13 April 2021), 'Trucks II', para. 66. In full, translated from German, the judgment reads: "While regression analysis – alongside other permissibly applicable methods – is in practice primarily used to determine the amount of damages, because it allows the difference to be calculated between prices on the market affected by the cartel and prices on the cartel-free comparator market, every positive difference above zero simultaneously implies that a cartel-induced price increase has occurred, and conversely a difference equal to or below zero rules out such a price increase. It is therefore equally suited to establishing the 'whether' of a loss. Provided it has been carried out methodologically correctly, on a sufficiently reliable data basis, and with significant results, it thereby also constitutes a relevant indication for or against the circumstance – to be established in the context of an interim judgment on the merits – that the claimant has probably suffered at least some damage in some amount as a result of the cartel infringement."

[5] In that case, the Stuttgart Regional Court criticised the statistically insignificant result presented by the defendants because it was based on a null hypothesis of no harm, which the Court considered to be in tension with the presumption of harm. However, the Court did not address the fact that the statistically significant result presented by the claimants was based on the same null hypothesis. The different treatment of the same methodological starting point suggests that the Court regarded the defendants' analysis as problematic not because of the null hypothesis itself, but because the defendants were unable to reject it. The implication appears to be that a null hypothesis of no harm is considered to contradict the presumption of harm only where it is not rejected.

[6] A confidence interval is calculated using a procedure with a specified long-run coverage probability. This means that, if the same procedure were applied repeatedly to new samples generated in the same way, a 95 percent confidence interval would contain the true effect in 95 percent of those samples. It does not mean that there is a 95 percent probability that the true effect lies within the particular interval obtained in an individual case. Nor does this long-run property alone establish that the width of the interval measures how precisely the true effect has been identified, or that the interval contains the substantively plausible values of the effect. In familiar regression applications, a conventional confidence interval may nevertheless provide a useful summary of statistical uncertainty in the defined sense because its width is determined by the estimated standard error. See Morey RD, Hoekstra R, Rouder JN, Lee MD, Wagenmakers EJ. The fallacy of placing confidence in confidence intervals. Psychon Bull Rev. 2016 Feb; 23(1):103-23), who distinguish and explain (1) the fundamental confidence fallacy, (2) the precision fallacy, and (3) the likelihood fallacy in interpreting confidence intervals.

[7] The concept of severity has been developed in particular by Deborah Mayo and Aris Spanos. For a concise introduction, see Deborah G Mayo and Aris Spanos, 'Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction' (2006) 57(2) *The British Journal for the Philosophy of Science* 323; for a more technical treatment, see Aris Spanos, *Probability Theory and Statistical Inference: Empirical Modeling with Observational Data* (2nd edn, Cambridge University Press 2019).

[8] In this sense, the severity approach can also be used to reconcile supposedly contradictory evidence. See Peter Bönisch and Roman Inderst, 'Using the Statistical Concept of "Severity" to Assess the Compatibility of Seemingly Contradictory Statistical Evidence (with a Particular Application to Damage Estimation)' (2022) 18(2) *Journal of Competition Law & Economics* 400.

[9] Abadie discusses the informational content of statistically insignificant findings within a Bayesian framework. Since this is not the framework typically applied by courts, we do not adopt a Bayesian perspective in this paper. The underlying conclusion is nevertheless similar: statistically non-significant results should not be dismissed as evidentially irrelevant, but may carry evidential weight depending on their precision, context and consistency with the other evidence. Alberto Abadie, 'Statistical Nonsignificance in Empirical Economics' (2020) 2(2) *AER: Insights* 193.

[10] See footnote 2 above.

11 The null hypothesis of no price effect is the most common null hypothesis in damages quantification, because the null hypothesis is typically a statement about the real world that the researcher seeks to test and, if justified by the data, to reject (see the next subsection). More generally, however, the null hypothesis need not be limited to no effect. It can be any statement about a parameter or effect of interest. The logic of the testing approach described in this paper would apply in the same way.