Does experts' adjustment to model-based forecasts contribute to forecast quality?

Philip Hans Franses
Rianne Legerstee

Econometric Institute
Erasmus University Rotterdam

Econometric Institute Report 2007-37

Abstract

We perform a large-scale empirical analysis of the question whether model-based forecasts can be improved by adding expert knowledge. We consider a huge database of a pharmaceutical company where the head office uses a statistical model to generate monthly sales forecasts at various horizons for various products in seven categories across thirty-five countries and where local managers can modify those model-based forecasts. To sensibly compare realizations and forecasts we develop a useful statistical methodology. Our main finding is that on average the model-based forecasts are about equally good with or without added expertise. We examine the possibility that the expert puts too much weight on his or her own contribution and we obtain strong evidence that this is the case.
Keywords: Judgemental adjustment; Expert forecasts

September 19 2007

We are very grateful to Sander Demouge, Youssef Boulaksil and Michel Smulders (N.V. Organon, Oss) for providing us with the database, and we thank Arnoud Swinkels, Enrico van Halm, Rob Schreuder and Willen Rutten for their helpful feedback on the presentation of our first results.

Address for correspondence:
Econometric Institute, Erasmus University Rotterdam, P.O.Box 1738, NL-3000 DR Rotterdam, The XXIII, franses@few.eur.nl

1. Introduction

It frequently occurs that model-based forecasts are adjusted by experts who have domain knowledge of the specific forecast situation at hand. Experts may know that institutions will change, that events have a specific impact that is not included in the model, or that a variable that is difficult to measure is missing in the model. There is scattered evidence that added expertise may yield better forecasts, see Blattberg and Hoch (1990) and Mathews and Diamantopoulos (1986, 1992) for early accounts, but most of this evidence concerns specific and detailed cases. In this paper we aim to shed light on this matter by examining a huge database containing model-based forecasts, expert-adjusted forecasts and realizations for monthly sales of pharmaceutical products, concerning thirty-five countries and seven product categories. An analysis of this enormous amount of information allows us to draw some generalizing statements on the relative contribution of expert on top of models.

The outline of our papers is as follows. First, in Section 2, we develop a useful and reliable statistical methodology to compare model-based forecasts and expert-adjusted forecasts. We need to do so, as it can easily be understood that expert-adjusted forecasts somehow nest the pure model-based forecasts, and hence the techniques in Clark and McCracken (2001, 2005) need to be implemented. Basically, other (and elsewhere already employed) methods would give biased results, in this case in favour of the model-based forecasts. Additionally, as we find those experts' adjustment shows strong autoregressive dynamics, all methods need to account for this. In Section 3, we apply our methodology to our large database. We test whether experts add knowledge in a systematic and significant manner, and next we examine if this would also lead to better forecast accuracy. We find that experts do matter, and that added knowledge and model balance each other on a 50-50 basis, but that forecast gain is not large. In fact, when experts do worse, they do seriously worse. On the positive side, when they do better it can be largely attributed to their adjustment. This finding seems to holds across forecast horizons, countries and categories. In Section 4 we examine the possibility that perhaps the experts exercise too much impact on the final forecasts, and we find overwhelming evidence for this hypothesis. In Section 5 we conclude with a summary of potential implications of our findings.

2. Methodology

In this paper we examine the accuracy of statistical model-based forecasts and of expert forecasts, where the expert has adjusted the model-based forecasts. We first deal with one-step-ahead forecasts and then with multi-step-ahead forecasts. We propose two statistical tests to examine if experts add knowledge that is relevant and whether it improves the quality of the forecasts.

2.1 One-step-ahead forecasts

In this study we examine whether experts make model-based forecasts better by adding domain knowledge. We first restrict our focus to one-step ahead forecasts. We consider the following variables

MF_{t+1}:	model-based forecast (made from origin t)
EF_{t+1}:	expert forecast (made from origin t)
S_{t+1}:	realization at time $t+1$

The variable S denotes sales here.

The model-based forecast is a linear function of past sales, where the weights are updated each month. We thus consider a recursive forecasting scheme. We can write

$$(1) \qquad MF_{t+1} = \mu_1 + \rho_1 S_t + \rho_2 S_{t-1} + \rho_3 S_{t-2} + ...$$

The recursive scheme means that the parameters are estimated for R in-sample data, and then a one-step-ahead forecast is made. Next, the sample is enlarged to R+1, parameters are re-estimated and again a one-step-ahead forecast is made. The number of forecasts thus obtained is denoted as P.

The expert receives the statistical model-based forecasts and quite often makes an adjustment. Franses and Legerstee (2007) show that part of that adjustment is based on past sales (again) and on other domain-specific variables, say X_t. Note that the inclusion of past sales in expert adjustment implies that there is some form of double-counting as the recursive scheme for the model-based forecasts already allows for additional impact of exceptional past sales data. In sum, the expert-adjusted forecast can be written as

$$(2) \qquad EF_{t+1} = \mu_2 + \delta_1 S_t + \delta_2 S_{t-1} + \delta_3 S_{t-2} + ... + \beta X_t + ...$$

As the expert did not write down how he or she modified the forecast, we do not have any specific information on X_t. Comparing (2) with (1) we see that the forecasting scheme of the expert nests the forecasting scheme of the model. If we were to call (2) the model used by the expert, then model (2) nests model (1). This observation is quite important as it has strong implications for the statistical methodology to be used below for testing whether (2) is better than (1) in terms of forecast accuracy.

There are two ways of defining the added contribution of the expert to the model-based forecast. The first is simply EF_{t+1} - MF_{t+1}, which assumes that both forecasts are independent and that the expert takes the model-based forecasts as given and adds his or her expertise. The second is by computing

$$(3) \qquad AD_{t+1} = EF_{t+1} - \lambda MF_{t+1},$$

where λ gets estimated from a linear regression, as it is recommended in Blattberg and Hoch (1990). In what follows below, we will use this second definition.

Does the expert's added value matter?

The first question that needs to be answered when comparing expert-adjusted and model-based forecasts is whether the added value of the expert actually matters (and this can be either in a positive or a negative way). This question can be answered by considering the following auxiliary test regression

$$(4) \qquad S_{t+1} = \alpha + \beta MF_{t+1} + \gamma AD_{t+1} + u_{t+1}$$

When the expert adds something that is relevant, the contribution of AD_{t+1} in (4) should be non-zero. Hence, a first relevant test is whether $\gamma = 0$ in (4).

A second question that is of interest is whether the contribution of the model and that of the expert are in balance, that is, does the 50% model – 50% manager rule (as advocated in Blattberg and Hoch, 1990) hold? It should be stressed here that there can be some discussion as to whether this 50%-50% rule concerns the forecast of the manager EF_{t+1} or the added value AD_{t+1} of the manager? We believe it should be the added value, and this then makes (4) the proper equation to consider. Indeed, if one were to consider the test regression

(5) $$S_{t+1} = \alpha + \beta MF_{t+1} + \theta EF_{t+1} + u_{t+1},$$

one should be aware that this means

(6) $$S_{t+1} = \alpha + (\beta + \theta\lambda) MF_{t+1} + \theta AD_{t+1} + u_{t+1}$$

In other words, if one were to find a 50%-50% balance between EF_{t+1} and MF_{t+1} using model (5), then this would not be informative for the balance between AD_{t+1} and MF_{t+1}, as this depends on the value of λ that has to be estimated too. So, we prefer to directly look at (4) and, based on this test regression, we examine the second interesting hypothesis that $\beta = \gamma$. In sum, we first consider the 50%-50% rule for the model versus added value of the expert. In Section 4 we will also consider another rule.

Is the RMSPE of the expert significantly lower than that of the model?

To test the null hypothesis that the root mean squared prediction error of the expert is equal to that of the model against the alternative hypothesis that the expert is better, we need to take account of the fact that model (2) nests model (1). As is convincingly explained in Clark and McCracken (2001), due to this nesting property the relevant test statistic does not have a standard normal distribution.

We follow the recommendation in Clark and McCracken (2001) and we will use the following procedure. We have R in-sample data, where in our cases below R concerns 5 years of monthly data, so R = 60. We have P recursively created out-of-sample forecasts, with P = 25. Hence, the fraction of forecasts over in-sample data is

$$\pi = \frac{P}{R} = 0.4$$

We need this value of π for the non-standard critical values of the upcoming test.

As said, we assume that model (2) nests model (1). We do not know how many variables are included in the additional set of regressors X_t, but for convenience we set that number k_2 equal to 2. Based on simulations concerning empirical size and power, Clark and McCracken recommend using the so-called ENC-NEW test, defined by

(7) $$ENC - NEW = P \frac{\frac{1}{P}\sum (u_{1,t+1}^2 - u_{1,t+1}u_{2,t+1})}{\frac{1}{P}\sum u_{2,t+1}^2}$$

The summation runs for the P one-step-ahead forecasts, and $u_{1,t+1}$ denotes the forecast errors for model-based forecasts (scheme (1)), and $u_{2,t+1}$ concerns the expert forecasts (scheme (2), which nests scheme (1)). The 5% critical values are given in Table 1 of Clark-McCracken (2001, page 92). For π = 0.4 and k_2 = 2 it is 1.481. Note that this test is a one-sided test of the null hypothesis that model (2) is equally good as model (1) against the alternative hypothesis that model (2) is better. So, the outcome of the test is whether the expert yields better forecasts or not.

2.2 Multi-step-ahead forecasts

When there is an interest in examining whether experts do better than models when it comes to *h*-step-ahead forecasts, we consider the variables

MF_{t+h}: model-based forecast (made from origin *t*)
EF_{t+h}: expert forecast (made from origin *t*)
$S_{t+h:}$ realization at *t+h*

Based on discussions with the relevant managers of our data provider, we will focus on the case where *h* = 6 in our empirical work.

As in the case of one-step-ahead forecasts we consider the test regression

(8) $$S_{t+h} = \alpha + \beta MF_{t+h} + \gamma AD_{t+h} + u_{t+h},$$

and we test the hypotheses that γ = 0 and that β = γ.

We can also compute the test statistic as in (7), but now a complication arises in terms of the asymptotic distribution of that test. For multi-step-ahead forecasts it is well known that the forecast errors are correlated, and this correlation needs to be included in the distribution. Clark and McCracken (2005) outline in detail how to do this in case the variables in X_t are known. One can then use bootstrap techniques to compute critical values for each particular situation at hand. In our case we face the problem that these variables, which are the additional variables used by the expert, are unknown. Fortunately, Clark and McCracken (2005, page 390) note that standard normal critical values would lead to reliable inference, provided that the forecast horizon is relatively short and π is also rather small. Our empirical work below seems to meet these requirements, so we will compute

(9) $$ENC-NEW_h = P\frac{\frac{1}{P}\sum(u_{1,t+h}^2 - u_{1,t+h}u_{2,t+h})}{\frac{1}{P}\sum u_{2,t+h}^2}$$

and we consider a one-sided test with a 5% critical value equal to 1.645.

In the next section we will use the methodology outlined in this section to see if experts' added expertise matters, whether it leads to better forecasts, and whether perhaps a little less added expertise would even do better.

3. Empirical results

We have data concerning products *i* within category *j* for country *c*. The data concern monthly sales for October 2004 to October 2006 of pharmaceutical products. The headquarters' office creates model-based forecasts and sends these to the local managers in each of the countries. Each country has its own expert dealing with all forecasts, so within a country we expect the same skills across products and categories. An expert is allowed to modify the model-based forecasts in a way he or she sees fit. We consider 35 countries and there are 7 product categories. We do not have all data for all categories for all countries. Also, within a category there are different numbers of products. In Table 1 we give a summary table of the amount of products within each category for each of the countries.

Insert Table 1 here

3.1. Preliminaries

When we run the regressions as in (4) for our one-step-ahead forecasts, we observe that the error term obeys AR(2)-type dynamics. Hence, we assume that the error term in (4) becomes

$$(1-\rho_1 L-\rho_2 L^2)u_{t+1}=\varepsilon_{t+1}$$

where L is the familiar lag operator.

We have access to a maximum of 25 monthly observations for each product, and when we allow AR(2) dynamics the effective sample reduces to 23 observations. This is quite small, and it seriously reduces the power of the tests on the parameters of interest. A simple solution is to pool the estimates for (4) across the products within a category. That is, we look at (3) per product per category, but we modify (4) into

(10) $$S_{1,t+1}=\alpha_1+\beta MF_{1,t+1}+\gamma AD_{1,t+1}+u_{1,t+1}$$

$$S_{2,t+1}=\alpha_2+\beta MF_{2,t+1}+\gamma AD_{2,t+1}+u_{2,t+1}$$

....

$$S_{n,t+1}=\alpha_n+\beta MF_{n,t+1}+\gamma AD_{n,t+1}+u_{n,t+1},$$

where n denotes the number of products within a category. The α parameter differs per product, but the β and γ parameters are common across products within a country-category combination. The model in (10) assumes independent equations with crossequation parameter restrictions. For the sake of computational simplicity, we assume the errors as independent.

Finally, the ENC-NEW and ENC-NEW$_h$ tests are computed for each product in each category for each country.

3.2 One-step-ahead forecasts

We first consider the 194 country-category combination for the one-step-ahead forecasts. The results appear in Tables 2 and 3, and a summary on the parameters in (4) appears in Table 4.

Insert Tables 2, 3 and 4 here

From Table 2 we learn that there are 104 of the 194 cases with P-values for $\gamma = 0$ that are smaller than 0.05. This concerns 53.6% o the cases. Next, from table 3, we can see that there are 118 of the 194 cases with P-values for $\beta = \gamma$ larger than 0.05, which amounts to 60.8% of the cases. Combining the results in Tables 2 and 3, we see that in 53 of 104 cases where the P-value for $\gamma = 0$ is smaller than 0.05 we find that the P-value for $\beta = \gamma$ is larger than 0.05. In words, when the expert significantly adds value, it holds in more than half of the cases that it obeys the 50%-50% rule. In Table 4 we give a summary of the estimated values of β, γ and β-γ, and we can see that β and γ are (on average) estimated to be about equally large and β-γ is estimated close to zero (on average).

Now we turn to a more detailed discussion of Tables 2 and 3. From Table 2 we see that the percentages with P-values for $\gamma = 0$ smaller than 0.05 across the seven categories are 67, 61, 42, 59, 55, 43 and 22, respectively. This means that for categories A, B, D and E the contribution of the experts is most prominent, while apparently for the categories C, F and G it is more difficult to add substantial expert knowledge.

When we look at the countries in Table 2, we notice that for most countries the experts sometimes have a relevant contribution across some (but not all) of the categories. Notable exceptions are IX and XIII where the expert never adds anything significant in a systematic way and XVII, XIX, XX and XXIX where the expert always seems to contribute additional to the model. Note that this does not necessarily mean that this contribution is positive, as γ can also be significantly negative.

From Table 3 we see that the percentages with P-values for $\beta = \gamma$ larger than 0.05 across the seven categories are 39, 61, 69, 56, 52, 80, and 100, which means that for the category A (with also the largest amount of products, see Table 1) the added contribution of the expert is not 50-50 with the model, while for the other categories this is more often the case.

When we look at the countries in Table 3, we notice that for most countries the experts and the models sometimes have a 50-50 contribution. Notable exceptions are XXII and XXVI where this balance always happens. So, again and in line with Table 2, strong and obvious differences across categories and countries do not seem to exist.

Insert Table 5 here

In Table 5 we report on the differences between RMSPE of the expert-adjusted forecasts versus the model-based forecast, when averaged over all products within a country-category combination. From this table we see that an average positive difference between expert and model (so the expert does better) across the seven categories occurs in 45, 45, 38, 56, 42, 37 and 44 per cent of the cases. This means that there seems to be no category that concerns a much better contribution of the expert. In other words, no category seems to be easier for an expert.

When we look at the countries in Table 5, we notice that for most countries the experts are sometimes better and sometimes worse than the models. Notable exceptions are XII and XXVII where the expert on average is always better and XXI and XXXII where the expert is always worse.

In general we can conclude from Tables 2, 3 and 5 that there are no systematic patterns across categories and countries, so we are tempted to summarize our results across all of these 194 combinations.

Table 5 did not yet consider whether differences in RMSPE were significant, so that is what we will do now. We therefore compute the median % improvement in RMSPE if the ENC-NEW test is significant and positive for the products in the category. This turns out to be 14.38%, with a minimum value of 0.05% and a maximum value of 84.92%. When this test is not significant, we get a median improvement (or better: deterioration) in RMSPE of -13.81%, with a minimum of -

324.4% and a maximum of 0.00%. These results suggest that if the expert is significantly better, the improvement is about equally large as in cases where the expert is not significantly better. Moreover, when the expert is not significantly better, the added contribution can be very bad with large negative outliers.

3.3 Six-step ahead forecasts

We now turn to the results for six-step-ahead forecasts. The main results on the tests for β and γ appear in Tables 6 to 8.

Insert Tables 6, 7 and 8 here

From Table 6 we learn that there are 89 of the 189 cases with P-values for $\gamma = 0$ smaller than 0.05, which is 47.1% of the cases. This is slightly smaller than the fraction of such cases for the one-step-ahead horizon. Table 7 indicates that there are 113 of the 189 cases with P-values for $\beta = \gamma$ larger than 0.05, that is 59.8% of the cases. Finally, in 39 of the 89 cases where the P-value for $\gamma = 0$ is smaller than 0.05, we find that the P-value for $\beta = \gamma$ larger than 0.05, that is 43.8%.

When we look at the summary in Table 8, we see that on average β seems to be larger than γ. However, if we only consider the cases when $\gamma = 0$ is rejected (last row of Table 8), we observe that β-γ is negative on average, and more so than in the case of one-step-ahead forecasts. This suggests that γ is then larger than β, meaning that for six-step-ahead forecasts the expert adds more to the model, on average, than in the case of one-step-ahead forecasts. This finding is of course not a surprise.

Overall, we see that the results for six-step-ahead forecasts are roughly the same as those for one-step-ahead forecasts, although now there is little less evidence for the 50%-50% rule. There is a small tendency towards more input of the expert.

A closer look at Table 6 indicates that the percentages with P-values for $\gamma = 0$ smaller than 0.05 across the seven categories are 52, 52, 48, 53, 52, 34 and 0, which means that for all but two categories the contribution of the experts is relevant in about half of the cases. For F and G it seems most difficult to add substantial expert knowledge.

When we look at the countries in Table 6, we notice that for most countries the experts exercise some contribution across some of the categories. Notable exceptions are VI, IX and XX where the expert never adds anything significant and XVII where the expert always seems to contribute additional to the model.

From Table 7 we see that the percentages with P-values for $\beta = \gamma$ larger than 0.05 across the seven categories are 61, 52, 60, 53, 64, 66, and 83, which means that for all categories the added contribution of the expert most often is in line with the 50%-50% rule.

When we look at the countries in Table 7, we notice that for most countries the experts and the models not always have a 50-50 contribution. Notable exceptions are VI, XX and XXII where this balance always happens, and XIX where this never happens. So, again, strong and obvious differences across categories and countries do not seem to exist.

Insert Table 9 here

Finally, we turn to differences in RMSPE of experts versus model. From Table 9 we see that a positive difference between expert and model (so the expert does better) across the seven categories occurs in 52, 58, 28, 47, 36, 31 and 29 per cent of the cases. This means that there seems to be no category that involves a much better contribution of the expert. Hence, again no category is easier for an expert. Comparing the results across Tables 5 and 9, we see 44% and 42% positive signs, respectively, suggesting that the forecast horizon does not matter much for the added value of the expert.

When we look at the countries in Table 9, we notice that for most countries the experts are sometimes better and sometimes worse than the models. Notable exceptions are I and XII where the expert is most frequently better and VI, XXII and XXXII where the expert is always worse, at least on average.

To see if the differences are significant, we resort to the ENC-NEW$_h$ test. We compute are the median improvement in RMSPE if the ENC-NEW$_h$ test is significant and positive for the products within the country-category combination. This turns out to be 16.84%, with a minimum value of 0.88% and a maximum value of 91.17%. When this test is not significant, we get a median improvement (deterioration) in RMSPE of -18.93, with a minimum of -458.8% and a maximum of 3.72%. This again confirms that the contribution of the manager in terms of forecast quality is not large. Also, when improvement is negative, it is very much skewed to the left, even more so than in case of one-step-ahead forecasts. In sum, the added value of the expert might be a little larger here, but then again, when the expert forecasts are worse, they are seriously bad.

4. Does the expert put too much weight on the own contribution?

From the results in the previous section we could learn that often the experts' added expertise and the model were in a 50%-50% balance. In this section we examine whether changes in this rule would lead to better results.

But first, we zoom in on some more detailed outcomes to sketch the issue. In Tables 10 and 11 we give the country-category combinations where the contribution of the expert is significant (that is γ is not equal to 0) and where the average improvement (across all products) is larger in an absolute sense than the average deterioration for one-step-ahead and six-steps-ahead forecasts, respectively. So, these cases concern cases where the expert does best.

Insert Tables 10 and 11 here

For the one-step-ahead forecasts in Table 10, there are 40 such cases (out of the 104 with significant γ parameters). The mean value of the estimated β and γ for these cases are 0.24 and 0.49. This suggests that when the expert forecast does better, it really is due to the added value of the expert. For the six-steps-ahead forecasts, there are 40 (out of 89) such cases, and there the average estimated β and γ parameters are 0.35 and 0.67. So, here the same conclusion can be drawn, that is, when it is better, it is due to the experts' added value, which then is about twice as large as the contribution of the model. In brief, when the expert does better, it is not a matter of luck, it is a systematic feature.

However, when we have a look at the cases where the contribution of the expert is significant but where the improvement is smaller than the (absolute) deterioration (104-40= 64 and 89-40=49 cases, respectively) we see another and perhaps more disturbing pattern. For the one-step-ahead forecasts the estimated β and γ are then 0.47 and 0.39, on average, while those for the six-step-ahead forecasts are

0.38 and 0.37. This suggests that when experts' forecasts are not as good as model-based forecasts, the added value of the expert is 50%.

Let us return to the test regression in (4). First, our unreported computations show that the average value of λ in the auxiliary regression (3) is about 0.4. When β = γ in (4), this means for (5) that the expert-model contribution is as 1.0 to 0.6, or when scaled to unity: 0.625 versus 0.375. So, the experts bring in quite some weight. To explain, look again at equation (6), that is

$$S_{t+1} = \alpha + (\beta + \theta\lambda)MF_{t+1} + \theta AD_{t+1} + u_{t+1}$$

When in this equation the weights of $\beta+\theta\lambda$ and θ are equal (as in the 50%-50% rule) and when $\lambda = 0.4$ (as we find on average), then we have that $\beta = 0.6\theta$. In other words, in

$$S_{t+1} = \alpha + \beta MF_{t+1} + \theta EF_{t+1} + u_{t+1}$$

the expert's forecast has weight θ = 0.625 versus the model forecast with a weight β is 0.375. Hence, across all our cases (countries, categories) we see that the experts add substantial value, on top of the model-based forecast, and our results in Tables 5 and 9 suggest that this added value might perhaps be too much.

To see if the added value of the expert can be improved, we then create new combined forecasts

(11) $$CF_{t+1} = \delta MF_{t+1} + (1-\delta)EF_{t+1}$$

with weight δ is 0.5. Notice that this means that the expert (added value) versus model contribution, given λ = 0.4 becomes 0.42 (experts' added value) versus 0.58 (model), approximately.

For the one-step-ahead forecasts, the results can be summarized as follows. The mean value of the numbers in Table 4 is -10.18%, with a minimum of -198.2% and a maximum value of 47.73%. So, on average the added value of an expert is not useful. Note that this is of course due large negative outliers, here very bad forecasts. In case we were to apply the 50-50 rule as in (11), this mean value becomes 5.47%, and the minimum and maximum values become -53.39% and 53.92%, respectively. In case the expert forecast is worse (on average, and in 108 of the 194 cases of Table 5), the difference between this 50-50 rule and the used rule would have a mean value of even 25.82 (with minimum 1.59% and maximum 144.8%), while for the case the expert forecasts were already better (194-108=86 cases) such improvement would have a mean value of only 2.89% (minimum of -20.04% and maximum of 16.47%).

This leads to the important conclusion that the 50-50 rule in (11) leads to much better forecasts overall and in particular in cases where the initial expert forecast was not very good. In sum, at present the experts exercise too much weight on the final forecasts. Would they impose less weight (downplaying it from 0.63 - 0.37 to 0.50 - 0.50), then overall forecast quality would seriously improve.

Similar results are obtained for the six-step-ahead forecasts. In case the expert forecast is worse (on average, and in 110 of the 190 cases of Table 9), the difference between this 50-50 rule and the used rule would have a mean value of even 28.35 (with minimum 2.25% and maximum 301.84%), while for the case the final expert forecasts were already better (190-110=80 cases) such improvement would have a mean value of only 1.34% (minimum of -22.79% and maximum of 15.56%).

In sum, yes, the experts seem to exercise too much weight in their final forecasts. Giving more credit to the model would lead to serious improvement.

5. Conclusions

In this paper we have put forward an effective and reliable methodology that allows us to investigate if experts' added knowledge to model-based forecasts is relevant and whether it leads to more accurate forecasts. The methodology builds on the latest developments in testing for equal forecast accuracy. We applied our methodology to a huge database concerning monthly sales of pharmaceutical products in various categories and various countries. Due to the fact that we have such a large database, we feel confident to draw some generalizing conclusions.

The first main conclusion that we can draw from our extensive analysis is that experts' <u>added value</u> frequently matters and that when it matters it also frequently occurs on a 50-50 basis. Note that this means that in combining model and expert the relative weights are 0.375 versus 0.625, and hence the impact of the expert is large.

This conclusion holds across all countries, categories and even across the two horizons, although the added value of the expert for longer term horizons is even slightly larger. Hence, even though there are individual differences in countries and categories which sometimes can be quite large, on average the added value of the expert versus the model is 50-50.

The second main conclusion is that when the expert yields a significant positive contribution to forecast quality, the final forecast's improvement is about equally large as the deterioration in case the expert does not significantly outperform the model. So, the 50%-50% rule, as apparently is used, does not yield substantial improvement.

The third main conclusion thus seems to be that experts put too much emphasis on their own added contribution. Indeed, when we give the added value of the expert less weight, we see strong improvement in final forecast quality.

It seems that our findings point towards one and the same major feature of expert adjustment and that is that experts put too much emphasis on their own judgement and too little on the model. It is perhaps misunderstood that the model captures recent events via the updated estimates of the parameters, or perhaps out-of-the-model events receive too much for too long a period. One implication of our findings is that the way the model works should be better communicated to the experts. A second is that experts should start documenting what they effectively do when they adjust model-based forecasts. A third, and which perhaps leads to a first immediate action, is that experts should become aware of the notion that they put too much weight on their expertise. When it is useful, it is no problem, but when it is not, forecasts can become dramatically bad. A case study of these exceptional forecasts may yield a first set of insights.

Table 1: Number of products in the categories for each country

	Category						
Country	A	B	C	D	E	F	G
I	3	2	1	4	2		
II	3	4		2	7	1	
III	3	5	6	10	8	4	1
IV	2	1	3	4	2	2	
V	11	6	7	9	5	4	1
VI	2			2			
VII	7	4		1	6	1	1
VIII	9	4	2	6	6	6	
IX	2	1		1		2	
X		1	1		6	3	
XI	10	4		8	7	4	
XII	7	9	4	7	8	1*	
XIII		2	2	2	3	3	
XIV	12	10	2	9	8	5	
XV	23	3		6	18	4	1
XVI	32	20	1	16	10	5	1*
XVII	7	2	4	2	11	3	
XVIII	12	4	2	5	8	4	
XIX	10	5	3	5	8	3	
XX	1		1	2	6		
XXI	6	1	2	4	9	5	
XXII	1		1	2	6		
XXIII	9	5	15	10	12	4	1
XXIV	6	9	2	3	6	1	
XXV	3	3	2*	1	2	2	
XXVI	6	4	2	6	3	4	
XXVII	11	3	7	4	8	3	
XXVIII	7	2		5	3	2	
XXIX	12	7	4	6	10	2	
XXX	15	7	6	8	7	4	1
XXXI	15	12	3	11	9	5	1
XXXII	1				3		
XXXIII	8	8	13	15	12	5	1*
XXXIV	7	8	2		6	2	
XXXV	2	5		2	7	3	

* These cases are only available for the one-step-ahead forecasts but not for the six-step-ahead forecasts. So, the one-step-ahead forecasts concern 194 country-category combinations, while for the six-step-ahead forecasts there are 190 such cases.

Table 2: P-value of the test for $\gamma = 0$ in the pooled test regression (10).
The case of one-step-ahead forecasts

	Category						
Country	A	B	C	D	E	F	G
I	0.017	0.000	0.956	0.257	0.003		
II	0.575	0.131		0.000	0.001	0.576	
III	0.000	0.000	0.000	0.060	0.000	0.095	0.003
IV	0.000	0.778	0.017	0.463	0.041	0.382	
V	0.364	0.036	0.540	0.051	0.543	0.000	0.719
VI	0.010			0.875			
VII	0.000	0.347		0.019	0.531	0.046	0.435
VIII	0.000	0.345	0.914	0.770	0.090	0.823	
IX	0.293	0.361		0.095		0.171	
X		0.012	0.001		0.309	0.468	
XI	0.000	0.000		0.000	0.000	0.162	
XII	0.000	0.000	0.022	0.631	0.108	0.432	
XIII		0.157	0.431	0.274	0.609	0.395	
XIV	0.892	0.000	0.715	0.084	0.000	0.202	
XV	0.000	0.887		0.010	0.000	0.562	0.905
XVI	0.000	0.000	0.080	0.000	0.357	0.000	0.265
XVII	0.000	0.000	0.003	0.000	0.000	0.004	
XVIII	0.000	0.598	0.671	0.000	0.003	0.412	
XIX	0.000	0.000	0.000	0.000	0.000	0.041	
XX	0.031		0.003	0.024	0.001		
XXI	0.000	0.462	0.976	0.000	0.013	0.018	
XXII	0.568		0.256	0.026	0.008		
XXIII	0.500	0.178	0.009	0.011	0.000	0.390	0.127
XXIV	0.775	0.000	0.891	0.416	0.291	0.963	
XXV	0.000	0.725	0.002	0.417	0.705	0.085	
XXVI	0.158	0.641	0.593	0.020	0.086	0.006	
XXVII	0.012	0.001	0.078	0.475	0.321	0.001	
XXVIII	0.063	0.045		0.000	0.699	0.006	
XXIX	0.000	0.006	0.004	0.000	0.000	0.003	
XXX	0.000	0.000	0.048	0.000	0.000	0.006	0.558
XXXI	0.002	0.000	0.747	0.002	0.150	0.071	0.018
XXXII	0.008				0.052		
XXXIII	0.641	0.006	0.548	0.000	0.000	0.109	0.069
XXXIV	0.000	0.000	0.547		0.000	0.000	
XXXV	0.744	0.000		0.000	0.397	0.031	

Table 3: P-value of the test for $\beta = \gamma$ in the pooled test regression (10).
The case of <u>one-step-ahead forecasts</u>

	Category						
Country	A	B	C	D	E	F	G
I	0.122	0.820	0.761	0.193	0.015		
II	0.078	0.182		0.154	0.000	0.565	
III	0.000	0.052	0.127	0.370	0.702	0.484	0.267
IV	0.042	0.488	0.048	0.000	0.000	0.035	
V	0.001	0.909	0.000	0.953	1.000	0.465	0.459
VI	0.000			0.687			
VII	0.000	0.031		0.045	0.034	0.501	0.989
VIII	0.000	0.250	0.070	0.071	0.035	0.178	
IX	0.213	0.011		0.000		0.721	
X		0.630	0.002		0.030	0.056	
XI	0.007	0.000		0.632	0.000	0.756	
XII	0.049	0.960	0.314	0.944	0.016	0.252	
XIII		0.707	0.056	0.004	0.971	0.433	
XIV	0.000	0.009	0.682	0.000	0.314	0.593	
XV	0.000	0.000		0.636	0.141	0.449	0.068
XVI	0.000	0.000	0.540	0.000	0.878	0.010	0.479
XVII	0.983	0.247	0.860	0.015	0.000	0.001	
XVIII	0.000	0.809	0.523	0.152	0.272	0.944	
XIX	0.000	0.034	0.095	0.017	0.072	0.000	
XX	0.012		0.000	0.329	0.002		
XXI	0.000	0.194	0.525	0.630	0.725	0.602	
XXII	0.515		0.525	0.869	0.417		
XXIII	0.501	0.003	0.477	0.776	0.179	0.392	0.703
XXIV	0.000	0.000	0.896	0.004	0.169	0.070	
XXV	0.442	0.312	0.000	0.539	0.014	0.546	
XXVI	0.065	0.905	0.827	0.127	0.160	0.264	
XXVII	0.000	0.640	0.295	0.273	0.030	0.000	
XXVIII	0.876	0.245		0.000	0.704	0.089	
XXIX	0.000	0.788	0.000	0.000	0.611	0.608	
XXX	0.145	0.001	0.388	0.501	0.007	0.000	0.512
XXXI	0.451	0.273	0.612	0.000	0.766	0.662	0.388
XXXII	0.672				0.000		
XXXIII	0.001	0.095	0.000	0.000	0.002	0.574	0.260
XXXIV	0.000	0.000	0.032		0.077	0.067	
XXXV	0.254	0.000		0.000	0.001	0.184	

Table 4: Summary of estimation results.
The case of <u>one-step-ahead forecasts</u> (194 observations)

			Statistics		
Estimates	Mean	Median	Maximum	Minimum	Stand.dev.
β	0.271	0.343	1.479	-2.173	0.537
γ	0.247	0.274	1.992	-4.843	0.594
$\beta - \gamma$	0.024	0.067	5.178	-3.327	0.734
$\beta - \gamma$ (two outliers deleted)	0.001	0.067	1.704	-2.147	0.589
$\beta - \gamma$ (if P-value of $\gamma = 0$ is < 0.05, 104 cases)	-0.060	0.039	1.704	-3.327	0.684

Table 5: Averaged difference in RMSPE of the expert forecast versus the model-based forecast. The case of <u>one-step-ahead forecasts</u>
(averaged over products in categories)
A positive number means that the expert is better.

	Category						
Country	A	B	C	D	E	F	G
I	4.43	8.26	-9.14	6.69	2.42		
II	-16.74	11.00		24.44	-3.57	-32.35	
III	-1.97	-2.36	-8.81	3.52	-32.56	-9.39	10.71
IV	5.55	14.76	24.17	-20.28	-10.60	8.66	
V	6.10	4.61	3.83	-11.30	-98.25	-21.95	31.00
VI	-27.61			-18.95			
VII	-59.82	-34.61		34.73	0.08	7.13	1.43
VIII	-39.42	-24.58	-97.05	-10.00	1.57	-51.50	
IX	-23.78	-5.81		7.59		-14.95	
X		-6.43	-2.40		13.59	-29.49	
XI	7.49	-31.27		-14.91	-10.16	-4.55	
XII	0.18	21.20	1.04	9.95	0.50	30.25	
XIII		-20.53	-0.45	4.18	-61.37	8.07	
XIV	-98.79	-113.0	1.52	-28.66	9.20	6.80	
XV	-22.61	-43.26		10.15	8.30	-11.63	-31.49
XVI	12.75	13.19	0.84	-5.82	1.74	0.61	37.82
XVII	-5.27	16.88	-7.71	31.77	-51.39	-39.31	
XVIII	14.05	-2.29	-1.96	10.65	4.13	-5.65	
XIX	12.94	-48.84	-2.84	4.93	-14.18	-82.60	
XX	8.95		-63.28	25.98	-6.09		
XXI	-1.42	-6.76	-30.89	-2.71	-9.32	-4.14	
XXII	-79.62		-198.2	28.77	-83.69		
XXIII	29.38	-1.47	4.16	0.91	-5.08	-1.38	-0.77
XXIV	-23.26	12.95	-4.64	1.53	2.39	-7.81	
XXV	-27.79	-15.39	-106.4	-13.79	-20.67	-87.70	
XXVI	8.79	10.17	-3.83	10.40	11.94	-4.71	
XXVII	6.62	22.14	12.62	1.66	4.17	47.73	
XXVIII	1.11	5.72		-36.92	0.40	1.21	
XXIX	-2.29	6.51	-8.77	-1.56	-5.30	7.95	
XXX	-8.13	-12.48	15.08	-25.41	-2.39	-8.74	-1.99
XXXI	-2.21	-0.87	1.49	-21.84	-3.25	6.69	-1.09
XXXII	-101.4				-55.74		
XXXIII	1.82	14.08	6.09	-7.42	22.29	-10.07	4.91
XXXIV	1.57	38.40	-25.07		-12.01	13.24	
XXXV	-61.98	-44.84		8.51	-121.4	-18.25	

Table 6: P-value of the test for $\gamma = 0$ in the pooled test regression (10) with (8).
The case of six-step-ahead forecasts

	Category						
Country	A	B	C	D	E	F	G
I	0.459	0.720	0.299	0.003	0.000		
II	0.001	0.300		0.000	0.000	0.801	
III	0.550	0.000	0.002	0.206	0.000	0.815	0.792
IV	0.001	0.487	0.001	0.000	0.077	0.048	
V	0.669	0.811	0.000	0.174	0.242	0.405	0.609
VI	0.308			0.772			
VII	0.830	0.009		0.001	0.321	0.307	0.608
VIII	0.000	0.029	0.974	0.000	0.124	0.253	
IX	0.378	0.842		0.717		0.661	
X		0.035	0.028		0.016	0.072	
XI	0.000	0.000		0.185	0.000	0.010	
XII	0.000	0.000	0.041	0.782	0.621		
XIII		0.025	0.840	0.312	0.767	0.514	
XIV	0.000	0.022	0.929	0.000	0.001	0.010	
XV	0.830	0.437		0.001	0.000	0.520	0.785
XVI	0.000	0.000	0.935	0.223	0.083	0.786	
XVII	0.001	0.000	0.000	0.000	0.000	0.001	
XVIII	0.485	0.815	0.281	0.018	0.468	0.524	
XIX	0.000	0.000	0.001	0.000	0.000	0.838	
XX	0.501		0.974	0.759	0.926		
XXI	0.000	0.206	0.798	0.000	0.718	0.714	
XXII	0.833		0.086	0.306	0.001		
XXIII	0.001	0.105	0.000	0.496	0.715	0.281	0.550
XXIV	0.000	0.000	0.356	0.415	0.091	0.135	
XXV	0.050	0.192		0.003	0.378	0.077	
XXVI	0.024	0.232	0.956	0.014	0.003	0.272	
XXVII	0.010	0.410	0.014	0.500	0.064	0.936	
XXVIII	0.000	0.216		0.559	0.013	0.001	
XXIX	0.184	0.026	0.001	0.000	0.000	0.001	
XXX*	0.000	0.067	0.077	0.000	0.010	0.000	
XXXI	0.113	0.000	0.997	0.925	0.017	0.049	0.753
XXXII	0.005				0.118		
XXXIII	0.509	0.095	0.006	0.000	0.002	0.002	
XXXIV	0.204	0.035	0.001		0.585	0.822	
XXXV	0.277	0.000		0.000	0.010	0.041	

* For XXX, G, the model parameters could not be estimated due to lack of data.

Table 7: P-value of the test for $\beta = \gamma$ in the pooled test regression (10) with (8).
The case of six-step-ahead forecasts

	Category						
Country	A	B	C	D	E	F	G
I	0.055	0.034	0.075	0.000	0.118		
II	0.860	0.174		0.182	0.007	0.223	
III	0.863	0.000	0.306	0.986	0.861	0.202	0.355
IV	0.007	0.997	0.001	0.000	0.990	0.002	
V	0.247	0.828	0.507	0.026	0.389	0.000	0.986
VI	0.686			0.673			
VII	0.085	0.001		0.000	0.003	0.017	0.615
VIII	0.774	0.078	0.403	0.053	0.024	0.006	
IX	0.344	0.120		0.779		0.468	
X		0.329	0.383		0.012	0.056	
XI	0.000	0.000		0.023	0.019	0.130	
XII	0.015	0.000	0.961	0.236	0.064		
XIII		0.050	0.850	0.917	0.905	0.540	
XIV	0.000	0.543	0.571	0.121	0.027	0.855	
XV	0.000	0.002		0.184	0.009	0.250	0.011
XVI	0.337	0.000	0.262	0.000	0.606	0.117	
XVII	0.000	0.021	0.000	0.975	0.032	0.001	
XVIII	0.996	0.275	0.038	0.217	0.145	0.240	
XIX	0.000	0.027	0.000	0.000	0.011	0.000	
XX	0.464		0.888	0.636	0.895		
XXI	0.000	0.050	0.532	0.170	0.824	0.293	
XXII	0.222		0.216	0.148	0.074		
XXIII	0.003	0.542	0.000	0.001	0.468	0.419	0.862
XXIV	0.136	0.012	0.053	0.103	0.664	0.010	
XXV	0.034	0.002		0.147	0.151	0.197	
XXVI	0.715	0.492	0.138	0.048	0.298	0.378	
XXVII	0.065	0.652	0.012	0.196	0.454	0.929	
XXVIII	0.160	0.475		0.007	0.469	0.007	
XXIX	0.155	0.789	0.000	0.000	0.884	0.429	
XXX*	0.000	0.909	0.019	0.025	0.089	0.000	
XXXI	0.088	0.040	0.884	0.000	0.000	0.242	0.233
XXXII	0.273				0.004		
XXXIII	0.619	0.577	0.040	0.011	0.053	0.001	
XXXIV	0.000	0.049	0.030		0.597	0.055	
XXXV	0.283	0.969		0.000	0.017	0.214	

* For XXX, G, the model parameters could not be estimated due to lack of data

Table 8: Summary of estimation results.
The case of six-step-ahead forecasts (189 observations)

	Statistics				
Estimates	Mean	Median	Maximum	Minimum	Stand.dev.
β	0.351	0.231	8.457	-2.962	1.019
γ	0.263	0.192	4.428	-4.801	0.671
β - γ	0.089	-0.007	8.632	-6.800	1.205
β – γ (four outliers deleted)	0.023	-0.008	2.198	-3.298	0.723
β - γ (if P-value of γ = 0 is < 0.05, 89 cases)	-0.145	-0.236	5.377	-6.800	1.139

Table 9: Averaged difference in RMSPE of the expert forecast versus the model-based forecast. The case of six-step-ahead forecasts
(averaged over products in categories)
A positive number means that the expert is better.

	Category						
Country	A	B	C	D	E	F	G
I	0.26	4.09	-35.19	4.56	0.41		
II	18.47	0.49		24.79	-21.61	-35.41	
III	3.45	-2.22	-11.44	-0.02	-2.85	-94.41	-1.73
IV	5.78	33.57	33.27	-16.32	-3.85	24.78	
V	4.74	-1.60	5.63	-4.02	-39.56	-18.49	-21.99
VI	-40.83			-13.79			
VII	-35.07	-320.9		56.28	-11.61	-7.38	3.72
VIII	-60.37	-82.15	-180.4	-11.13	11.96	-46.15	
IX	-25.15	8.47		-6.24		-15.61	
X		-14.36	-22.34		19.25	-2.73	
XI	23.08	-8.87		-3.94	7.10	6.84	
XII	3.87	2.35	-0.99	11.13	4.46		
XIII		15.44	-4.64	3.51	-19.82	14.32	
XIV	-46.25	-30.99	-10.06	-4.94	7.60	11.96	
XV	-25.61	-106.3		4.28	13.41	2.79	-0.60
XVI	0.51	29.19	-13.55	4.44	-8.56	-9.04	
XVII	-24.62	31.50	17.13	31.09	-62.64	-52.08	
XVIII	12.10	-40.76	-0.56	4.50	-13.88	-5.30	
XIX	25.05	16.90	-11.34	22.69	-15.53	-70.42	
XX	26.29		-16.79	29.92	-41.32		
XXI	7.40	16.01	-55.41	9.12	-14.18	0.88	
XXII	-74.25		-458.8	-33.53	-94.62		
XXIII	12.89	8.67	8.77	-0.92	7.15	-3.24	-17.11
XXIV	9.26	6.25	-9.64	-5.83	4.89	-24.42	
XXV	-23.42	-1.81		7.25	-29.33	-45.87	
XXVI	-12.11	7.24	4.62	-1.30	18.51	-34.52	
XXVII	-8.12	15.84	3.33	-29.24	-11.83	53.69	
XXVIII	5.02	-1.02		-27.96	11.84	7.09	
XXIX	-8.52	-0.01	-2.41	4.69	-9.04	-12.91	
XXX	4.29	32.42	-0.01	-22.53	-6.98	2.07	7.63
XXXI	-59.14	1.74	8.24	-20.03	-13.27	-0.31	-39.20
XXXII	-45.83				-14.82		
XXXIII	10.18	5.87	-2.74	-44.50	11.64	-22.61	
XXXIV	-10.32	42.39	-25.04		-5.82	-22.61	
XXXV	-116.9	-59.05		25.13	-109.1	-23.42	

Table 10: Cases where improvement in RMSPE of expert over model when ENC-NEW test is significant AND LARGER IN ABSOLUTE SENSE THAN deterioration in RMSPE when the test is not significant (for significant γ)
The case of <u>one-step-ahead</u> forecasts

Country	Category A	B	C	D	E	F	G
I	X	X			X		
II							
III	X						
IV	X		X				
V							
VI							
VII				X		X	
VIII							
IX							
X					X		
XI	X						
XII		X	X				
XIII							
XIV							
XV				X	X		
XVI	X	X				X	
XVII		X		X			
XVIII	X			X	X		
XIX	X			X			
XX	X			X			
XXI							
XXII				X			
XXIII			X				
XXIV		X					
XXV							
XXVI				X		X	
XXVII					X	X	
XXVIII		X				X	
XXIX						X	
XXX							
XXXI							
XXXII							
XXXIII					X		
XXXIV		X				X	
XXXV				X			

Table 11: Cases where improvement in RMSPE of expert over model when ENC-NEW$_h$ test is significant AND LARGER IN ABSOLUTE SENSE THAN deterioration in RMSPE when the test is not significant (in case γ is significant)
The case of six-step-ahead forecasts

Country	Category						
	A	B	C	D	E	F	G
I				X	X		
II	X			X			
III							
IV	X		X			X	
V							
VI							
VII				X			
VIII							
IX							
X					X		
XI	X				X	X	
XII	X						
XIII	X						
XIV						X	
XV				X	X		
XVI		X					
XVII		X	X	X			
XVIII				X			
XIX	X	X		X			
XX							
XXI	X			X			
XXII					X		
XXIII			X				
XXIV	X	X					
XXV				X			
XXVI					X		
XXVII			X				
XXVIII	X				X	X	
XXIX							
XXX						X	
XXXI							
XXXII							
XXXIII							
XXXIV		X					
XXXV				X			

References

Blattberg, Robert C. and Stephen J. Hoch (1990), Database models and managerial intuition: 50% model + 50% manager, Management Science, 36, 887-899

Clark, Todd E. and Michael W. McCracken (2001), Tests of equal forecast accuracy and encompassing for nested models, Journal of Econometrics, 105, 85-110

Clark, Todd E. and Michael W. McCracken (2005), Evaluating direct multi-step forecasts, Econometric Reviews, 24, 369-404

Franses, Philip Hans and Rianne Legerstee (2007), Dynamics of expert adjustment to model-based forecasts, Econometric Institute Report 2007-35, Erasmus University Rotterdam

Mathews, Brian P. and Adamantios Diamantopoulos (1986), Managerial intervention in forecasting: An empirical investigation of forecast manipulation, International Journal of Research in Marketing, 3, 3-10

Mathews, Brian P. and Adamantios Diamantopoulos (1992), Judgmental revision of sales forecasts: The relative performance of judgmentally revised versus non-revised forecasts, Journal of Forecasting, 11, 569-576