Judgment and Decision Making, Vol. 13, No. 2, March 2018, pp. 202–211

Comparative evaluation of the forecast accuracy of analysis reports and a prediction market

Bradley J. Stastny* Paul E. Lehner†

Abstract

This paper summarizes an empirical comparison of the accuracy of forecasts included in analysis reports developed by professional intelligence analysts to comparable forecasts in a prediction market that has broad participation from across an intelligence community. To compare forecast accuracy, 99 event forecasts were extracted from qualitative descriptions found in 41 analysis reports and posted on the prediction market. Quantitative probabilities were imputed from the qualitative forecasts by asking seasoned professional analysts, who did not participate in the prediction market, to read the reports and to infer a quantitative probability based on what was written. These readers were also asked to provide their personal probabilities before and after reading the reports. There were two statistically significant results of particular interest. First, the primary result is that the prediction market forecasts were more accurate than the analysis reports. On average prediction market probabilities were 0.114 closer to ground truth than the analysis report probabilities. Second, in cases where analysts (readers) updated their personal probabilities in a direction opposite to what the reports implied, analysts tended to update their probabilities in the correct direction. This occurred even though, on average, reading the reports did not make readers more accurate.

Keywords: prediction market, forecasting, probability judgment, intelligence

1 Introduction

When forecasting events, it is common practice to express certainties with verbal phrases such as, "The probability is high that . . . ", "It is likely that . . . ", or "there is a fair chance . . . "(e.g., Gardner, 2010). This propensity to express probabilities in qualitative language is common in the intelligence community where analysts are responsible for developing forecasts concerning complex geopolitical events (Lehner, Michelson, Adelman & Goodman, 2012; Kent, 1964). This preference for qualitative forecast statements is contrary to the preference of many consumers of forecasts. They prefer numerical forecasts such as, "There is a 70% chance that. . . " (Erev & Cohen, 1990).

There are a variety of efforts underway among intelligence professionals to develop crowd wisdom methods and to integrate those methods with more traditional intelligence analysis (ACE, 2016). Among these efforts was the fielding of a prediction market within a group of intelligence agencies, where the prediction market has thousands of participants with several hundred who frequently and regularly participate. In this paper this market will be referred to as the intelligence community prediction market (ICPM). Like many prediction markets, the ICPM uses non-monetary points to "buy" or "sell" shares on questions of intelligence interest. The resulting market "price" is a quantitative estimate that serves as the market consensus prediction for each question. The impetus behind the ICPM was to create a functionality that would allow members to quickly collaborate and settle on a numerical consensus.

Although a number of studies have already compared the accuracy of crowd wisdom methods to expert judgements (e.g., Surowiecki, 2004), these comparisons typically required that the experts generate quantitative forecasts that were not their "natural" mode of communication. It could be argued that such an approach unfairly disadvantages experts in such comparisons.

This study endeavored to compare expert and crowd wisdom forecasts by comparing the analysis reports produced by expert analysts to crowd wisdom forecasts that were drawn from forecast events listed in those reports. The approach is based on the method described in Lehner et al. (2012) for evaluating the accuracy of imprecise qualitative forecasts. Based upon the idea that individuals can impute numeric

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) via contract 2009-0917826-016, and is subject to the Rights in Data – General Clause 52.227–14, Alt. IV (DEC2007).

Any views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright annotation therein.

Dr. Lehner is currently on an Interagency Personnel Agreement from MITRE to IARPA serving as the Chief Officer for Test and Evaluation. This work was performed before Dr. Lehner moved to IARPA.

A comment and reply to this paper have been published.

*Capital One

†Corresponding author. Intelligence Advanced Research Projects Activity. Email: paul.lehner@iarpa.gov.

https://doi.org/10.1017/S1930297500007105 Published online by Cambridge University Press

probabilities from qualitative language (Mosteller & Youtz, 1990), Lehner et al. asked analysts to read qualitative forecasts extracted from strategic intelligence assessments and to impute probabilities based upon those forecasts. For instance, consider this fictional example created to mimic the types of qualitative forecasts made in geopolitical intelligence assessments:

> (1) We assess with moderate confidence that Country X will be more at risk of widespread internal violence in the next year. We cannot rule out that Group 1 elements might seek to confront Group 2's militia. Such efforts by Group 1 could prompt a violent response from Group 2, leading to widespread fighting.

Based only on what is written in the report, multiple readers could be asked to assign a numeric probability to "Country X will have widespread internal violence in the next year." If reader imputed probabilities clustered tightly, then the mean of that cluster is a fair reflection of what is written in the report irrespective of what the authors may or may not have intended. If reader imputed probabilities varied widely, then that is evidence that the report did not in fact make a meaningful forecast.

In this study the approach in Lehner et al. was modified to support a direct empirical comparison of the accuracy of qualitative and quantitative forecast statements. Specifically, qualitative forecasts from analysis reports were compared to quantitative forecasts on the ICPM. This study examined 41 analysis reports covering diverse topics in more than 30 geopolitical regions. These reports were considered high quality products where each reflected contributions and consensus judgments of multiple analysts from multiple agencies, where these analysts were considered among the most expert in the topic of the report. Very broadly described, these analysis reports were composed of two elements: an in-depth assessment of the key causal drivers that influence the evolution of events and a collection of forecast statements that are warranted by the key driver assessment. Ninety-nine forecast statements were selected and then transcribed into forecast questions that could be posted on the ICPM. Statement (1) above is representative of the 99 statements that were selected.

Now, consider this statement:

> (2) Will there be a lethal confrontation between Group 1 and Group 2 before 1 January 2014? (Here a "lethal confrontation" is defined as a conflict that causes at least 100 combined deaths of Group 1 and Group 2 personnel within a one-week period, with at least 10 deaths occurring on both the Group 1 and Group 2 sides).

Questions such as this are typical of forecast questions found on the ICPM. They are narrow in scope and are often indicators of the analysis for a much wider topic.

Five experienced analysts were asked to impute quantitative probabilities in response to 99 such questions after reading the analysis report from which each question was derived. Though considered experienced analysts, none of the readers were as knowledgeable as the authors of the analysis reports. The forecast questions were also posted to the ICPM. After the forecast questions were resolved, the imputed and ICPM probabilities were examined for relative accuracy.

2 Method

A seven-step process was used to extract forecast questions from the analysis reports, obtain imputed probabilities from the readers of the reports and the ICPM, resolve questions based upon ground truth, and analyze the data. Two types of forecast questions were developed. There were 71 precisely-worded forecast question and 28 fuzzy questions. The steps used to develop both types of questions are described below.

Note that below we occasionally refer to a "product" rather than a report. This is because in the standard vernacular of the intelligence community a written analysis is usually referred to as an analytic product, so where we describe the instructions provided to participants we use the term that we used with them: product.

2.1 Analyze the analysis reports

Forecasts were accessed from a collection of analysis reports published between October 2012 and May 2014. Reports in this collection are widely considered to be high quality analyses. Each report contained a variety of qualitative forecast statements. Some forecast statements have more supporting background information than others and each statement was assessed for the potential to create resolvable forecast questions and for whether there was sufficient information for readers to impute probabilities. Consider again this fictional passage that resembles passages in actual reports:

> We assess with moderate confidence that Country X will be more at risk of widespread internal violence in the next year. We cannot rule out that Group 1 elements might seek to confront Group 2's militia. Such efforts by Group 1 could prompt a violent response from Group 2, leading to widespread fighting.

This passage asserts that the consensus of multiple analysts is "moderate confidence" that Country X would experience widespread internal violence. Further, the passage specifies that Group 1 might confront Group 2, leading to widespread fighting. Thus, this passage would be flagged as a promising passage within a report that may allow for the creation of a forecast question with a resolvable outcome.

https://doi.org/10.1017/S1930297500007105 Published online by Cambridge University Press

2.2 Develop forecast questions

The precisely-worded and fuzzy forecast questions were developed from flagged passages using language pulled directly from the reports.

For the precisely-worded questions, if the language was vague (e.g., the term "widespread violence" from the example in Step 1), then language was selected that was appropriate for the ICPM but related to the language of the original passage from a report. Consider this candidate question that was developed from the example passage in Step 1:

> Will there be a lethal confrontation between Group 1 and Group 2 before 1 January 2014?

This question captures the essence of the original passage in a way that also makes the question better suited for posting on a prediction market.

Fuzzy questions, were not precisely worded and were intended to more closely reflect language used in the reports. For example, from the passage in Step 1 the following would be a characteristic fuzzy question:

> Will elements of Group 1 seek to confront Group 2's militia?

Here hard to resolve phrases such as "elements of" and "seek to confront" are kept in the forecast question.

2.3 Create resolution language for the questions

For the precisely worded questions, resolution language was developed to ensure that questions would be "empirically resolvable". An empirically resolvable question is one that has well defined parameters that provide conditions for ground truth assessment. Here is an example of resolution language developed for the precisely worded question above:

> For positive resolution, a major news source must confirm that there has been a significant lethal confrontation between Group 1 and Group 2 members on or before 01 January 2014. "Before" should be interpreted to mean at or prior to the end (23:59:59 ET) of the previous day. For example, "before 10 Oct" means any time up to 23:59:59 ET on 9 Oct. A "lethal confrontation" is defined as one that causes at least 100 combined deaths of Group 1 and Group 2 personnel within a one-week period. "Group 1 member" refers to militia members considered to be associated with the Group 1 political party in Country X. Individuals who are Group 1 militia members do not need to be citizens of Country X. "Group 2 member" refers to militia members considered to be associated with the Group 2 political party in Country X. Individuals who are Group 2 militia members do not need to be citizens of Country X.

This resolution language clarifies potentially confusing language in the question, including determinations of group membership, the specifics of lethal attacks, and the timeframe in which the question should be judged.

The fuzzy questions did not have associated resolution language. Rather the ICPM administrators, at their discretion, determined whether there was a clear resolution. In general, because of their lack of specificity, fuzzy questions are more likely than precisely worded questions to be unresolved. However, all 28 fuzzy questions posted for this study were resolved.

The authors of this paper did not participate in the resolution of the fuzzy forecast questions.

2.4 Final review of the questions and resolution language

For each forecast question, the relevant passage (from which a specific question was developed) and the resolution language were submitted to the government organization that produced the report to review and possibly edit the question and resolution language. The government reviewers were independent assessors who were not ICPM participants or participants in the report analyses described in 1–3 above. The reviewers had broad policy and analysis experience. Government edits focused on the definitions of vague terms in the questions and the resolution language.

2.5 Data collection

Once the final versions of the questions and resolution language were settled, two separate data collection procedures were initiated in parallel.

ICPM Data Collection: ICPM administrators posted the questions and associated resolution language to the ICPM. ICPM participants were then free to make predictions during the life of the question.

Imputed Data Collection: Five analysts were recruited to read the analysis reports and to provide imputed probabilities for the forecast questions based upon a reading. All five readers had significant intelligence analysis or policy experience. The readers were asked to read the reports and to provide four estimates in the following order:

1. Initial Personal Probability: *Before* reading a given analysis report, they provided their personal probability that the events in question would occur. In other words, this probability was based solely on the readers' personal beliefs about the likelihood of an event.
2. Imputed Probability: They read the entire report and provided their interpretation of the likelihood the report implied that the events in question would occur.

https://doi.org/10.1017/S1930297500007105 Published online by Cambridge University Press

Readers were explicitly instructed, "When making this estimate, consider only what is written in the analytical product and report the estimate that you feel the product implies about the likelihood of the event." Readers could refer to the entire report to arrive at their interpretation.

3. Imputed Probability in Light of Current Events: Readers added an imputed probability based upon the report and current events. Since they may be reading a report months after it was written, this probability represents readers' interpretation of how the analysis report applies to the current situation and allows them to incorporate their knowledge of events that occurred after the reports were published. Again, readers could refer to the entire analysis report to arrive at their interpretation.
4. Updated Personal Probability: Readers provided a second personal probability after they provided the imputed probabilities. This second personal probability reflects the readers' updated personal estimate of how likely they believed the events in question would occur. The fact that these updated estimates came after they read the analysis reports allowed us to determine how the reports influenced their personal beliefs.

Although all five readers received the analysis reports, none of the reports or forecast questions were reviewed by all five of the readers. Of the 99 forecast questions, there were 17 questions where four readers provided imputed and personal probabilities, 26 questions where three readers provided probabilities, 40 where two provided probabilities and 16 questions where just one reader provided probability judgments. Remember that the readers were themselves seasoned analysts who were among the limited population of individuals who are allowed access to all of the analysis reports. They had significant current responsibilities and provided as much assistance to our research as they could.

2.6 Question resolution

The ICPM has well-established procedures for assessing the resolution of posted questions. Below is an example of the language used to specify the assessment of ground truth:

> Outcome will be resolved based on reporting from BBC News or Reuters or Economist Online (http://www.bbc.co.uk/news/ or http://www.reuters.com/ or http://www.economist.com), or at least two independent products or reports. Administrator reserves the right to use other sources as needed (e.g., CIA World Factbook, Wikipedia), provided those sources do not directly contradict concurrent event reporting from BBC News, Reuters, or Economist Online, or multiple classified products. In cases of substantial controversy or uncertainty, Administrator may refer the question to outside subject matter experts, or we may deem the question invalid/void.

In essence, questions were resolved based upon available reporting. In cases of potential controversy, the ICPM administrators planned to contact subject matter experts to help resolve the questions. This did not occur for questions created for this study. There were two questions in this study where the ICPM administrator determined that the questions could not be resolved. These two questions were removed from this study and are not included in the 99.

3 Data analysis

In total we developed 105 questions for this study. As noted above two questions were removed because they could not be resolved. In addition, four questions from one report were excluded due to researcher error in distributing these questions to the readers. Thus, 99 questions were included in the analyses; 96 questions were binary and 3 had three possible outcomes. In addition, one of the readers did not properly follow directions for one question, and thus, this reader's estimates for that question was removed from the statistical analyses.

3.1 Data analysis procedures

In order to statistically analyze the accuracy of estimates made from analysis report imputations and ICPM forecasts, t-tests were conducted on three question categories: All questions (both Non-Fuzzy and Fuzzy), Non-Fuzzy questions only, and Fuzzy questions only.

ICPM forecast probabilities are updated continuously whenever a participant made an investment. The ICPM administrators took a daily snapshot of the probabilities for each question and maintained a history of those daily snapshots. For the data analyses in this paper we selected the daily ICPM forecast probability on the day that the readers submitted their imputed probabilities. So if three readers submitted their imputed probabilities on three separate days, then we matched those imputed probabilities to the ICPM probabilities for those three days.

Most ICPM questions were open for several months to a year, but readers typically returned their imputations within a month. As a result, imputed probabilities were compared to ICPM probabilities early in the posting period.

Two types of error scores were calculated: absolute and squared error (Brier score). Absolute error has the advantage of being easy to comprehend (and equally sensitive over the 0–1 probability range), while the Brier score is often used because it is a proper scoring rule – expected error is minimized by stating one's true beliefs (Brier, 1950).

https://doi.org/10.1017/S1930297500007105 Published online by Cambridge University Press

In addition, when comparing ICPM and imputed probabilities, error scores were calculated using both an average-of-errors and an error-of-averages approach. In the average-of-errors approach for each question the error score for each reader's probabilities are calculated first, and then averaged. So if three readers submitted their probabilities on three different days, then we would calculate three different error scores for their probabilities, and also for the corresponding three ICPM probabilities, and then average those errors. For the error-of-averages approach for each question the average probabilities was calculated first and then the error score for each probability was calculated. Again, if we had three readers, then we would average the three imputed probabilities, and also average the three corresponding ICPM probabilities, and then calculate the error score for those averages.

Results were substantively identical irrespective of how the error scores were calculated or averaged. Consequently, we show only absolute error results below, and when comparing ICPM to imputed probabilities we show the error-of-averages results.

Finally, we note that all significance tests reported below are two-tailed.

4 Results

Results are partitioned into two sections. The first section addresses the relationship between personal and imputed probabilities. The second section examines the comparative accuracy of the ICPM and the analysis reports.

4.1 Relationship between personal and imputed probabilities

The data analyses below examine the relationship between personal and imputed probabilities. Specifically examined are whether (a) readers' personal probabilities biased their imputations, (b) readers' personal probabilities were affected by their reading of the reports and (c) reading the analysis reports led readers to be more accurate.

4.2 Effect of personal probabilities on interpretation of analysis reports

Readers were asked to provide estimates in a specific order, personal probabilities first, followed by their imputed, imputed + current, and updated personal probabilities. To examine whether personal probabilities influenced readers' imputations, the frequency that personal and imputed beliefs were in the same direction relative to the average of imputed beliefs was examined. To illustrate, imagine that three readers had personal probabilities of 30%, 50% and 60%, and imputed probabilities of 20%, 30% and 70%. The average imputed probability is 40%. The first reader's personal probability was below 40% and so was her imputed; the second reader had personal and imputed probabilities that were above and below average respectively, while the third reader had personal and imputed probabilities that were both above 40%. The first and third reader had imputed and personal probabilities that were in the same direction while for the second reader they were in the opposite direction.

In general, if imputed probabilities are not biased by readers' personal probabilities then personal and imputed probabilities should be equally likely to be in the same or opposite direction. Note that this logic applies only to the initial imputation where readers are asked to interpret the analysis report as written. The second imputation, where readers incorporate events that occurred after the report was written, should be "biased". This is because a reader's knowledge of events that occurred after the report was written should influence both their personal and imputed + current probabilities.

Table 1 shows that about 61% of the imputations were in the same direction as the personal probabilities and 39% were in the opposite direction. This analysis considered only forecast questions where there were at least two readers provided probabilities. For the three questions that were non-binary (probabilities assigned to more than two possible outcomes) only the probabilities assigned to the true outcome were counted. A ratio of 61:39 compared to 50:50 corresponds to a Cohen's d effect size of approximately 0.11. This would typically be described as a "small" effect.

It is important to note that the extent to which imputations may be biased is limited by the extent that reports are clear in their probability statements. If, instead of a qualitative forecast, a report stated "70% chance" then it is unlikely that readers would impute anything other than "70%". Only when reports leave room for differing interpretations is there room for the readers' personal views to affect their imputations. In this study, there were 83 forecast questions where two or more readers provided probabilities. In 22 of these 83 questions the imputed probabilities differed by .5 or more. Clearly the reports left substantial room for substantially differing interpretations.

On balance these results suggest that the professional analysts who were our readers did a reasonable job of putting aside their personal views when making imputation judgments, but that they were not immune from this effect.

4.3 Effect of interpretation of analysis reports on personal probabilities

The second question is whether reading analysis reports influenced the readers' personal probabilities. This was assessed by examining the change in readers' personal probabilities before and after reading the analysis reports. If the analysis reports affect readers' estimates, one would expect that updated estimates would move in the direction of imputed probabilities. If readers are largely ignoring the

https://doi.org/10.1017/S1930297500007105 Published online by Cambridge University Press

Table 1: Direction of initial personal probability relative to imputed.

Personal to imputed	Same Direction	129
	Different direction	82
	Sign test	<.002
Personal to imputed + current	Same direction	136
	Different direction	81
	Sign test	<.001

Table 2: Directional changes in updated personal probabilities from initial personal probabilities.

	Shift in personal probabilities	
Change relative to imputed probabilities	In direction of imputed?	153
	Away from imputed?	37
	Sign test	<.001
Change relative to imputed + current probabilities	In direction of imputed?	179
	Away from imputed?	9
	Sign test	<.001

analysis reports, one would expect to see updated estimates moving in the direction of imputed estimates and away from the imputed estimates at the same rate.

As is shown in Table 2, when a shift occurred, readers' updated personal probabilities shifted in the direction of imputed probabilities around 80% of the time and in the direction of imputed + current event probabilities 95% of the time. These results demonstrate that readers are taking what they learned in the reports and using that information to update their personal beliefs.

Comparing the ratios in Tables 1 and 2 (129:82 vs. 153:37, p<.001, z-ratio), suggests that the influence that analysis reports had on reader judgments was somewhat stronger than the influence that reader prior judgments had on their interpretation of the reports.

4.4 Effect of analysis report on accuracy of updated personal probabilities

The third question is whether or not reading the analysis report improved the accuracy of readers' personal probabilities. Overall updated probabilities were more accurate in 113 instance and less accurate in 91 instances. This difference is not statistically significant.

Note that on average the readers' initial probabilities were *more* accurate than the imputed probabilities, with an average absolute error of 0.371 vs. .416 (p<.05, paired t-test). This is not too surprising. Previous research has robustly shown that deep expertise does not result in more accurate geopolitical forecasts (Tetlock, 2005). So even though the report authors were more knowledgeable than the readers, there was no reason *a priori* to expect that their forecasts would be more accurate. The fact that our readers' forecasts were a little more accurate than the analysis reports probably reflects the fact the readers were in fact professional seasoned analysts, as well as the variance involved in the imputation task itself.

Given that initial probabilities were more accurate than imputed probabilities, and that updated probabilities moved in the direction of the (less accurate) imputed probabilities, it's perhaps surprising that the updated probabilities were even a little more accurate than the initial. The explanation has to do with the pattern of how readers updated their personal probability judgments.

Table 3 partitions the 190 instances shown in Table 2 where readers updated their probabilities in the same or opposite direction of a report's forecasts. Of particular note are the 37 forecasts where readers updated their judgments by moving their personal probabilities in the *opposite* direction of the imputed probabilities. For example, in one instance a reader had an initial probability of 60%, read the report and imputed 75%, and then revised their personal probability down to 25%. In 32 of these 37 instances the updated probabilities were more accurate and in 5 instances the updated probabilities were less accurate. This 32:5 ratio is statistically significant (p<.001, sign test) as is the comparison of the 72:81 to 32:5 ratios (p<.001, Fisher exact).

Keeping in mind that this is only a small subset of the data, the results for these 37 forecasts do suggest that readers were influenced by more than a report's forecasts. They could digest the analysis of key drivers and sometimes use that analysis to correctly revise their personal conclusions in the opposite direction of what was concluded in the report.

4.5 Accuracy profile of analysis reports and ICPM estimates

Overall the ICPM was more accurate than the imputed probabilities. The mean absolute error for the ICPM and Imputed probabilities was .302 and .416 respectively. That is to say, compared to ground truth, the mean ICPM forecast was 69.8% while the mean imputation was 58.4%. This difference is explored below.

4.5.1 Empirical comparison of estimates from analysis reports and the ICPM

Table 4 contains a summary of each t-test analysis performed with absolute error as an index of accuracy. The ICPM scores were significantly lower than the imputed scores for both non-fuzzy and fuzzy questions

https://doi.org/10.1017/S1930297500007105 Published online by Cambridge University Press

Table 3: Directional accuracy of updated personal probabilities partitioned by direction of update.

	Updated personal probability more accurate than initial	Updated personal probability less accurate than initial	Total
Personal probability revised in *same* direction as imputed probability	72	81	153
Personal probability revised in *opposite* direction of imputed probability	32	5	37
Total	104	86	

Table 4: Comparative accuracy of imputed and ICPM estimates.

	N	Mean error imputed	Mean error ICPM	p (2 tailed)	Difference
All questions	99	0.416	0.302	<.0001	0.114
Non-fuzzy questions	71	0.412	0.305	<.0004	0.107
Fuzzy questions	28	0.427	0.300	<.004	0.127

The analysis reports represented the analysis community's best current analyses and forecasts, but the fact that the ICPM was more accurate than the reports does not necessarily entail that this difference was due to the ICPM methodology. ICPM forecasts were updated daily and reflected the latest available information. By contrast the analysis reports were static and reflect only the information that was available before the report was published. Usually there was at least a one-month delay between the time the report was published and when the forecast questions were posted on the ICPM. Consequently, the superior performance of the ICPM could result from fact that the ICPM forecasts were based on additional recent information. The two analyses below examine this possibility.

4.5.2 Impact of the delay between publication of analysis reports and posting on the ICPM

For all questions, there was a delay between the time the analysis report was published and the time that the forecast questions were posted to the ICPM. Across the 99 forecast questions, this posting delay ranged from 13 to 237 days. Because of the delay, ICPM participants had information available to them that was not available to the authors of the analysis reports. Consequently, the accuracy advantage of the ICPM over analysis reports may be due to this additional information. If this is the case, one would expect that longer posting delays would yield an increasing accuracy advantage for the ICPM.

Table 5 depicts the relationship between the posting delay and the relative accuracy of the ICPM and analysis reports. The forecast questions were partitioned into three bins based upon the posting delay. Irrespective of the posting delay, the ICPM outperformed the analysis reports, where the ICPM advantage *decreased* with longer delays. This result is robust irrespective of how the delays are binned.

For the questions in this study, longer posting delays yielded on average a lesser advantage for the ICPM, not greater. This result suggests that the accuracy advantage of the ICPM is due to the forecasting method and not the additional information available to ICPM participants.

4.6 The influence of current events on imputation accuracy

Recall that readers were asked to make two imputations, the first based only on what they read in the report and the second to estimate what they thought the report implied given current information. These Imputed+Current estimates incorporate the same additional information that is available to participants in the ICPM. Consequently, if the ICPM advantage is due only to better information, then the ICPM advantage should disappear. In fact, it does not. The mean error score for the ICPM, Imputed, and Imputed+Current probabilities was .302, .416, .394 respectively. The error score for the imputations does decrease when current information is included, but there remains a significant difference between ICPM and the Imputed+Current estimates (p<.001, paired t-test). This analysis suggests that, at most, only part of the ICPM advantage could be attributed to additional information being available to ICPM participants. (However, the set of pooled information of all the ICPM participants might have been greater than the information available to the analysts. This is a characteristic of prediction markets.)

https://doi.org/10.1017/S1930297500007105 Published online by Cambridge University Press

TABLE 5: Comparison of ICPM and analysis report accuracy for different posting delays.

	Number of days until posted		
	10 to 35	36 to 50	More than 50
Number where ICPM more accurate	18	37	14
Number where analysis reports more accurate	4	16	10
Average difference in absolute error	0.170	0.117	0.017
Sign test	<.01	<.01	n.s.

TABLE 6: A calibration analysis of imputed and ICPM estimates.

		Bin midpoint				
		10%	30%	50%	70%	90%
Imputed Estimates	Number of questions contained in a bin	19	29	22	24	5
	Percentage in bin that occurred	21%	3%	9%	33%	60%
ICPM Estimates	Number of questions contained in a bin	33	35	10	15	6
	Percentage in bin that occurred	3%	11%	30%	27%	100%

4.7 Calibration analysis

Calibration is the extent to which observed frequencies match the forecasts. A forecaster is well-calibrated, for example, if 20% of the events forecasted with 20% certainty occur. A forecaster is poorly calibrated if say 70% of those events occur. Knowing the extent to which a forecaster is calibrated is useful information for the consumer of a forecast, as it helps the consumer to assess how much confidence she should put into forecasts from that forecaster.

Mathematically calibration is only one of several elements that are incorporated into error scores. Consequently, even though the ICPM was more accurate than the analysis reports, it might still be that the reports were better calibrated. If this were the case, then it could be argued that the forecasts in the reports are more useful than the ICPM forecasts.

The calibration of the imputed and ICPM probabilities was examined by placing the probability estimates into bins and examining how many of the events in each bin occurred. Bins were created to try to provide a reasonable number of data points for equally sized bins. Table 6 shows these results in tabular form, where the range for each bin was 20% (e.g., the first bin included any estimate that ranged from 0% - 20%).

ICPM forecasts were better calibrated than the analytic reports, but both the report and ICPM forecasts exhibited poor calibration. Both exhibited overestimation of the likelihood of event occurrence. In many cases, the Imputed and ICPM probabilities were statistically significantly lower than perfect calibration:

For Imputed estimates:

1. For the 30% bin, 1 out of 29, p < .001.
2. For the 50% bin, 2 out of 22, p < .001.
3. For the 70% bin, 8 out of 24, p < .001.

For ICPM estimates:

1. For the 30% bin, 4 out of 35, p < .02.
2. For the 70% bin, 4 out of 15, p < .002.

4.8 Comparison of calibration to other similar studies

The data suggest that the analysis reports and ICPM participants were poorly calibrated and overly certain in many of their forecasts. Of interest is how this compares with other studies where professional analysts generated geopolitical forecasts. One way to make this comparison is to compare the Calibration Index (CI) of the estimates from each effort. CI is calculated by taking the sum of the squared deviations of estimates from ground truth relative frequencies, weighted by the number of questions that fall within a given bin. The best possible CI score is 0, when there is no deviation from ground truth frequencies. The CI for the current study is .097 for Imputed estimates, and .047 for ICPM estimates. These compare poorly to other studies where analysts made similar types of forecasts. Lehner et al. found that probabilities imputed from unclassified analysis reports exhibit a CI of .018. Studies by Tetlock (2005), Mandel, Barnes and Hannigan (2009), and Mandel and Barnes (2014), where experts were asked to make quantitative probability judgements, were all better calibrated than the forecasts in this study, at .025, .014, and .016, respectively.

https://doi.org/10.1017/S1930297500007105 Published online by Cambridge University Press

There is no obvious explanation for the relatively poor calibration found in this study. It seems likely to us that it has something to do with the forecast questions themselves. The forecast questions seemed to the authors to be particularly challenging. Along these lines note that Mandel and Barnes, who assessed the calibration of quantitative probabilities in Canadian intelligence analysis reports, found that the analysts' estimates were better calibrated for easier-to-forecast questions.

5 Discussion

Summarizing the important results. First, on a collection of 99 forecasts derived from 41 analysis reports, the probability forecasts in a prediction market were significantly more accurate than the imputed forecasts made by analysts who read the reports. This result is robust even after accounting for the possibility that ICPM participants had access to more recent information than the report authors. This result is consistent with other research showing crowd wisdom to be more accurate than either individual or consensus expert judgments. Second, readers who updated their personal probabilities after reading the reports did not on average become more accurate. However, in the small number of instances where readers updated their probabilities in a direction opposite to what a report implied, they were significantly more likely to update their probabilities in the correct direction. This could only occur if readers could glean forecasting relevant information from the reports other than the forecasts.

Below we discuss the limits and implication of this research and possible future directions.

Forecast clarity: This research focused on evaluating expert forecasts that are expressed in their normal format – qualitative forecasts in analysis reports. We chose analysis reports that addressed sensitive topics of substantial geopolitical importance where each report reflected the contribution of multiple seasoned and respected intelligence analysts with diverse areas of relevant substantive expertise. Because of their sensitivity there is only a limited community of individuals who are granted access to these reports, and we drew our readers from this community. Thus, there were only two or three readers for most of the reports – and these readers sometimes disagreed substantially on their imputations. This source of variability does not affect the validity of any statistical significance results described in this paper. However, we conjecture that reports with clearly stated forecasts are also likely to be more accurate. Because we had a small and varying number of readers for each report, and our measure of clarity would be a function of the variability of reader imputations, we could not test this conjecture with our data. It would be interesting in a future study to assess whether greater clarity in forecasts is an indicator of greater accuracy or just false confidence.

Are imputed probabilities fair? Some reviews of this paper may express concern that imputed probabilities do not reflect the probabilities of the authors who wrote the analysis reports and therefore do not represent a fair test. In response, we note that in previous studies comparing expert judgment to crowd wisdom forecasts, experts were required to provide numerical probabilities. This could also be viewed as an unfair comparison because experts do not normally provide forecasts as numbers. This study endeavored to evaluate expert forecasts in their natural form – qualitative certainty expressions in carefully written analysis reports. Results were consistent with previous studies. In general crowd wisdom forecasts are more accurate than expert forecasts, no matter how the expert forecasts are expressed. Still, inference of numerical probabilities from verbal statements is inherently noisy sources of information about probability, especially when they are not written in order to allow someone else to extract a numerical probability.

Integration of crowd wisdom and traditional analysis. In addition to relative accuracy, the results presented in this paper do provide some evidence that analysis reports helped readers to understand a substantive domain and to improve their personal probabilities; and that this benefit could not be attributed to adjusting personal probabilities in the direction of a report's conclusions. This indicates that there is value to the written analyses in the reports beyond the forecasts; and therefore, a possible integration of the two approaches where crowd wisdom methods become critical to forecasting *what* events will occur, but traditional written analysis reports remain critical to helping analysts and decision makers to understand *why* those events may occur. The question of how to combine these two approaches into a single integrated approach is an important research question of considerable practical importance.

References

ACE (2016). Aggregative Contingent Estimation (ACE), retrieved on 7/15/2016 from https://www.iarpa.gov/index.php/research-programs/ace.

Beyth-Marom, R. (1982). How probable is probable? A numerical translation of verbal probability expressions. *Journal of Forecasting, 1,* 257–269.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78,* 1–3.

Erev, I. & Cohen, B. L. (1990). Verbal versus numerical probabilities: Efficiency, biases, and the preference paradox. *Organizational Behavior and Human Decision Processes, 45,* 1–18.

https://doi.org/10.1017/S1930297500007105 Published online by Cambridge University Press

Gardner, D. (2010). *Future babble: Why expert predictions fail – and why we believe them anyway.* Toronto: McClelland and Stewart.

Kent, S. (1964). Words of estimative probability: History of a semantics problem. *Studies in Intelligence, 8*(4), 49–66.

Lehner, P., Michelson, A., Adelman, L., & Goodman, A. (2012). Using inferred probabilities to measure the accuracy of imprecise forecasts. *Judgment and Decision Making, 7*, 728–740.

Mandel, D. R. & Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. *The Proceedings of the National Academy of Sciences, Early Edition,* 1–6.

Mandel, D. R., Barnes, A., & Hannigan, J. (2009, February). A calibration study of an intelligence assessment division. Paper presented at the *Global futures forum community of interest for the practice and organization of intelligence Ottawa - What can the cognitive and behavioural sciences contribute to intelligence analysis? Towards a collaborative agenda for the future*. Meech Lake, Quebec.

Mosteller, F. & Youtz, C. (1990). Quantifying probabilistic expressions. *Statistical Science, 5*(1), 2–34.

Surowiecki, James. *The wisdom of crowds*. New York: Doubleday, 2004.

Tetlock, P. (2005). *Expert political judgment*. Princeton: Princeton University Press.

https://doi.org/10.1017/S1930297500007105 Published online by Cambridge University Press