ERASMUS UNIVERSITEIT ROTTERDAM

Erasmus University Rotterdam
Master thesis

Predicting a Pandemic

Testing the wisdom of crowd in forecasting cases during the Covid-19 outbreak

Author: Gaoyi Fang

Student number: 409392

Supervisor: A. Baillon

Specialization: Behavioral Economics

Abstract

Over the past decades, the wisdom of crowd phenomenon has become an increasingly relevant and well-studied topic thanks to the rapid development of the internet. Instead of relying on the answers of individuals, performance is generally better when relying on the group answer. That is, combining individual's answers by mathematical aggregation. In this thesis I study crowd wisdom in predicting corona (Covid-19) infections at the end of May, where the period of collecting data is between the 10th of April and the 13th of May. I ask subjects to guess the number of accumulated Covid-19 infections in eight different countries, one question asking the number of hours the virus is able to survive on plastic, and when the first Covid-19 infection was detected in the Netherlands. Firstly, I analyse the effect of diversity by constructing "diverse" and "non-diverse" groups, where the difference in distribution is tested with a Mann-Whitney U test. Subsequently, I analyse individual performance by testing the correlation between the given confidence interval size and the individual error, and the correlation between source reliance and the individual error. A random effects model is used for this. The results of this paper show that diversity has no significant effect on group performance and social media reliance does not have a significant (negative) impact on individual performance. However, a wider confidence range is found to be negatively correlated with individual performance. Moreover, a wider confidence range is correlated with poor individual performance.

Table of Contents

1. Introduction

Relying on the judgement of large groups as opposed to an individual has been a commonly known method to increase accuracy in prediction outcomes for a while. Back in 1906, an English scientist made a remarkable discovery at a fair. Of 800 people, he collected their guess about the weight of a dead ox. On the individual level, no one guessed the exact answer. However, when he aggregated all the answers and took the average, all those 800 guesses came down to an answer of 1197 pounds (Galton, 1907). Considering the real weight being 1198 pounds, the outcome is very noteworthy. This event perfectly captures the term *The Wisdom of Crowds*. A crowd of people is often more accurate than an individual person, as the errors will likely cancel each other out when the group is large enough. Since then, the idea of crowd wisdom has been applied to diverse fields like economics and politics. For example, Franch (2013) applies the wisdom of crowds to political markets by using social media platforms like Facebook, Twitter, YouTube and Google to predict the election outcomes. In his paper, he aggregates the political beliefs by specific social media platform and overall as well. He found that aggregating the information from social media platforms serves as a very reliable measure to predict the election outcome. Similar to the social media application for wisdom of crowd testing by Franch (2013), others have already used an analogous method to forecast the spreading of Covid-19. Turiel & Aste (2020) analysed the number of Covid-19 related tweets and discovered a significant relationship between the trend of social media attention and the cumulative deaths caused by the virus in Spain and Italy. They infer that social media can be used as a wisdom of crowd platform to predict the virus spreading. Therefore, predictions are not only useful in business, but are also vital regarding disease trends. Besides, how a pandemic will develop in the future will certainly influence business decisions. However, when no historical data are available, forecasting becomes challenging. One way to

decrease such uncertainty is to create a prediction market. By crowdsourcing opinions from the public and/or experts and aggregating them, one could make a forecast prediction of unknown future events. To illustrate, in 2003, contracts have been drawn up within the Defence Advanced Research Projects Agency (DARPA) based on probability outcomes regarding economic and political events (Wolfers & Zitzewitz, 2004). The idea was that payoffs would be provided to those whose estimated guess turned out to be the correct one. In other words, individuals participating in a prediction market are asked to buy and sell shares in outcomes and will be rewarded if the outcome they bet on turns out to be the correct one. Despite its limitations, prediction markets have sparked interest since some successful applications have occurred, such as predicting the outcome of football matches and presidential elections (Wolfers & Zitsewitz, 2004). Hewlett-Packard (HP) has used internal prediction markets to forecast the sales of its printers, which has actually outperformed the forecasts of the company's own sales forecast (Polgreen et al., 2007).

Li et al. (2016) demonstrate the use of prediction markets for predicting the trend of epidemic diseases. They gathered the participation of health care professionals in Taiwan and found more accurate results compared to Taiwan's Center for Disease Control. Information gathered from aggregated opinions can therefore be very useful, especially in the absence of historical data sources. Not only using experts' opinions, but also information gathering from the common crowd can prove its usefulness (football matches and presidential elections). Some studies have shown that a crowd can make better decisions compared to experts. Nofer & Hinz (2014) found superior performance of the online community over institutional investors in the stock market. The results of their analysis showed that a stock prediction community outperformed the stock market in terms of excess returns. An important explanation they give is that crowds

benefit from independence in their answers, whereas experts have a higher chance of discussing possible outcomes with each other, causing their input to be influenced by other members.

Platforms such as Wikipedia and Reddit have shown that taken together, people acquire a large amount of knowledge on diverse topics. The internet is a very suitable and accessible platform for many people from different places in the world. The power of the internet also becomes conspicuous when used for predicting future events.

In this paper, I test the effectiveness of the wisdom of the crowd on a topic that is currently affecting everyone around the world, the coronavirus (Covid-19). By asking respondents several predictive questions relating to the number of cumulative virus infections, I test whether the crowd is able to make a relative reliable prediction compared to the individual level, and in general. Therefore, to test the potential wisdom of the crowd regarding Covid-19 infections, I will make use of the "common" crowd, where the main population of my sample will be students. Thus, the research question is stated as follows:

How reliable is the crowd's prediction regarding infections in the Covid-19 pandemic?

The next section will dive deeper into the existing literature, followed by the methodology, data analysis, results, discussion and conclusion.

2. Theoretical Framework

2.1 Wisdom of the crowd

The existing literature is rather rich in studies about how consumers' judgement has improved. Many studies have proved the wisdom of crowd hypothesis to be valid. In the case of the Ox example presented in the introduction, the crowds' mean average of 1197 pounds only deviated 1 pound from the true weight. This event has become the starting point in the literature presenting the power of collective intelligence. Similarly, Treynor (1987) conducted a bean jar experiment where he asked a group of 48 and 56 students to guess the number of beans in the jar. In the first group, the average estimation was 841 beans, where the jar was actually holding 810 beans. Only two students guessed closer to the actual number compared to the mean guess. Subsequently, in the second group, the jar contained 850 beans, whereas the average guess was 871. One person managed to guess a number lying closer to the real number compared to the average guess. Additionally, in their paper *Word-of-mouth and the forecasting of consumption enjoyment*, He & Bond (2013) discuss consumer knowledge of services and products for the forecasting of consumer enjoyment. They argue that the average rating of all consumers proves to be more valuable in making precise forecasts compared to the rating of a single reviewer. This idea is in line with the main principle of the wisdom of crowds.

2.2 Conditions

What determines a wise crowd? According to Surowiecki (2005), the diversity and independence of a group are the two most important factors in order to obtain the best crowd performance. Diversity is defined as the private and individual information and interpretation each

person holds, whereas the independence of a crowd indicates that an individual's answer should not be dependent on the opinion of someone else. Thus, when answering the questions, every person should answer without consulting or discussing with anyone.

Diversity allows access to different sources of information that in turn influence the answer given. This enhances overall performance, since the errors a large diverse set of answers can cancel each other out to reach the correct answer. Nofer & Hinz (2014) study the wisdom of crowd effect on predictions in the financial stock market. More specifically, studied the impact of independence and diversity on prediction accuracy. They found that crowds have the ability to outperform the market and a small group of financial analysts. Additionally, they found a positive and significant effect of independence on the risk-adjusted returns for the crowd, indicating that the independence of participants in a crowd has a positive effect on their quality of stock recommendations. On the other hand, they found no significant effect on diversity. This may be explained due to the exclusion of gender diversity, since females represented less than 5% of their sample. It is expected that in large groups, diversity is likely to occur. However, diversity can be defined in more than one way. It is a multi-dimensional construct that can imply age or gender, but also one's education, knowledge and skills (Arazy et al., 2006). In their research, Arazy et al. (2006) test to what extent the conditions of the wisdom of crowd determine the quality of Wikipedia articles. They define *diversity* as the diversity in opinions, which they measure in the following ways: (1) the word count in the discussion page of an article and (2) the number of "edit wars" (three edits by one user within 24 hours). The authors found a significant causal relationship between diversity and the quality of the articles. This illustrates the importance of different opinions integrated into an article for the output to be of high quality. Another way to measure diversity is taking into account the variance. Hong et al. (2016) define diversity as the distance between opinions, which in turn is calculated as

the variance across all observations on a given day. They argue that a crowd exhibiting high levels of opinion distance implies that crowd will display high levels of opinion diversity. Their hypothesis is in line with their results, as their findings suggests that opinion distance is highly significant and positively correlated with crowd performance. This strengthens the argument that diversity is needed for a crowd to perform well.

Based on the findings in the literature, the first hypothesis will be stated as follows:

H_1: *Diversity has a positive effect on the accuracy of crowd wisdom.*

Despite the fact that the wisdom of the crowd is a statistical phenomenon, social influence can impair the quality of the crowd, because it influences individual decision making. Lorenz et al., (2011) studied the effect of social influence in an experiment where respondents are asked to answer factual questions. They found that by revealing what the answers of others were, subject's propensity to convert their answer increased, which decreased diversity of the crowd and did not improve accuracy. Dependent answers cause the aggregate answer to deviate away from the centre. An independent crowd prevents individuals from making correlated mistakes and brings a variety of perspectives and ideas to the table. Therefore, making the group more reliable. Furthermore, opinion leaders can weigh down the efficiency of crowd wisdom (Golub & Jackson, 2010). If participants depend too much on the information coming from such small groups, the aggregate answer is likely to be biased. This idea is related to the third condition Surowiecki deems important; decentralization of the crowd, meaning that no one is imposing the crowd's answer. Every person learns and absorbs information from their own sources. An example of this is the aforementioned Wikipedia, which employs an open platform system where users can independently submit or edit

webpages. Wikipedia's accuracy has proven to be similar to that of the Encyclopedia *Britannica* (Arazy et al., 2006).

2.3 Overconfidence

According to Grushka-Cockayne et al., (2017), overconfidence is one of the main reasons for poor forecasts. Additionally, psychologists have noticed a general tendency of providing overconfident forecasts by individuals. Especially when there is a competitive element involved, people have the propensity to be overconfident in their forecasts (Grushka-Cockayne et al., 2017). A lot of studies have found that generally speaking, people often overstate the capacity of their knowledge accuracy. When it comes to forecasting, overconfidence becomes even more apparent when past events have led a person to (falsely) believe that his/her predictions will be accurate. For example, Hilary & Hsu (2011) investigate overconfident behaviour in managers' forecasts. They found a positive and significant correlation between the managers' current forecast error and the number of correct predictions in the past. This indicates that too much weight is placed on personal information and too little on public signals. To what extend individuals exert overconfidence is also related to the way confidence is elicited. For example, in binary questions, one can choose which one is correct and subsequently state his/her confidence about the correctness of his/her answer: "Which city lies more north, New York or Rome?" "New York, I am 80% sure". However, how should you validate this level of confidence? For example, if the above answer was to be incorrect, can you state that person was overconfident? A common way to approach this is to assess the calibration of a persons' confidence. There are many studies where participants are asked to provide an upper and lower limit of their confidence interval, where numeric answers are involved. For example: "How many reported crimes were there in Spain in

2008?" "I am 80% sure that the answer lies between 300 and 400." According to Soll & Klayman (2004), people are more likely to provide overconfident answers in the latter example. Subjects report confidence intervals that are too narrow. In other words, when asked to provide a subjective confidence interval, such that the person is 80% sure their answer lies in their confidence range, their answers actually lie in the confidence range less than 80% of the time (Soll & Klayman, 2004). Perhaps those subjects are not aware of how much they know, or in this case, of how much they do not know. Overconfidence seems to be especially present in quantitative confidence interval estimates. Also, Russo & Schoemaker (1992) found that even experts have a hard time correctly calibrating their confidence interval. When asking their subjects to provide a 90% confidence interval, more than 50% of their answers were incorrect. This indicates the general difficulty in calibrating a subjective confidence interval. In addition, Liu & Tan investigate the effect of overconfidence on forecast accuracy in the financial market. In their experiment, participants are incentivised to provide their confidence interval on their stock price forecast prediction. Their study found that overconfident participants make the least accurate predictions, compared to those that are less confident. Thus, they established a negative relation between overconfidence and forecast accuracy. Furthermore, it has been found that overconfidence increases with the degree question difficulty; this effect is known in the literature as the hard-easy effect. (Fischhoff et al., 1977). In other words, difficult tasks are associated with a higher level of overconfidence, whereas underconfidence is associated with easier tasks. Considering that participants will be asked to guess the number of corona infections per country in the future, it would be safe to label the required task for the purpose of this paper as hard. Therefore, the second hypothesis will be stated as follows:

***H_2**: overconfidence has a negative impact on the accuracy of predictions.*

2.4 Infodemic

On the 11th of March 2020, the World Health Organisation (WHO) declared the coronavirus a global pandemic. More than 90 per cent of all countries have been affected by Covid-19 (Hua & Shaw, 2020). Such an interconnected event goes hand in hand with the diffusion of a tremendous amount of information, both reliable and unreliable. The spreading of information has a critical effect on how people choose to behave during this pandemic and can have a negative effect on the mitigation effects implemented by government bodies. Therefore, many forecasting models account for information consumption and population behaviour. Thus, besides combating a pandemic, there is also a battle against an infodemic, which is the spreading of misinformation during the handling of a pandemic (Cinelli et al., 2020). Especially nowadays, the speed of misinformation and rumours spreading is amplified through the use of social media platforms. The dispersion of questionable information is also potentially dangerous, since it can cause people to act inappropriately to the situation, which could result in exaggerated panic and unnecessary deaths. Additionally, some fake news circulating on social media platforms can look like it comes from health institutions, by falsely referring to an expert. This makes it seem like the information comes from a reliable source, while the information itself might not be correct. For this reason, WHO's risk communication team launched an information platform right after the infection was declared a humanitarian crisis, in order to nudge people into actions that control the outbreak and softens the blow (Zarocostas, 2020).

Rovetta & Bhagavathula (2020) investigate online search behaviour in Italy during the Covid-19 pandemic with the aim to discover different forms of misguided information. They

define flawed information as *infodemic monikers* and found several infodemic monikers that were widespread throughout Italy. With the use of Google Trends, a significant increase in web searches was observed when the WHO declared Covid-19 a global pandemic. Additionally, search queries increased linearly with the increase in the number of cases in Italy, suggesting anxious behaviour in combination with a growing need to collect information to stay protected and healthy. This in combination with false information, likely has harmful consequences. The openness and accessibility of social media make it susceptible to the exposure of pernicious information, and also the publication of it. Nevertheless, social media is used by many as a source of news, especially by the younger demographic (Nielsen et al., 2020). In their paper, Nielsen et al (2020) research various news sources about Covid-19 and how these are perceived as trustworthy, alongside the sample's knowledge about the virus. They asked participants several factual questions about the disease and ran a regression analysis to determine potential correlations between news reliance and knowledge, while controlling for political orientation, education and age. Their findings showed a correlation between the reliance on news organizations and a higher level of knowledge. No correlation was found between social media reliance and knowledge level in this study; however, another finding is that there is a lot of expressed concern about social media messaging, stating that false and misleading information has been seen on the platforms. Therefore, the third hypothesis that will be tested in this paper is stated as follows:

H_3: *relative stronger reliance on social media channels will result in less accurate predictions*

3. Research methodology

3.1 Research design & data sampling

I collect primary data through the distribution of an online survey. From the 10th of April until the 13th of May, participants are asked what they think will be the number of reported accumulated Covid-19 infections in eight different countries at the end of May (31th of May). Additionally, two questions are asked about the coronavirus itself which required a numeric answer: how many hours the virus can survive on a plastic surface, and on which date the first virus infection was detected in the Netherlands. Thus, the respondents are asked a total of ten questions, of which the first eight are prediction questions for random countries, and the last two are more "general" questions. The questionnaire starts with a short introduction and a graph that shows the trend of how the number of virus infections has developed over previous weeks prior to the first day of the survey distribution. The graph is shown with every prediction question to provide some guidance to the subject about the direction of the past trend in the eight countries. Furthermore, in order to elicit the respondent's 80% confidence range, I ask them to provide the lower and upper limit of their confidence interval, which make them 80% sure that the correct number will lie within these ranges. For example, the question about corona infections in the Netherlands is formulated as follows:

1. What do you think will be the total (accumulated) number of reported corona infections in the **Netherlands** at the end of May?
2. Please provide a lower and upper limit, such that you are 80% sure that the number of total reported infections at the end of May in the **Netherlands** will lie between these numbers:

> **For example**: if you would choose 10 as the lower limit and 50 as the upper limit, this would mean you are 80% sure that the total reported infections at the end of May in the Netherlands is going to be between 10 and 50 infections.

The survey ends with questions about demographics such as age, gender, occupation and ethnicity. The answers will be compared to the actual number of contaminations reported on the 31st of May, of which the data is retrieved from John Hopkins University, whose map maintains data of patients diagnosed with Covid-19. This map is based on information sourced from the World Health Organization (WHO) and the European Centre for Disease Prevention and Control (ECDC).

The primary data sample consists of 224 respondents in total, of which each answered ten questions (excluding the demographic questions). Those providing wrong answers systematically, which implied a lack of understanding or laziness, were removed from the sample. For example, someone's level of confidence is measured by a confidence range, where subjects report the lower and upper limit of their interval, providing their 80% confidence range. Some subjects gave multiple answers where they provided a lower limit that was higher than their upper limit, resulting in a negative confidence range. After removing those, a useable dataset of 192 subjects is left.

3.3 Data analysis

In this section I will elaborate on the analysis methods applied. The aim of this paper is to test whether a group's prediction can make a relatively reliable prediction regarding the Covid-19 infections.

3.2.1 Dependent variable

The dependent variable measures the absolute distance between the actual number of infections reported on the 31st of May and the predicted number of infections at the end of May. Similarly, for the last two questions this means that the dependent variable measures the actual number of hours the virus can survive on a plastic surface (the actual date the first infections was detected in the Netherlands), and the guessed number of hours the virus can survive on a plastic surface (the guessed date the first infections was detected in the Netherlands). What is noticeable when asking such prediction questions, is that the errors people make are inevitably enormous, and vary greatly per question. For example, in general, people might have more general knowledge about the development of the virus in the US (at the time of survey distribution) than the virus trends in Russia, since the former was more spoken about in the news than the latter. Likewise, errors made in the last two questions will be much smaller compared to the first eight. For this reason, I will standardise the errors by dividing the absolute error by the maximum error made for that question. This way, the errors per questions will be relative to the worst answer given (greatest error) and will be between a range of 0 and 1. The dependent variables *ASGE* (absolute standardised group error) and *ASE* (absolute standardised error) can be formulated as follows:

For group I :

$$grouppredictio n_{It} = \frac{1}{N}\sum_{i\ in\ group\ I} prediction_{it}$$

$$absolutegrouperror_{It} = |actual_t - \ grouppredictio n_{It}|$$

$$absoluteerror_{jt} = \left| actual_t - \ prediction_{jt} \right|$$

$$ASGE_{It} = \frac{absolutegrouperror_{It}}{MAX_{\{j\in\{1,\ldots,192\}\}}\{absoluteerror_{jt}\})} \quad (1)$$

Where $prediction_{it}$ is the prediction given by individual i for question t, and $grouppredictio n_{It}$ is the average of the summed-up predictions per individual i of group I for question t. Subsequently, $absolutegrouperror_{It}$ is the absolute difference between the actual number of question t ($actual_t$) and the group prediction of group I for question t. $Absoluteerror_{jt}$ is the absolute difference between the actual number of question t and the prediction made by group j per question t. $ASGE_{It}$ is the absolute standardised group error for group I per question t, which is standardised by dividing the absolute group error of group I for question t by the highest given absolute error of individual i of group J per question t. Formula (1) will be used to test H1.

Formula (2) will be used to test H2 and H3:

For individual $i \in \{1, \ldots, 192\}$:

$$absoluteerror_{it} = |actual_t - \ prediction_{it}|$$

$$ASE_{it} = \frac{absoluteerror_{it}}{MAX_{\{j\in\{1,\ldots,192\}\}}\{absoluteerror_{jt}\})} \quad (2)$$

Where the $absoluteerror_{it}$ of individual *i* for question *t* is calculated by taking the absolute difference between the actual number of question *t* and the prediction made by individual *i* for question *t*. To standardize the absolute error, it is divided by the highest absolute predicted number of group *J* per question *t*. Then, we come to ASE_{it}, which is the absolute standardised error for individual *i* per question *t*.

3.2.2 Diversity

As stated by Surowiecki (2005), diversity is an important determinant in obtaining a well-performing group. In this paper, I will test the effect of diversity on group performance (H1) by randomly constructing 60 groups of 32 subjects. All question will be taken into account for the distribution, which means there are 1920 data points (10 questions x 192 individuals). The randomly constructed groups will represent the *diverse* groups (see appendix B1 for an example). Next, I will take the same dataset and rank on *gender*, *age*, *occupation* and *ethnicity*, in this specific order. Based on these rankings, 60 *non-diverse* groups of 32 subjects are formed. So, for example, I will have a group of 32 subjects with only female students aged 19-22 of which the majority is white (see appendix B2 for an example). This gives me a total of 120 groups, of which 60 are diverse and 60 non-diverse. Thus, with a sample of 1920, I construct "diverse" and "non-diverse" groups, where the distribution between the two groups of the *ASGE* (absolute standardised group error) will be tested with a Mann-Whitney U test.

3.2.3 Overconfidence and source reliance

In the literature, overconfidence is described as a common phenomenon where people are too certain of their own judgements. Overconfidence tends to be even more pronounced when

people are faced with difficult questions. I will test the effect of overconfidence on the prediction accuracy of subjects, by asking the subject's 80% confidence range. More specifically, after guessing the number of Covid-19 infections, subjects report their lower and upper bound so that they are 80% sure that their predicted number will lie between these numbers. This means that someone is overconfident if less than 8 of the 10 questions answered lies within their provided confidence range. In other words, I will test whether the size of the confidence range is correlated with the accuracy of the prediction. Similar to the *ASGE*, I will also standardise the confidence range by dividing it by the largest confidence range reported for that specific question.

In order to test H2, I will treat my data as panel data, where the questions represent the time variable (1 to 10) for each subject *i*. This means that I will have 1920 data points (10*192). Furthermore, I will run a random or fixed effects model where I take ASE_{it} as the dependent variable, and C_{it} (standardised confidence range for individual *i* per question *t*). C_{it} is calculated as follows:

For individual $i \in \{1, \dots, 192\}$:

$$confidencerange_{it} = Upperlimmit_{it} - Lowerlimmit_{it}$$

$$C_{it} = \frac{confidencerange_{it}}{MAX_{\{j \in \{1,\dots,192\}\}}\{confidencerange_{jt}\})} \quad (3)$$

Where the *upperlimmit* and *lowerlimmit* are the upper and lower limit of the individual's 80% confidence range per question *t*, which is then standardised by dividing the confidence range by the maximum confidence range given in group *j* for question *t*.

Additionally, H3 states that subjects whose social media reliance is relatively high compared to news outlet sources, are less accurate in their predictions. This is because a lot of false information is circulating through social media. Subjects are asked to give a 1-10 score for both social media and news outlet sources. That is, how much they rely on both for Covid-19 related information.

Consider a regression model for individual i,…, N, who is observed at time t = 1 until t = T. Then the model can be written as follows.

$$ASE_{it} = \alpha_i + \beta_1 C_{it} + \beta_2 SMreliance_i + \beta_3 Informed_i + \beta_4 Newsoutlet_i + \beta_5 Age_i + \beta_6 Occupation_i + \beta_7 Female_i + \beta_8 Ethnicity_i + \epsilon_{it} \quad (4)$$

See table 1 below for further explanation of the variables.

Table 1. Variables explanation.

Notation	Name	Explanation
ASE_{it}	Absolute standardised error	Absolute standardised error of individual i for time t
C_{it}	Confidence range	The confidence range of individual i for question t
$SMreliance_i$	Social media reliance	Takes the value zero if the amount of social media reliance relative to news outlet reliance is less or equal to 65%, and takes the value one if it is more than 65%
Age_i	Age	Age of the individual
$Informed_i$	Informed	How well the respondent thinks he/she is informed on a scale from 1-10
$Newsoutlet_i$	News outlets	Respondents report a number from 1-10 indicating how much they rely on news outlets for information
$Occupation_i$	Occupation	The occupation of the individual
$Female_i$	Female	Takes the value one if the individual is female and zero otherwise
$Ethnicity_i$	Ethnicity	The ethnicity of the individual

4. Results

In this section, I will discuss the results obtained by the analysis. First, non-parametric tests are conducted to test the difference in distribution between diverse and non-diverse groups. Subsequently, the analysis will be run over panel data.

4.1 Descriptive statistics

The survey brought a usable sample size of 192 subjects. The division in gender is equally represented, having 50.52% and 49.48% females and males, respectively. Furthermore, table 2 shows the division in occupation between the subjects. It can be seen that students represent the vast majority of the sample followed by full-time employees as the second largest group (13.02%). The remaining occupation represents no more than two per cent, whereas there are no retirees present in the sample. Table 3 shows that the average age is 24 years, with a minimum and maximum of 14 and 60 years, respectively.

Table 2. Descriptive statistics on occupation.

Occupation	Frequency	Per cent
Student	159	82.81
Part-time employee	2	1.04
Full-time employee	25	13.02
Unemployed/Work seeking	3	1.56
Retired	0	0
Other	3	1.56

Table 3. Descriptive statistics on age.

	Observations	Mean	Std. Dev.	Min	Max
Age	192	24.08	4.74	14	60

4.2 Crowd average

The main idea behind crowd wisdom is that the group (the average) performs better than the individual. In this case it would imply that overall, the total absolute standardised error of the crowd (*total* $ASGE_i$), where group *I* is the whole sample, should be smaller than the total absolute standardised error of most individuals (*total* ASE_i). These errors are summed up over the ten questions to get a total number per individual. In other words, the following should hold:

For individual $i \in \{1, \dots, 192\}$

$$total\ ASGE_i < total\ ASE_i$$

Where:

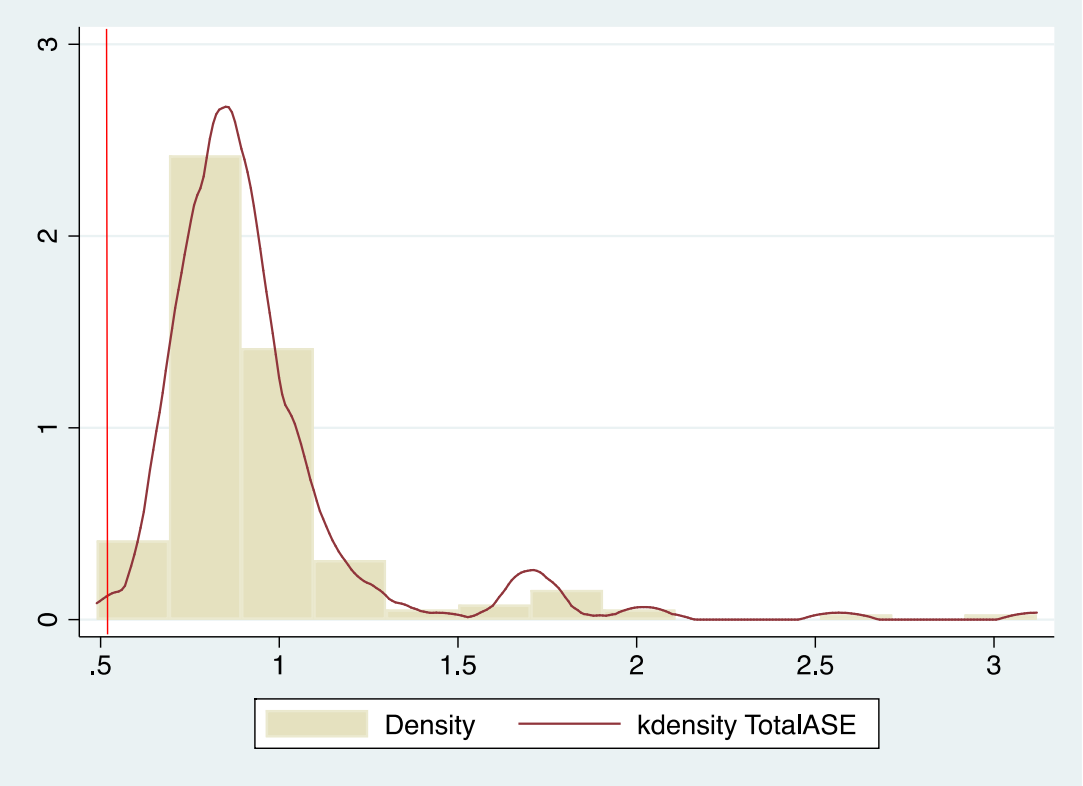
$$total\ ASGE_i = \sum_t^{10} ASGE_{it} \quad (5)$$

and:

$$total\ ASE_i = \sum_t^{10} ASE_{it} \quad (6)$$

The distribution seen in figure 1 illustrates that generally, the total absolute standardised group error is indeed smaller than the individual absolute standardised error, where the vertical red line indicates the total absolute standardised group error. More specifically, 190 out of the 192 individuals perform worse than the crowd average. Thus, the data shows that crowd wisdom is present in this case.

Figure 1. Distribution of total absolute standardised error (ASE)



4.3 Distribution differences

I want to test whether the diversity of a group improves its performance (prediction accuracy). The distribution of the *ASGE* between the diverse and non-diverse groups will be analysed. Statistical procedures that include parametric tests (such as t-tests) have the assumption that the data follows a normal distribution (Ghasemi & Zahediasl, 2012). From figure 1, it can be seen that the asymmetry of the distribution is quite evident. In other words, the distribution of the *ASGE* is skewed to the right. Non-parametric tests do not require the assumption of normal distributions. Therefore, a Mann-Whitney U test is qualified to test such difference between two groups and will be used to determine the difference in the distribution of the average absolute error between diverse and non- diverse constructed groups. In other words, it tests the following:

H_0: *distribution ASGE of diverse groups = distribution ASGE of non-diverse groups*

H_a: *distribution ASGE of diverse groups ≠ distribution ASGE of non-diverse groups*

Table 6. Applying a Mann-Whitney U test to test the null hypothesis that the distribution of absolute standardised error between a diverse and non-diverse group is the same.

	Observations	**Sum of Ranks**	**p-value**
Diverse	60	3655	
Non-diverse	60	3605	
Non diverse = diverse			0.8956

Where diverse groups are the control group and non-diverse groups are the treatment group. The results in table 6 suggest the null hypothesis cannot be rejected, since the p-value is bigger than 0.05 (p=0.8956). This means that there is no significant difference in average absolute error between the diverse groups and the non-diverse groups. Therefore, based on these results, no evidence is found in favour of H1, which stated that diversity has a positive influence on group performance.

4.4 Panel data

The original dataset consists of 192 subjects answering 10 consecutive prediction/guess questions. In theory, basic panel data analysis observes the individual at different points in time. Taking question one until ten as a timing variable, where the individual's prediction is observed after several seconds, the data can be used as panel data. This indicates we have data across seconds and individuals. To test the effect of the confidence range on prediction accuracy, either a fixed- or random-effects model can be run. A Hausman test is performed to see whether a random effects model can be used. The null hypothesis of the Hausman test states that the difference in coefficients between the random- and fixed-effects model is not systematic. If the Hausman test provides a p-value smaller than 0.05, there is a systematic difference in the coefficients and the fixed effects model is preferred. The Hausman test yields a p-value of 0.7165 (see table 7 and Appendix B3). This means the null hypothesis cannot be rejected, so there is no systematic difference between the coefficients of the two models. Thus, the random effects model can be used. Besides, to test H3, a random effects model approach is necessary since *SM_reliance* is individual-specific and fixed over seconds, as well as the demographic variables added to the model. In order to test whether source reliance has an impact on prediction accuracy, *SM_reliance* is added as an independent variable.

One important aspect to note is that only four subjects from the dataset do not display overconfidence. That is, 98 per cent of the entire sample guessed less than eight answers correctly within their provided confidence range. A significant majority of the sample is overconfident in their predictions/guesses, if not all. Therefore, H2 cannot reliably test the effect of overconfidence, since there is no large enough base group to compare it to. Instead, the impact of the size of the confidence range can be tested.

Table 7 shows the results of the models. The evidence suggests that there is a positive significant correlation between the confidence range, *C*, and the *ASE*. The effect is significant at the 1% level across all four models. For example, model four illustrates that a one-point increase in the standardised confidence range will increase the standardised absolute error by 0.344 points. This means that having a larger confidence range will increase the absolute error, implying a decrease in prediction accuracy (performance). In other words, a larger confidence range is correlated with a lower prediction accuracy. This contradicts H2, which states that overconfidence (a small confidence range) has a negative impact on prediction accuracy. The evidence suggests the opposite: as the size of the confidence range increases, prediction accuracy decreases. Therefore, the H2 can be rejected. Additionally, *SM_reliance* remains insignificant across the models. This suggests that (relatively strong) reliance on social media to obtain information about Covid-19 is not correlated with lower prediction accuracy. In other words, people who use social media (quite a lot) as a news source do not provide significantly worse predictions. So, H3 can be rejected as well.

The possibility also exists that social media reliance and the provided confidence range are strongly correlated with each other. For example, it may be that someone who relies strongly on social media is less confident in his/her prediction. If this is the case, then multicollinearity is likely to be a problem, making the regression output unreliable. Firstly, another regression will be run without C_{it} to capture the effect of *SM_reliance*. The results can be seen in model (5) of table 7. *SM_reliance* remains insignificant, indicating that *SM_reliance* and C_{it} are not correlated with each other. This is confirmed by the correlation table in Appendix B4, where it can be seen that there is a very weak correlation (-0.0286) between social media reliance and the confidence range given. Therefore, no multicollinearity is present between these variables.

Table 7. Random effects model on ASE (absolute standardised error). Model 1 displays the impact of the standardised confidence range (independent variable) on the absolute standardised error (dependent variable). Model 2 adds social media reliance, its reliance (SM_reliance) relative to the reliance of news outlets, and news outlet reliance. Furthermore, model 3 adds the demographic variables age, occupation, ethnicity and gender.

Variables	Model (1) random effects	Model (2) fixed effects	Model (3) random effects	Model (4) random effects	Model (5) random effects
C	0.342***	0.332***	0.343***	0.344***	
	(0.0429)	(0.0515)	(0.0430)	(0.0431)	
SM_reliance			0.00768	0.00738	0.00620
			(0.00936)	(0.00951)	(0.00966)
NewsOutlets			0.000466	0.000354	0.00123
			(0.00261)	(0.00268)	(0.00272)
Informed			0.000351	0.000412	-0.000329
			(0.00245)	(0.00251)	(0.00254)
age				-7.72e-05	-4.60e-05
				(0.00101)	(0.00103)
Occupation				0.00158	0.00230
				(0.00497)	(0.00505)
Ethnicity				-8.08e-05	-0.000319
				(0.00361)	(0.00367)
Female				0.00395	0.000669
				(0.00914)	(0.00928)
Constant	0.0824***	0.0827***	0.0736***	0.0721**	0.0827**
	(0.00467)	(0.00495)	(0.0236)	(0.0352)	(0.0358)
R^2	0.0321	0.0321	0.0324	0.0326	0.0004
Observations	1,920	1920	1,920	1,920	1920
Number of ID	192	192	192	192	192
Hausman p-value	0.7165				

Standard errors in parentheses

*Statistical significance: *** p<0.01, ** p<0.05, * p<0.1*

4.5 Time difference

An additional test will be carried out where the prediction accuracy of individuals and groups over a span of a month regarding Covid-19 infections on the 31st of May is tested. The media intense and relevant nature of this topic causes new information and announcements to be broadcasted every day through the news, social media channels and articles. For example, the trend of the virus is communicated on a daily basis, and announcements of additional Covid-19 rules (or withdraw) are regularly updated. Such information may cause subjects to have better knowledge about the situation over time. Someone doing the survey on the 10th of May could have a much better indication of how many infections there will be on the 31st of May than someone who did the survey on the 10th of April, for instance. Therefore, I will use a Mann-Whitney U test in order to see whether there is a significant difference in the distribution of the prediction error between the first two weeks and the last two weeks when the survey was filled in. In other words, it tests the following:

H_0: *distribution* ASE_{it} *in the first half of the period = distribution* ASE_{it} *of the second half of the period*

H_a: *distribution* ASE_{it} *in the first half of the period ≠ distribution* ASE_{it} *of the second half of the period*

Table 8 displays the results of the test. A p-value of 0.3817 means that there is no significant difference in prediction error between the first and second half of the period of survey distribution. Therefore, in this case, time has no impact on prediction accuracy.

Table 8. Applying a Mann-Whitney U test to test the null hypothesis that the distribution of absolute standardised error between the first time period and the last time period is the same.

	Sum of Ranks	Sum of Ranks	p-value
First half (t)	141	13309	
Second half (t)	51	5219	
First half = second half			0.3817

5. Discussion

This section will discuss the obtained results and their implication. In addition, the limitations will be discussed as well.

5.1 The results

Diversity is a well-known requirement for a good performing crowd in the literature. Without it, the difference in perspectives would be at a minimum, which prevents contradicting errors from cancelling each other out. This makes it a common belief that by increasing diversity, one can increase crowd accuracy. However, in this research, I found that diversity does not cause a significant difference in collective group error, where diversity is defined mainly as differences in gender and age. There are several possible explanations why in this case diversity did not turn out to be a significant factor for prediction performance. Firstly, other dimensions of diversity may be more determining than gender or age. For example, Watson et. al., (1993) emphasizes the importance of cultural diversity on the provision of a large range of perspectives. Whereas culture was not included in the study, it could be a more important factor for (individual) group performance than age or gender. Furthermore, as the majority of the sample is represented by

students, it could be that a similar mindset towards the severity of the virus prevents opposing opinions. To illustrate, younger people tend to be less worried about the consequences of the virus compared to elders, which might be reflected in the predictions they give. Secondly, task difficulty may outweigh diversity, putting a constraint on performance despite having individual differences. This idea is in line with Keuschnigg & Ganser (2017), whom found that increasing the task difficulty by one standard deviation decreases the probability of a correct group performance by 9.8%. Furthermore, they found that in the case of high task difficulty without experts, individual ability determines group performance only in large groups. Considering that predicting the exact number of Covid-19 infections in the near future can be labelled as a very hard task, without presenting the question to any "experts", diversity may have less of an impact on group performance than initially hypothesised. Thirdly, it is argued that the effects of diversity are more pronounced in large groups. According to Davis-Strober et. al. (2015), diversity becomes more important for group accuracy as group size increases. It may be the case that groups of 32 individuals are too small to detect the effect of diversity. Perhaps making bigger crowds and decreasing task difficulty could demonstrate the effect of diversity better.

Furthermore, looking at the results, we can say that it contradicts H2. Whereas H2 states that overconfidence has a negative effect on prediction accuracy, the data shows that a larger confidence is correlated with a larger error, significant at the 1% level. This means that providing a larger confidence range, which implies you are less sure about your prediction (hence, less confident), decreases prediction accuracy. Many researchers have found that overconfidence actually decreases prediction accuracy. How is it possible that the results show something else? Note that 98% of the sample already displays overconfident behaviour. Of all the people providing their 80% confidence range, two per cent of the subjects had the correct answers at least eight

times within their given confidence range. So, the results actually imply that an increase in the confidence range of overconfident people, leads to a decrease in the accuracy of their predictions. On the other hand, the formulation of the question may have influenced the subject's behaviour in reporting his/her confidence range. They were asked to state their lower and upper limit, such that they are 80% *confident* their answers lie within these ranges. However, such phrasing maybe a nudge in itself to push the individual to be as confident as possible, perhaps leading to overconfidence. Also, the two per cent not displaying overconfidence may be pure by chance. Still, it seems that prediction accuracy is correlated with the degree in overconfidence. In other words, for overconfident people, whose confidence range is too small since their prediction falls outside of their given range, giving a larger confidence range is actually correlated with a higher error. Therefore, it seems that a higher degree in overconfidence is positively correlated with prediction accuracy. Another explanation for the correlation between large confidence ranges and low prediction accuracy, is that large confidence ranges maybe a reflection of the person's knowledge. It could be that people giving larger confidence ranges are truly less knowledgeable/more clueless instead of being less overconfident, which translates into a larger prediction error. Instead of overconfidence, the confidence range might be a measure of knowledge.

Additionally, no evidence was found in favour of H3. A relative stronger reliance on social media compared to official news outlets is not correlated with higher individual error. One explanation is that social media outlets provide reliable news as well. If many people rely on such news outlets, it maybe because social media also contains a considerable amount of truthful information. Another explanation might be that both official news outlets and social media do not improve the knowledge in predicting the number of infections that will take place in the future.

Someone can follow the official news very closely, but that does not mean it will improve his/her prediction skills. Even experts have difficulty predicting such fickle events.

5.2 Limitations and future recommendations

This research also has its limitations that should be acknowledged. Firstly, the use of a survey research methodology restricts the study to a limited sample size. With a sample of 192, it limits the group size able to construct, in order to test for differences between groups. For example, the importance of diversity is more noticeable in large groups. Therefore, a significant effect of diversity might be more likely in groups of for example 60 people, compared to groups of 32. Furthermore, subjects were not monetarily compensated for filling in the questionnaire. Considering the difficulty and the amount of thought having to put into answering the questionnaire, subjects may not be motivated to think hard enough about the questions, which could lead to insincere answers. Consistently irrational answers have been deleted, however, the possibility that predictions are filled in randomly without genuinely thinking about the question remains. Lastly, the prediction questions may not be entirely independent from each other. Countries have been chosen all over the world and at random, however, it is possible that subjects use the first prediction number as an (irrelevant) anchor to predict the next one. Additionally, Lorenz et. al. (2011) argues that individual answers may not be entirely independent either, since nowadays it is normal for people to be embedded in social media networks, which makes them also more susceptible for bias. Research extending on a similar study should take into account the use of a larger sample size and ensure the motivation of respondents to diminish insincere responses.

The results of this research can potentially be extended in several ways. One possible direction is to identify the best performing individuals and increase their weight to the aggregate number. Likewise, one could decrease the weight of the outliers, which in this case are the least performing individuals. Furthermore, to dive deeper into the effect of diversity, a larger sample size should be gathered taking into account variables like religious and political beliefs as well.

6. Conclusion

The wisdom of the crowd phenomenon has been applied to many topics and fields in the literature. In this study, crowd wisdom was applied to find out whether a group would be able to make a relatively reliable prediction about the Covid-19 infections midst the pandemic. Subsequently, several hypotheses were studied to see what determined (crowd) performance. No correlation was found between diversity and group error. It is likely that other determinants of diversity (besides gender and age) play a more prominent role in group performance. However, I believe the most interesting finding of this research is the positive correlation between the confidence range size and the absolute individual error given by the random effects model. I believe the most convincing explanation for this relation is that in this case, the size of the confidence range not only reflects their (over)confidence, but also their knowledge. Subjects who reported the largest confidence range, where probably the most clueless when providing a prediction.

The research question of this study was how reliable the crowd's prediction is regarding the Covid-19 infections. Looking at the distribution of the total *ASE* (absolute standardised error), it can be seen that the group indeed performs better than the individuals taken separately. Therefore,

the evidence suggests that crowd wisdom is present. Since the crowd performs better than the individual subjects, it implies that the crowd prediction certainly holds value. However, errors compared to the true value are still present, for some questions much larger than others. There is still room for constructing a wiser crowd, for example by testing other sides of diversity, or including only the most “knowledgeable” people.

7. Appendix

A1. Survey design

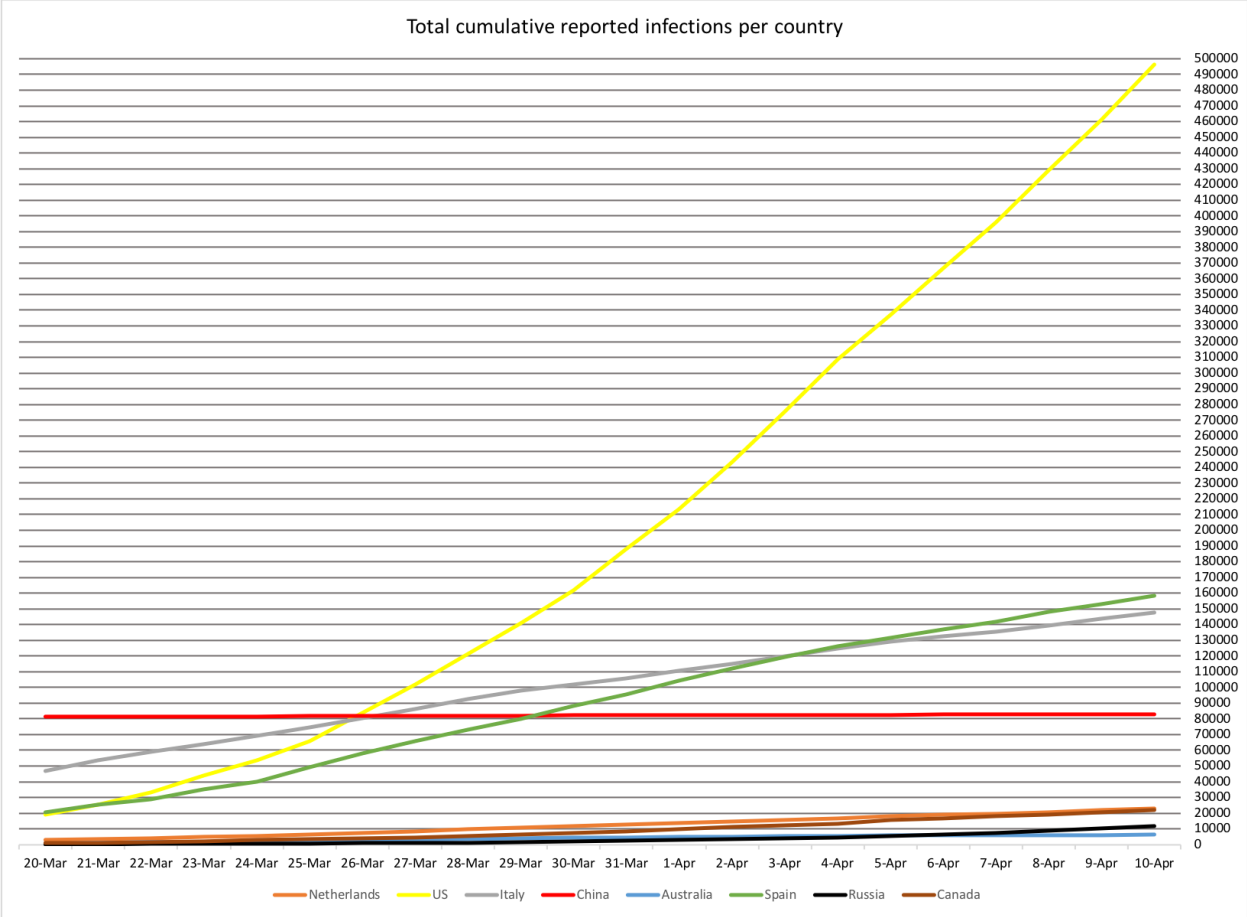
Data will be sampled through a survey. Two different questionnaires will be randomly distributed among respondents. Whereas one starts with information from official news outlets, the other will include news that has been widely distributed through social media outlets like WhatsApp and Facebook. The design is as follows:

Thank you for participating! Your response will be completely anonymous.

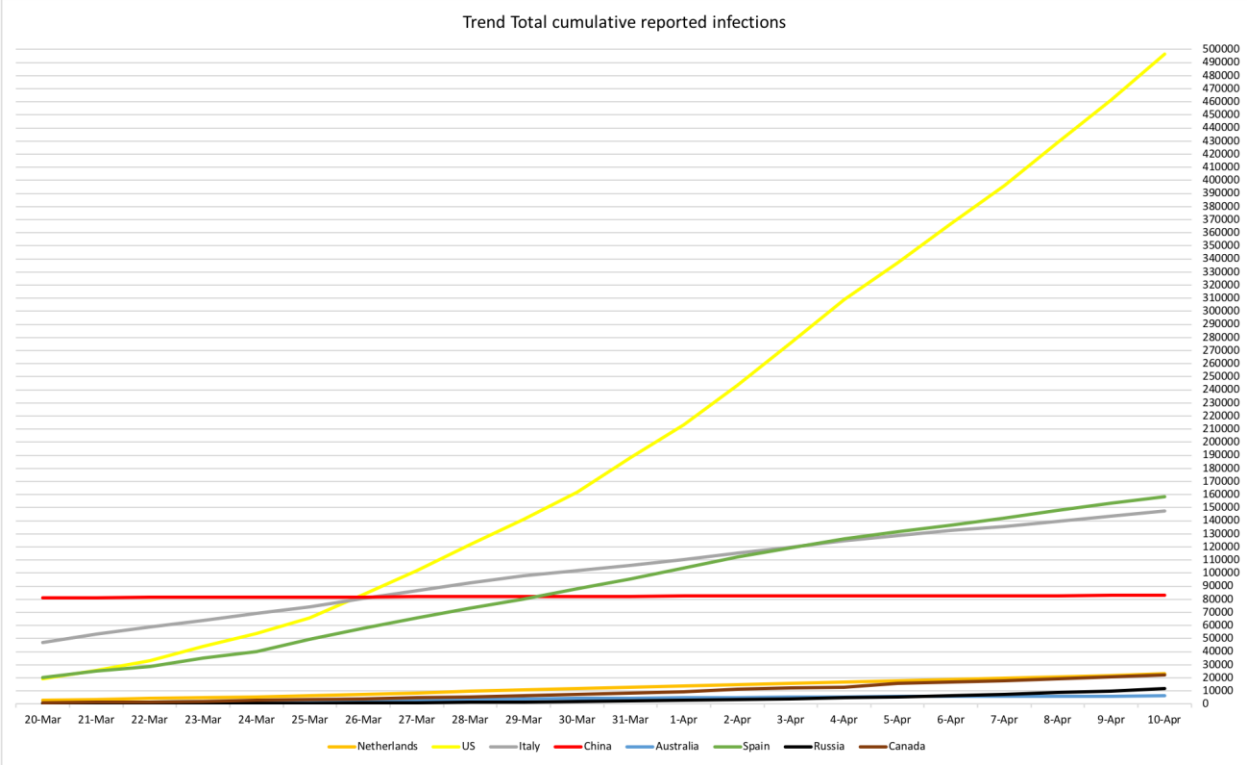
I hope you are doing okay in these strange and extreme times. As you know, the Corona virus is affecting everyone around the world. This survey will ask you to guess the number of total accumulated reported Corona infections in different countries **at the end of May**. So, based on what you have read/seen on the news, internet, social media, or even just your intuition, try to guess what you believe could be the number.

The survey will take around 5 - 10 minutes!

The virus is evolving differently per country. In the table below, you see the trend per country regarding the total reported corona infections from **March 20th** until **April 10th**:


Total cumulative reported infections per country
500000
490000
480000
470000
460000
450000
440000
430000
420000
410000
400000
390000
380000
370000
360000
350000
340000
330000
320000
310000
300000
290000
280000
270000
260000
250000
240000
230000
220000
210000
200000
190000
180000
170000
160000
150000
140000
130000
120000
110000
100000
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
20-Mar
21-Mar
22-Mar
23-Mar
24-Mar
25-Mar
26-Mar
27-Mar
28-Mar
29-Mar
30-Mar
31-Mar
1-Apr
2-Apr
3-Apr
4-Apr
5-Apr
6-Apr
7-Apr
8-Apr
9-Apr
10-Apr
Netherlands
US
Italy
China
Australia
Spain
Russia
Canada

What do you think will be the total (accumulated) number of reported corona infections in the Netherlands at the end of May?

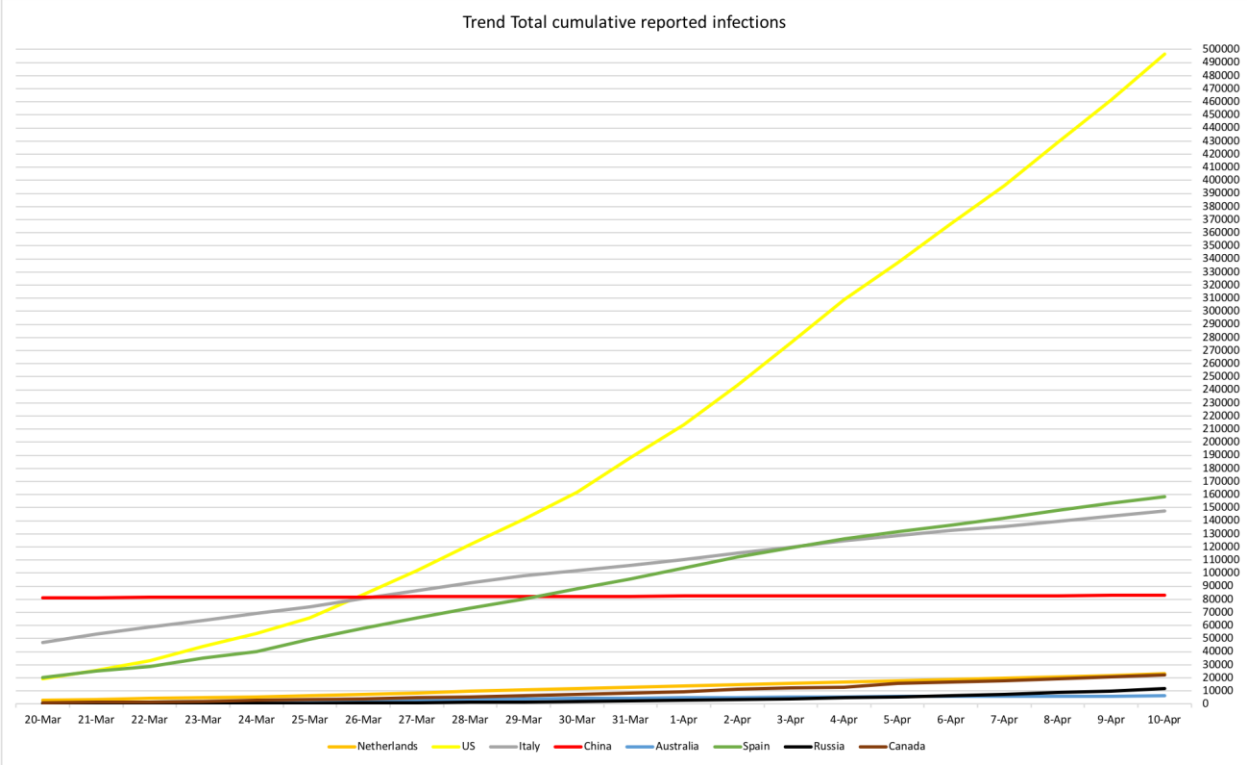


Please provide a lower and upper limit, such that you are 80% sure that the number of total reported infections at the end of May in the Netherlands will lie between these numbers:

Lower limit:

Upper limit:

What do you think will be the total (accumulated) number of reported corona infections in the US at the end of May?

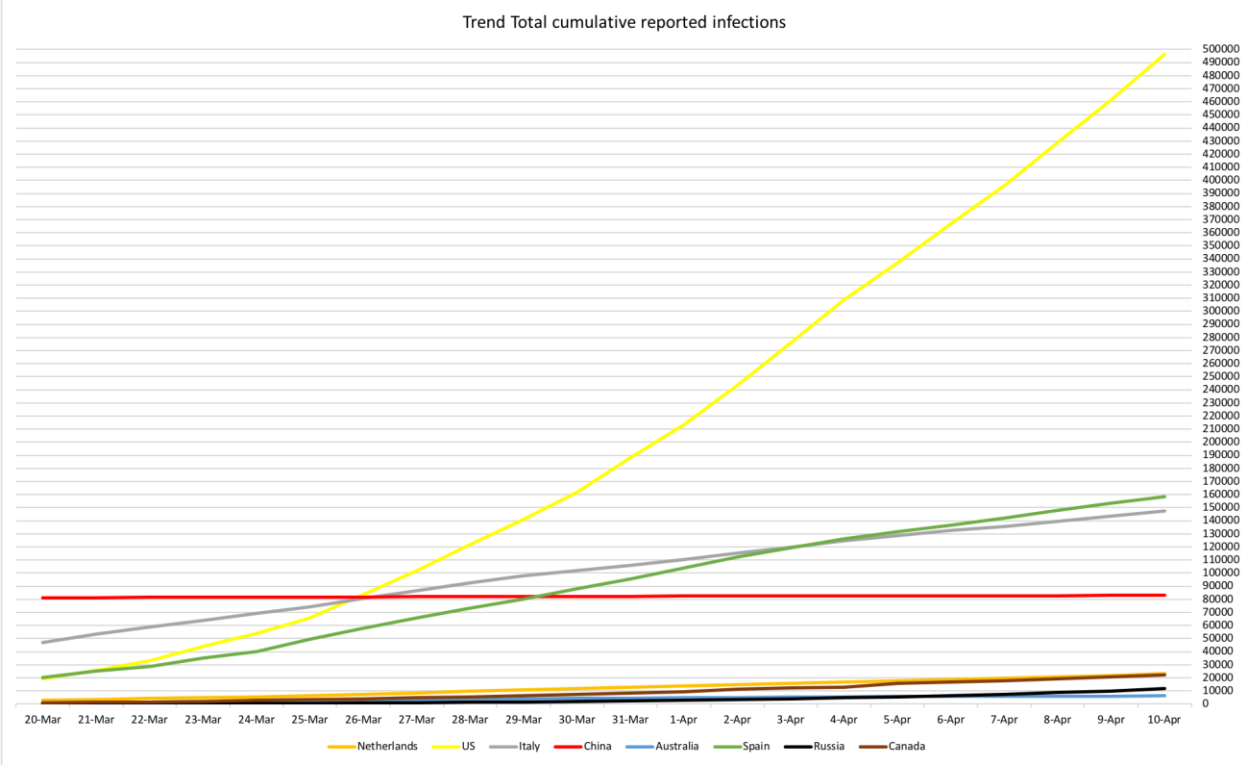


Please provide a lower and upper limit, such that you are 80% sure that the number of total reported infections at the end of May in the US will lie between these numbers:

Lower limit:

Upper limit:

What do you think will be the total (accumulated) number of reported corona infections in Italy at the end of May?

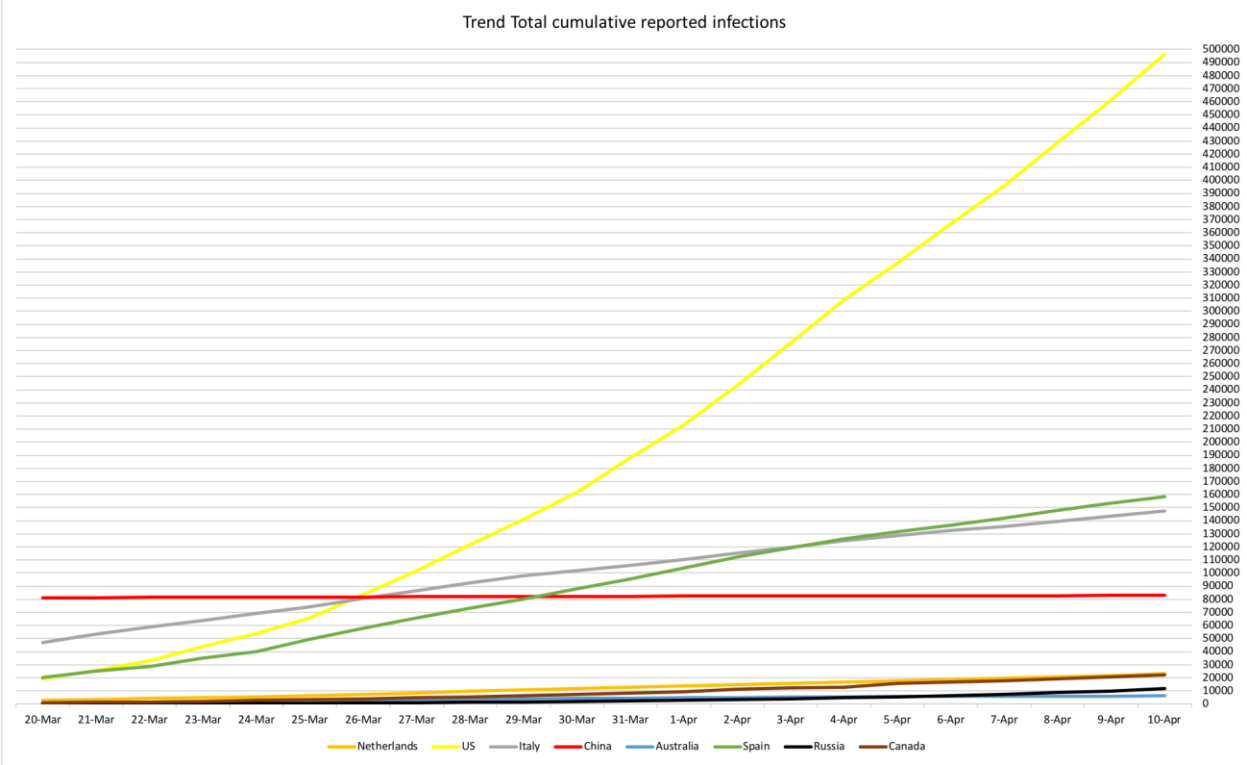


Please provide a lower and upper limit, such that you are 80% sure that the number of total reported infections at the end of May in Italy will lie between these numbers:

Lower limit:

Upper limit:

What do you think will be the total (accumulated) number of reported corona infections in China at the end of May?



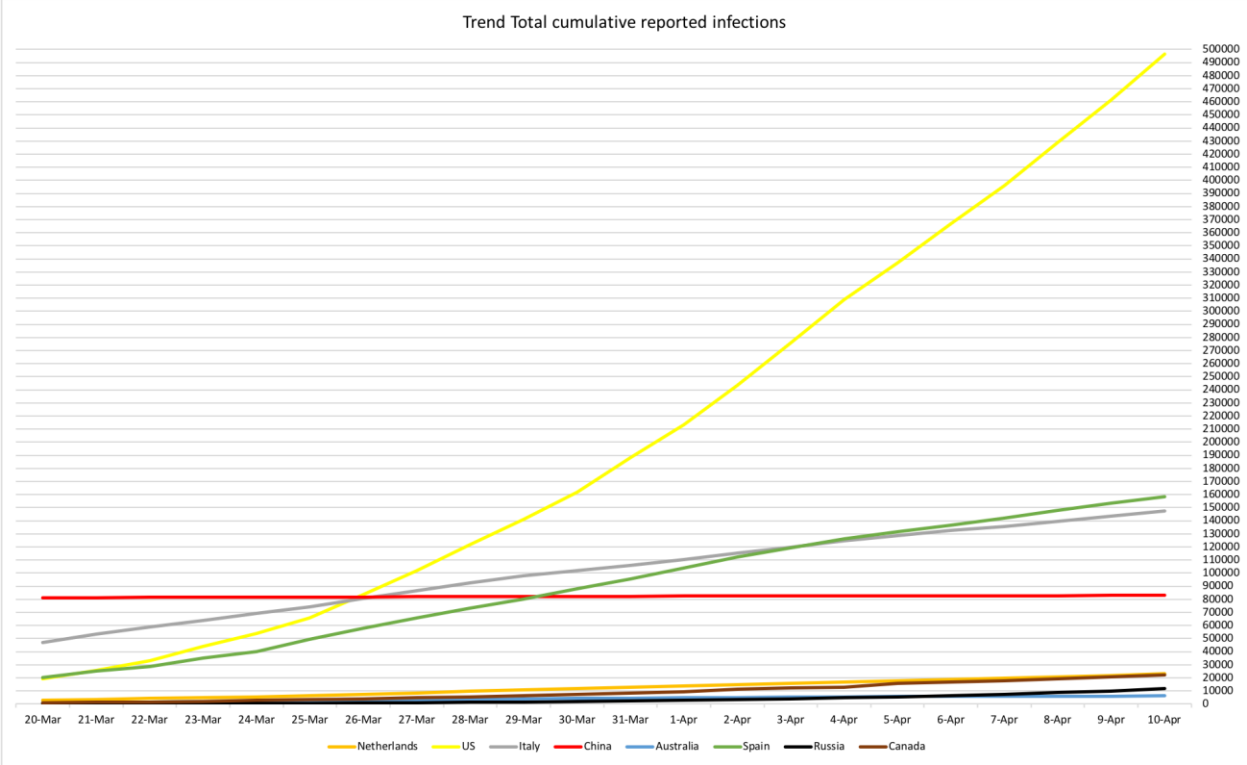
_____

Please provide a lower and upper limit, such that you are 80% sure that the number of total reported infections at the end of May in China will lie between these numbers:

Lower limit:

Upper limit:

What do you think will be the total (accumulated) number of reported corona infections in Australia at the end of May?

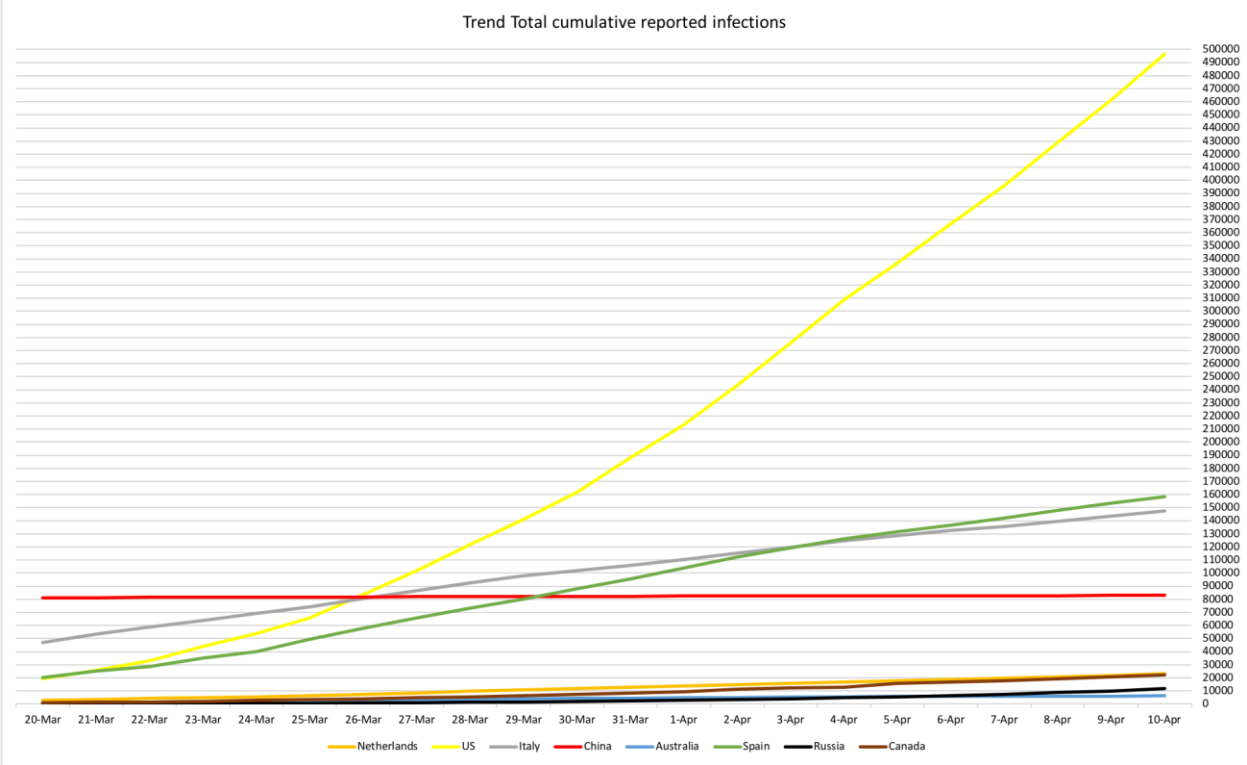


Please provide a lower and upper limit, such that you are 80% sure that the number of total reported infections at the end of May in Australia will lie between these numbers:

Lower limit:

Upper limit:

What do you think will be the total (accumulated) number of reported corona infections in Spain at the end of May?

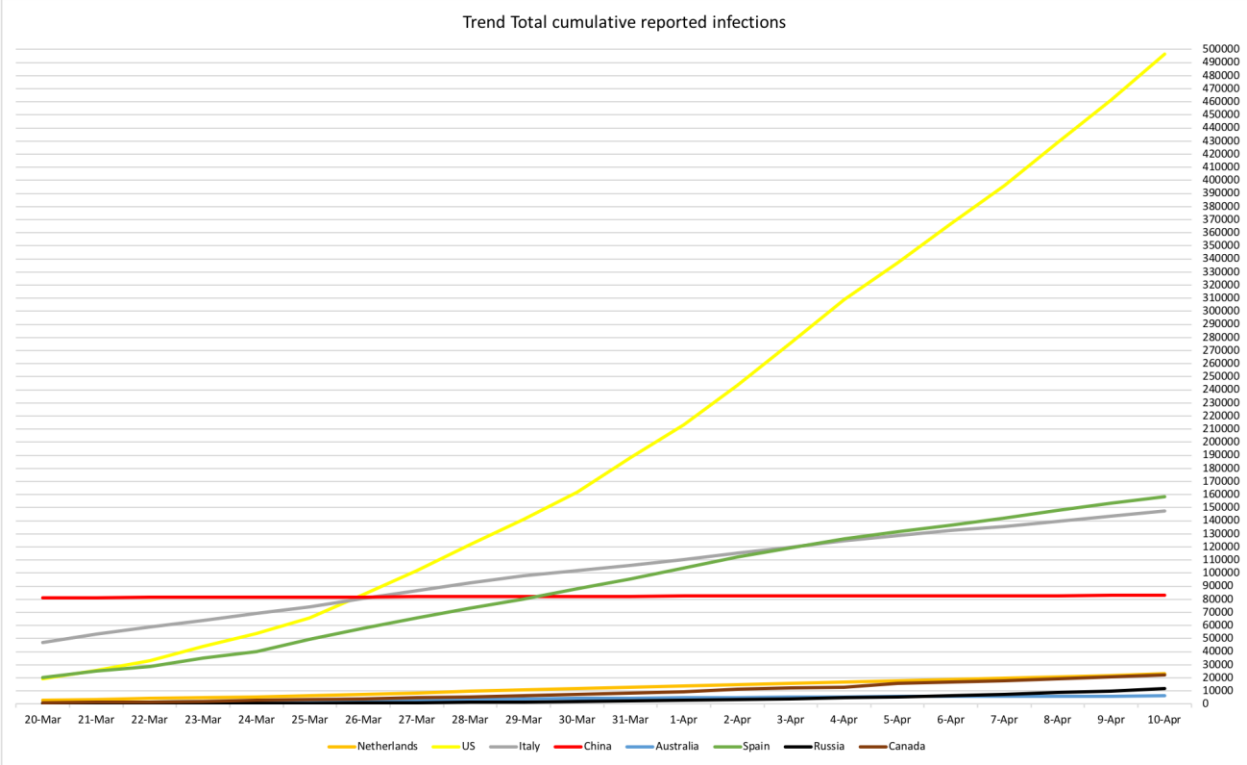


Please provide a lower and upper limit, such that you are 80% sure that the number of total reported infections at the end of May in Spain will lie between these numbers:

Lower limit:

Upper limit:

What do you think will be the total (accumulated) number of reported corona infections in Russia at the end of May?

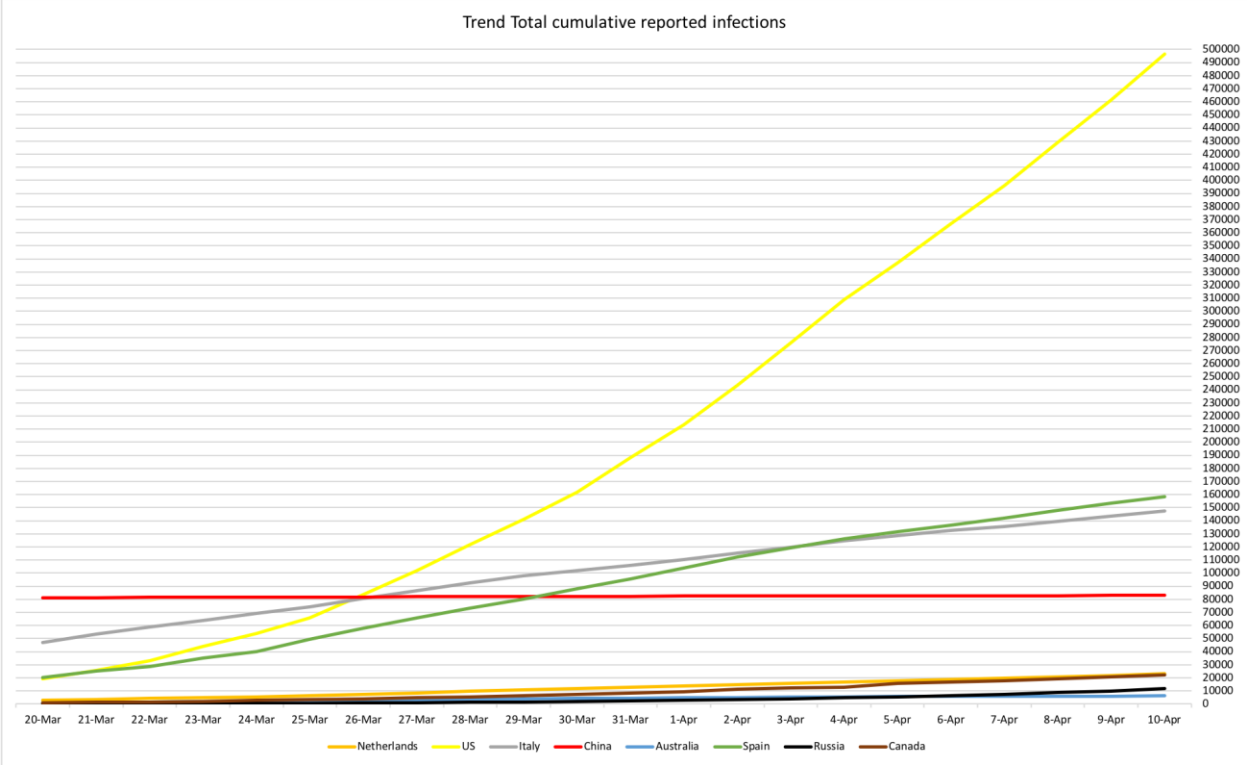


Please provide a lower and upper limit, such that you are 80% sure that the number of total reported infections at the end of May in Russia will lie between these numbers:

Lower limit:

Upper limit:

What do you think will be the total (accumulated) number of reported corona infections in Canada at the end of May?



Please provide a lower and upper limit, such that you are 80% sure that the number of total reported infections at the end of May in Canada will lie between these numbers:

Lower limit:

Upper limit:

What is the maximum number of hours that the coronavirus can survive outside the body on a surface such as plastic?

_____

Please provide a lower and upper limit, such that you are 80% sure that the actual number of hours will lie between these numbers

Lower limit:

Upper limit:

When was the first case of the coronavirus detected in the Netherlands? Please fill in the according date

_____

Please provide a lower and upper limit (date), such that you are 80% sure that the actual date lies between these dates

Lower limit:

Upper limit:

On a scale from 1 to 10 (with 1 being very little and 10 being very much), how much do you rely on information from social media channels (such as Facebook, WhatsApp etc.) regarding the coronavirus?

On a scale from 1 to 10 (with 1 being very little and 10 being very much), how much do you rely on information from institutions and news outlets (such as NOS, RIVM, WHO, newspapers etc) regarding the coronavirus?

My gender is

- *Male*
- *Female*
- *I'd rather not say*

My age is

_____

I am currently

- *A student*
- *A part-time employee*
- *A full-time employee*
- *Unemployed/work seeking*
- *Retired*
- *Other*

My ethnicity is

- *Black*
- *Asian*
- *White*
- *Hispanic*
- *Other*

B1. Example diverse group

date	ID	group_ID	t	country	Q	Female	age	Occupation	Ethnicity
11apr2020	11	1	1	_NL	60000	1	22	1	2
12apr2020	51	1	1	_NL	27000	1	35	2	2
21apr2020	128	1	1	_NL	30000	1	31	3	3
12apr2020	43	1	1	_NL	30000	0	22	1	3
26apr2020	143	1	1	_NL	40000	0	23	1	2
12apr2020	28	1	1	_NL	25009	0	23	1	8
08may2020	183	1	1	_NL	40000	1	24	1	3
11apr2020	4	1	1	_NL	65000	0	25	1	3
12apr2020	46	1	1	_NL	45000	0	19	1	3
11apr2020	18	1	1	_NL	40000	1	21	1	3
14apr2020	78	1	1	_NL	40000	0	60	3	3
16apr2020	101	1	1	_NL	80000	0	24	1	3
15apr2020	95	1	1	_NL	21000	0	21	1	3
13apr2020	54	1	1	_NL	150000	0	33	3	3
12apr2020	35	1	1	_NL	60000	0	21	1	2
12apr2020	42	1	1	_NL	40000	0	26	3	2
15apr2020	93	1	1	_NL	30000	1	53	1	2
11apr2020	16	1	1	_NL	5000	1	23	1	3
30apr2020	165	1	1	_NL	28000	0	25	1	3
12apr2020	49	1	1	_NL	30000	1	25	2	2
18apr2020	116	1	1	_NL	400000	0	21	1	3
09may2020	186	1	1	_NL	22000	0	22	1	3
20apr2020	123	1	1	_NL	34000	0	22	4	2
12apr2020	25	1	1	_NL	50000	1	24	1	3
30apr2020	161	1	1	_NL	35000	0	22	1	3
30apr2020	163	1	1	_NL	25000	0	22	1	2
11apr2020	14	1	1	_NL	40000	1	23	1	2
16apr2020	99	1	1	_NL	75000	0	29	3	3
17apr2020	105	1	1	_NL	35000	1	25	6	3
30apr2020	159	1	1	_NL	33000	0	21	1	3
12apr2020	41	1	1	_NL	20000	0	23	1	3
13may2020	192	1	1	_NL	22500	1	24	1	3

B2. Example non-diverse group

date	ID	group_ID	t	country	Q	Female	age	Occupation	Ethnicity
12apr2020	46	1	1	_NL	45000	0	19	1	3
26apr2020	150	1	1	_NL	20000	0	20	1	1
11apr2020	19	1	1	_NL	5000	0	20	1	2
03may2020	175	1	1	_NL	24000	0	21	1	2
12apr2020	35	1	1	_NL	60000	0	21	1	2
11apr2020	12	1	1	_NL	50000	0	21	1	2
30apr2020	164	1	1	_NL	47000	0	21	1	3
09may2020	187	1	1	_NL	58000	0	21	1	3
30apr2020	162	1	1	_NL	10000	0	21	1	3
18apr2020	112	1	1	_NL	30000	0	21	1	3
28apr2020	153	1	1	_NL	50000	0	21	1	3
23apr2020	135	1	1	_NL	45000	0	21	1	3
06may2020	178	1	1	_NL	20000	0	21	1	3
30apr2020	159	1	1	_NL	33000	0	21	1	3
15apr2020	95	1	1	_NL	21000	0	21	1	3
18apr2020	116	1	1	_NL	400000	0	21	1	3
30apr2020	163	1	1	_NL	25000	0	22	1	2
21apr2020	133	1	1	_NL	40000	0	22	1	2
09may2020	186	1	1	_NL	22000	0	22	1	3
01may2020	172	1	1	_NL	3200	0	22	1	3
11apr2020	6	1	1	_NL	200000	0	22	1	3
12apr2020	43	1	1	_NL	30000	0	22	1	3
14apr2020	76	1	1	_NL	20000	0	22	1	3
13apr2020	66	1	1	_NL	25000	0	22	1	3
30apr2020	161	1	1	_NL	35000	0	22	1	3
08may2020	185	1	1	_NL	30000	0	22	1	3
11apr2020	3	1	1	_NL	1000000	0	22	1	3
11may2020	188	1	1	_NL	22000	0	22	1	3
30apr2020	160	1	1	_NL	26000	0	22	1	3
24apr2020	140	1	1	_NL	60000	0	22	1	3
04may2020	177	1	1	_NL	50000	0	22	1	3
01may2020	171	1	1	_NL	30000	0	22	1	3

B3. Hausman test

```
Fixed-effects (within) regression               Number of obs      =       1,920
Group variable: ID                              Number of groups   =         192

R-sq:                                           Obs per group:
     within  = 0.0235                                         min =          10
     between = 0.3495                                         avg =        10.0
     overall = 0.0321                                         max =          10

                                                F(1,1727)          =       41.49
corr(u_i, Xb)  = 0.0409                         Prob > F           =      0.0000
```

ASE	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
stdCI	.3320308	.0515471	6.44	0.000	.2309295	.4331321
_cons	.0827307	.0049508	16.71	0.000	.0730205	.0924409
sigma_u	.0261459					
sigma_e	.2024251					
rho	.0164094	(fraction of variance due to u_i)				

```
F test that all u_i=0: F(191, 1727) = 0.17                     Prob > F = 1.0000
```

```
Random-effects GLS regression                   Number of obs      =       1,920
Group variable: ID                              Number of groups   =         192

R-sq:                                           Obs per group:
     within  = 0.0235                                         min =          10
     between = 0.3495                                         avg =        10.0
     overall = 0.0321                                         max =          10

                                                Wald chi2(1)       =       63.61
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =      0.0000
```

ASE	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
stdCI	.3423927	.0429303	7.98	0.000	.2582508	.4265345
_cons	.0823729	.0046656	17.66	0.000	.0732284	.0915174
sigma_u	0					
sigma_e	.2024251					
rho	0	(fraction of variance due to u_i)				

	Coefficients (b) fe	(B) re	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
stdCI	**.3320308**	**.3423927**	**-.0103619**	**.0285323**

b = consistent under Ho and Ha; obtained from xtreg
B = inconsistent under Ha, efficient under Ho; obtained from xtreg

Test: Ho: difference in coefficients not systematic

chi2(**1**) = (b-B)'[(V_b-V_B)^(-1)](b-B)
= **0.13**
Prob>chi2 = **0.7165**

B4. Correlation graph

	SM_reliance	stdCI
SM_reliance	**1**	
stdCI	**-0.0286**	**1**

8. Bibliography

Arazy, O., Morgan, W., and Patterson, R. 2006. "Wisdom of the Crowds: Decentralized Knowledge Construction in Wikipedia " in Proceedings of the 16th Annual Workshop on Information Technologies & Systems (WITS).

Bassamboo, A., Cui, R., & Moreno, A. (2015). Wisdom of Crowds in Operations: Forecasting Using Prediction Markets. *Available at SSRN 2679663*.

Cinelli, Matteo, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. "The covid-19 social media infodemic." arXiv preprint arXiv:2003.05004 (2020).

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise?. *Decision*, *1*(2), 79.

Davis-Stober, C. P., Budescu, D. V., Broomell, S. B., & Dana, J. (2015). The composition of optimally wise crowds. *Decision Analysis*, *12*(3), 130-143.

Franch, F. (2013). (Wisdom of the Crowds) 2: 2010 UK election prediction with social media. *Journal of Information Technology & Politics*, *10*(1), 57-71.

Galton, F. (1907). Vox populi.

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism*, *10*(2), 486.

Golub, B., & Jackson, M. O. (2010). Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, *2*(1), 112-49.

Grushka-Cockayne, Y., Jose, V. R. R., & Lichtendahl Jr, K. C. (2017). Ensembles of overfit and overconfident forecasts. *Management Science*, *63*(4), 1110-1130.

He, S. X., & Bond, S. D. (2013). Word-of-mouth and the forecasting of consumption enjoyment. *Journal of Consumer Psychology*, *23*(4), 464-482.

Hilary, G., & Hsu, C. (2011). Endogenous overconfidence in managerial forecasts. *Journal of Accounting and Economics*, *51*(3), 300-313.

Hong, H., Du, Q., Wang, G., Fan, W., & Xu, D. (2016). Crowd wisdom: The impact of opinion diversity and participant independence on crowd performance.

Hua, J., & Shaw, R. (2020). Corona virus (Covid-19)"infodemic" and emerging issues through a data lens: The case of china. International journal of environmental research and public health, 17(7), 2309.

Keuschnigg, M., & Ganser, C. (2017). Crowd wisdom relies on agents' ability in small groups with a voting aggregation rule. *Management Science*, *63*(3), 818-828.

Lehrer, R., Juhl, S., & Gschwend, T. (2019). The wisdom of crowds design for sensitive survey questions. *Electoral Studies*, *57*, 99-109.

Liu, B., & Tan, M. (2019). Overconfidence and forecast accuracy: An experimental investigation on the hard-easy effect.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the national academy of sciences*, *108*(22), 9020-9025.

Nielsen, R. K., Fletcher, R., Newman, N., Brennen, S. J., & Howard, P. N. (2020). Navigating the 'infodemic': How people in six countries access and rate news and information about coronavirus. Reuters Institute.

Polgreen, P. M., Nelson, F. D., Neumann, G. R., & Weinstein, R. A. (2007). Use of prediction markets to forecast infectious disease activity. *Clinical Infectious Diseases*, *44*(2), 272-279.

Russo, J. E., & Schoemaker, P. J. (1992). Managing overconfidence. *Sloan management review*, *33*(2), 7-17.

Rovetta, A., & Bhagavathula, A. S. (2020). Covid-19-related web search behaviors and infodemic attitudes in italy: Infodemiological study. JMIR public health and surveillance, 6(2), e19374.

Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 299.

Sunstein, C. R., & Hastie, R. (2015). *Wiser: Getting beyond groupthink to make groups smarter*. Harvard Business Press.

Surowiecki, J. 2005. The Wisdom of Crowds, New York, PublishedUnited States by Anchor Books.

Treynor, J. L. (1987). Market efficiency and the bean jar experiment. *Financial Analysts Journal*, *43*(3), 50-53.

Turiel, J., & Aste, T. (2020). Wisdom of the crowds in forecasting COVID-19 spreading severity. *arXiv preprint arXiv:2004.04125*.

Watson, W. E., Kumar, K., & Michaelsen, L. K. (1993). Cultural diversity's impact on interaction process and performance: Comparing homogeneous and diverse task groups. *Academy of management journal*, *36*(3), 590-602.

Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. *Journal of economic perspectives*, *18*(2), 107-126.

Zarocostas, J. (2020). How to fight an infodemic. The Lancet, 395(10225), 676.