UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Detection of COVID-19 cases in Belgium using participatory syndromic surveillance data (Infectieradar)

Ilyas Sahli

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

SUPERVISOR :

dr. Lisa HERMANS

2021
2022

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Detection of COVID-19 cases in Belgium using participatory syndromic surveillance data (Infectieradar)

Ilyas Sahli

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

SUPERVISOR :
dr. Lisa HERMANS

HASSELT UNIVERSITY

Abstract

Master of Biostatistics

Detection of Covid19 cases in Belgium using participatory syndromic surveillance data (Infectieradar.be)

by Ilyas SAHLI

Recent risks to public health have sparked significant innovation in infectious disease monitoring, which has led to various syndromic surveillance techniques. Since March 2021, a web-based system called Infectieradar has kept track of cohorts of volunteers who self-report cases of influenza-like illness (ILI) across Belgium. This study explores insights into probable COVID-19 cases and compares them to the official laboratory-confirmed incidence in Belgium using data from the weekly survey of Infectieradar. Methods applied are ARIMA models to prewhiten the time series before inferring from their cross-correlation function and risk factor analysis to determine whether the Infectieradar population's ILI risk is equivalent to the findings of the literature. Results showed a significant correlation between the official laboratory-confirmed incidence and incidence trends from Infectieradar. The results of the risk factor analysis are in line with the literature except for smoking. Infectieradar surveillance platform showed to be a crucial supplemental monitoring method that is more timely, provides data at the individual level, and directly assesses the incidence of ILI in the community.

Acknowledgements

First and foremost, I am indebted to my supervisor Dr.Lisa Hermans for her essential guidance, support, and patience throughout my thesis project. Her vast knowledge and wealth of expertise on this particular topic have aided me substantially throughout my study. Finally, I want to express my heartfelt gratitude to my family for their unwavering love, support, and assistance. Special thanks to my wife for putting up with me being stuck in the office for hours at a time and for offering advice and a sounding board when needed. My parents will always be responsible for providing me with the opportunities and experiences that have shaped who I am. They selflessly inspired me to take risks in life and pursue my own goals.

Contents

Chapter 1

Introduction

1.1 Background

The Covid-19 pandemic raised more concern about monitoring the spread of infectious diseases in the general population. An accurate disease surveillance system is crucial to detect and anticipate abnormalities in real-time. A spike usually follows influenza and Covid-19 outbreaks in hospitalizations and fatalities. In addition to surveillance based on laboratory-verified cases or general practitioners' reports, it is also conceivable to employ symptom surveys for the general population, which may be completed online, administered periodically, and obtained rapidly. Early anticipation of such events can significantly benefit public health officials. It can anticipate shortages in intensive care capacities and inform decision makers to take temporary measures according to the developed situations. Traditional surveillance systems collect clinical and virological data from ILI (Influenza-like illness) patients that visit their physicians. While for Covid-19, it relies particularly on PCR and Antigen tests.

Influenzanet [10] is a monitoring system for ILI in voluntary participants of internet users. It is known as De Grote Griepmeting or the Great Influenza Survey (GIS). Influenzanet is a syndromic monitoring platform that utilizes volunteers and the Internet to track the activity of influenza-like illness (ILI). This cutting-edge monitoring method is based on the voluntarily online input of the public, which answers an online survey concerning flu symptoms on a weekly basis. The first Influenzanet, De Grote Griepmeting, was introduced by the Netherlands and Belgium during the winter of 2003/2004, and it drew over 30,000 participants in its first year. Since then, the Dutch Great Influenza Survey has been conducted annually. For the 2005–2006 season, it was deployed in Portugal; subsequently, in 2008, the Italians adopted Influweb; and finally, in 2012, the French adopted GrippeNet. The survey aimed primarily at assessing the incidence level in the community and making scientific information accessible to the public and students.

Infectieradar.be is part of Influenzanet, a European partnership between various universities and government institutions. This collaboration includes Hasselt University. Influenzanet's goal is to map and track the symptoms of infections in Europe, such as coronavirus (Covid-19) and flu. The information is utilized in scientific studies of illnesses. Via this community participatory surveillance survey, individuals can report signs and symptoms and report whether they seek health care or not. The spread of the ILI is monitored by looking at symptom burden in real-time.

1.2 Objectives and research questions

The main objective of this master thesis is to assess the incidence of Covid-19-ILI in Flanders and to describe insights into the symptom burden of Covid-19 disease and ILI. Furthermore, to investigate the relationship of the incidence trend from Infectieradar with the one from Sciensano's official Covid-19 cases data. The secondary objectives are to explore the survey's participation and representation of the Flemish population. Great focus will be on investigating whether Infectieradar data can be utilized as anticipation of the Covid-19-ILI outbreak in Belgium by correlating the symptom burden over time to the official incidence data for Covid-19 obtained from Sciensano. If both the timing and relative intensities of epidemics from Infectieradar are congruent with those reported by Sciensano, we may be able to establish the Infectieradar system as a reliable sentinel for Covid-19-ILI surveillance.

Chapter 2

Dataset

The Infectieradar survey began on March 29, 2021, and continues today. It is a voluntary survey that is filled in online by all Belgian citizens who are aged 16 years or older (children under 18 years could participate under supervision whether they were under the legal guardian's care or if the legal guardian took action on their behalf) and is accessible in four languages (Dutch, French, German and English). Participants of Infectieradar receive an initial registration form with questions about their background: employment, age, and existing diseases and conditions. Later, a weekly email is sent with a link to symptoms questionnaires, where questions are asked if the participant has had any symptoms in the past week and, if so, which ones: a runny nose, coughing, fever, chills, sore throat, cough, dyspnoea (shortness of breath), nausea, loss of sense of smell/taste, vomiting, diarrhea, stomachache, sneezing, a high temperature, headache, muscle/joint pain, chest pain, and malaise. The questionnaire also covers testing behaviors, hospitalization, immunization uptake, and other associated activities. Although the survey is still ongoing, the data analyzed in this study only lie within 28 Mars 2021 until June 19, 2022.

To define a Covid-19 case based on survey symptoms, we employ the case definition established by Sciensano[3]. The latter was validated for various reasons, particularly for surveillance or as a diagnostic guideline to determine which people should be tested. Next to the case definition from the European Centre for Disease Prevention and Control (ECDC) [4], This definition demonstrates the capacity to estimate concurrent incidence, has the best specificity and sensitivity, and has the highest association with confirmed cases. Accordingly, a possible case of Covid-19 is a person with at least one of the following major symptoms without other obvious cause: acute onset, cough, dyspnea, chest pain, anosmia or dysgeusia, or at least two of the following three minor symptoms with no other obvious cause: fever, muscle aches, fatigue, rhinitis, sore throat, headache, anorexia, watery diarrhea without apparent cause, acute confusion, sudden fall without apparent cause, or worsening of chronic respiratory symptoms (Chronic obstructive pulmonary disease, asthma, chronic cough) without any other obvious cause.

Weighting or sample balancing are used to improve the quality and analytical strength of survey data after it has been gathered. It allows us to produce more representative outcomes for a larger population. The population statistics are provided by StatBel [13]. We Weighted our data according to 3 stratification variables: age, sex, and province. Each unique combination of variables would be isolated to compute the final weights. In this case, we end up with 75 strata. Adjusting for this is relatively straightforward: calculating a multiplier, or weight, for each stratum. This process is called sample balancing, or sometimes "raking" the data. The formula to calculate the weights is $W = T/A$, where "T" represents the "Target" proportion,

"A" represents the "Actual" sample proportions, and "W" is the "Weight" value. A minimum and maximum size of weights are assigned as .5 (a 50 % weighting) and 2.0 (a 200% weighting), respectively. All statistical analyses were carried out using weighted data.

Because of the low participation rate in Brussels and Wallonia regions, we acquired a relatively small dataset, thereby, higher weights will be derived, and the weighting cost will be great in terms of accuracy reduction. Here we stress that weighting is most effective as a tool for making minor improvements; it should not be used to try to save a bad sample design [6]. As a result, we decided to only include data from Flanders.

Another data source that will be used for this project is the official laboratory confirmed Covid-19 cases in Belgium as reported by Sciensano [11] in Flanders, the Belgian health institute. This publicly available data set contains aggregated case numbers for the Belgian population. For Infectieradar data, a ratio is determined by dividing the number of Covid-19-cases defined per week by the number of participants on the same week multiplied by 1000. Thus, we obtain the number of weekly Covid-19 cases per 1000 participants. As for Sciensano data, the ratio is determined by dividing the weekly aggregated Covid-19 cases by the number of residents of Flanders multiplied by 1000. As a result, we obtain the number of weekly Covid-19 cases per 1000 residents.

2.1 Ethical Statement

According to the legislation and GDPR laws, Infectieradar processes personal data legally. The University of Hasselt Medical Ethics Committee and the UZA Ethics Committee [14] have both given their clearance for this project, and an official waiver for ethical approval was obtained. All participants signed informed consent before participating in the survey and can stop participating without feeling obligated to continue by simply not responding to their weekly emails. Anonymity and confidentiality of the participants were guaranteed by removing all identifying information from the study material and report. A participant cannot in any way be linked to their data. The participants of Infecetieradar are the main stakeholders in this study. By volunteering to provide valuable health data, they contribute to assessing the population's circulation and presence of Covid-19. The coronavirus pandemic has dramatically impacted human life worldwide, not just because of its mortality but also due to the strict measures imposed by governments to contain the virus. It presents an unprecedented challenge to public health. This study remains highly societally relevant in this context as it attempts to track and monitor covid-19 incidence. An individual contributes to research on the spread of infectious illnesses, and the novel coronavirus in particular, by providing personal information to Infectieradar.

Chapter 3

Methodology

In this study, data were collected every week for 66 weeks. This type of data is defined as time series data, where a time series is a sequence taken at successive equally spaced points in time [5]. Auto-Regressive Integrated Moving Average (ARIMA) is a statistical method that captures the standard temporal dependencies specific to time series data. This abbreviation is descriptive, capturing the model's major features. They are: AR (p) stands for autoregression: The dependent relationship between an observation and a set of lagged observations. I (d) represents "integrated": To stabilize the time series by differencing raw observations (i.e., subtracting an observation from the preceding time step). MA (q) stands for Moving Average: A moving average model's dependency between observation and residual errors is applied to lagged observations. Each of these components is explicitly specified in the model as a parameter. The general ARIMA model is defined as follows:

$$Y_t = \mu + \sum_{i=1}^{p} \phi_i Y_{t-i} - \sum_{j=1}^{q} \theta_j \epsilon_{t-j}$$

Where Y_t is the data on which the ARIMA model is to be applied, μ represents the intercept, ϕ_i is the coefficient of the autoregressive component for lag i, θ_j stands for the moving average component to be estimated, ϵ_t are the unknown random errors that are assumed to follow a normal distribution. The differencing components are integrated by replacing the Y_t with the required difference.

In this study, we follow Box and Jenkins' [1] approach to prewhiten the data and develop the final model. First, a unit root test will be performed to determine whether the data is stationary; if any, we differentiate the time series to make it stationary and remove the trend. As for the other parameters, we use the Autocorrelation Function (ACF) to set the value for p. The ACF plot is a bar chart representing the correlation coefficients between a time series and its lagged values. It mainly describes how a time series' present value and its previous values are associated. In comparison, the PACF partial autocorrelation function summarizes the correlations for an observation with lag values that are not accounted for by prior lagged observations. The process of choosing the appropriate AR and MA orders should be defined, considering the ACF and PACF plots simultaneously. We anticipate that the ACF plot for the AR process will gradually decline while the PACF should simultaneously experience a sharp decline after 'P' significant lags. The ACF and PACF plots should exhibit the opposite behavior to establish an MA process, i.e., the ACF should show a sudden decrease after a given 'Q' number of lags, whereas the PACF should show a geometric or steady declining trend. On the other hand, both MA and AR are regarded if both the ACF and PACF plots show a trend of steady declining values. The final choice of the appropriate set of parameters will be checked using the Akaike information criterion (AIC) to decide which models fit the data well. Finally, a model diagnosis will be performed by exploring the autocorrelation of the

model residuals. The residuals would leave no temporal structure in the time series for a good fit.

Initially, a graphical comparison between the time series of the incidence trend from Infectieradar and the incidence trend from Sciensano will be made to visualize any common tendencies, if there are any. Next, we will conduct a correlation analysis by obtaining the cross-correlation between the two-time series at different lags. Although both the time series are not long enough to show any seasonality, we expect them to be autocorrelated and share a common trend. Therefore, as the last comparison approach, both time series will be prewhitened using ARIMA models before analyzing their cross-correlation. The latter is a method for objectively evaluating how both time series match up and, particularly, where the best match occurs. It can also make any periodicities in the data visible. Measurements will be carried out using the Pearson correlation coefficient to see how well one time series predicts the values in the other. Then the time series are shifted, and the process is repeated. This means we look at whether there is a greater correlation between the two-time series when a time lag is factored in.

Another approach to validate the data from Infectieradar is to investigate the risk factors based on survey data to see if they are consistent with risk factors associated with Covid19-ILI in literature. Consequently, we adopt a Log-Binomial regression model to evaluate the relationship between numerous covariates and the chance of having Covid19/ILI. The likelihood of a person experiencing Covid19-ILI is modeled as a function of several factors of interest: Age, gender, household situation (with or without children), mode of transportation, chronic conditions (asthma, diabetes), and smoking status. Including the covariates in the model ensures we obtain adjusted risk ratios. Adjusting removes the impact of factor correlations on the ratios and accounts for confounding. The selection of these covariates is based on the literature on similar studies, most of which have been found to be associated with increased ILI risk in earlier research.

In cohort studies, binary outcomes are frequently examined using a logistic regression model to provide odds ratios for comparing groups with various factors. Although sometimes this is appropriate, ILI is an illness that is quite prevalent. Thus, logistic regression techniques are not suited to estimate relative risk since the odds ratio (OR) does not accurately represent the relative risk in this situation [12]. To get directly accurate estimates of the relative risk, we use a log-link instead of the standard logit link, converting our model to a log-Binomial model. Since an individual can report symptoms that match the covid19 case definition multiple times, appropriate methods are required to account for non-Independence in the data. The Generalized Estimating Equations (GEE), proposed by Liang and Zeger Liang [16], are a parameter estimation method for correlated data. It is a robust alternative to maximum likelihood estimation in mixed models. GEE is a modified version of the maximum likelihood that adjusts for the additional variability in the variance structure. The covariance matrix of the GEE is robustified due to the so-called sandwich covariance. This yields parameter estimates consistent even if the correlation structure is misspecified. The GEE aims to estimate the average response over the population ("population-averaged" effects). The chosen regression model is formulated as follows:

$$log_(\pi) = \beta_0 + \sum_{k=1}^{p} \beta_k$$

Where π stands for the probability of success, the parameter β_0 denotes the fixed intercept. The parameters β_k denote the coefficients of the rest of the covariates.

3.1 Related Work

A lot of similar studies have been found in the literature; most of them relate Internet-based monitoring to confirmed Covid19-ILI cases and investigate whether self-reported platforms can be reliable as a valid anticipation method for the spread of Covid19-ILI. Vandendijck et al. [15] aimed to examine The Grote Corona Survey population's representativeness and evaluate the survey's reliability in terms of ILI incidence. Throughout eight influenza seasons, the researchers filtered the GIS incidence time series using a Random Walk model of first order and compared its fitted values against two other monitoring systems: The Belgian Sentinel Network and The Google Flu Trends. Moreover, they also performed a risk factor analysis to see if the risks of contracting ILI in the GIS population are comparable to the findings in the literature. Ellsiepen [4] utilized the same data source in a master thesis project with a similar study with more focus on the comparison of multiple Covid19 case definitions, including but not limited to WHO (World Health Organization) and ECDC (European Centre for Disease Prevention and Control), and to determine which case definition is the most optimal according to sensitivity and specificity. Van Noort et al. [9] conducted an analogous study where incidence from Influenzanet was contrasted against ILI data from the European Influenza Surveillance Network in various European countries. Coherence was investigated and validated through risk factor analysis and cross-correlation before and after prewhitening with ARIMA models. Unlike similar studies, both compared time series were prewhitened with separate models instead of one for both. All studies concluded that the voluntary online platforms represent valuable additional surveillance networks for ILI monitoring.

Chapter 4

Results

4.1 Descriptive statistics

Since the implementation of the Infectieradar platform on 29 Mars 2021 until 19 June 2022, 37566 weekly reports were submitted from 1747 individuals. Of these, 58 (3.21%) did not return any weekly surveys, and 1689 (96.79%) responded to at least one weekly email and filled in the symptom questionnaire. 1550 (91.04%) individuals had returned two or more weekly surveys. At the same time, the individuals that participated at least three times account for 85.87% of the individuals. Overall, the mean number of reports per week is equal to 542. The proportion of individuals that participated only once is 7.96%. Meanwhile, 3463 (9.22%) weekly reports had Covid19 symptoms, and 26% of those performed a Covid19 test. 35 (2%) participants had Covid19 symptoms on their first weekly survey. 60.76% of the participants were female, and 43.45% were male. The highest participating age categories were 60-69 and 50-59, with 30.56% and 20.71% of the total participants, respectively. As for the geographical region, the province of Antwerp comes first with 43.68%, and West-Vlaanderen is the province with the fewest participants (see Figures 4.1 and 4.2).

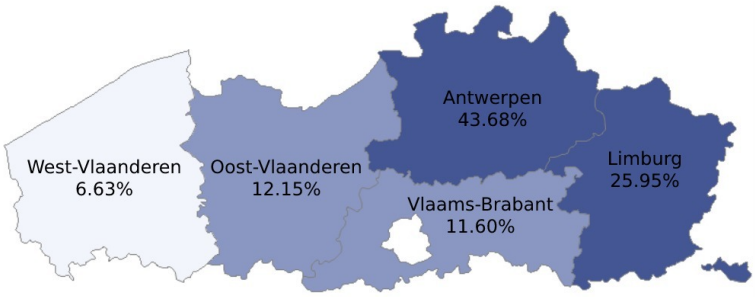

FIGURE 4.1: Distribution of participants according to region

As for the number of participations per week in Figure 4.5, it was susceptible to many variations due to various reasons: a data loss occurred during the last week of august 2021, and multiple advertising campaigns in different periods were put in place. The first explains the steep downhill in participation line in late august, while the second justifies the sudden inclinations observed in the plot. We notice that the Infectieradar survey is reaching more people as time passes.

In order to get insights into the representativeness of the Infectieradar sample, we compare the number of individuals from each stratum in age, gender, and province with the number of residents in Flanders for the same stratum. Figures 4.3 and 4.4

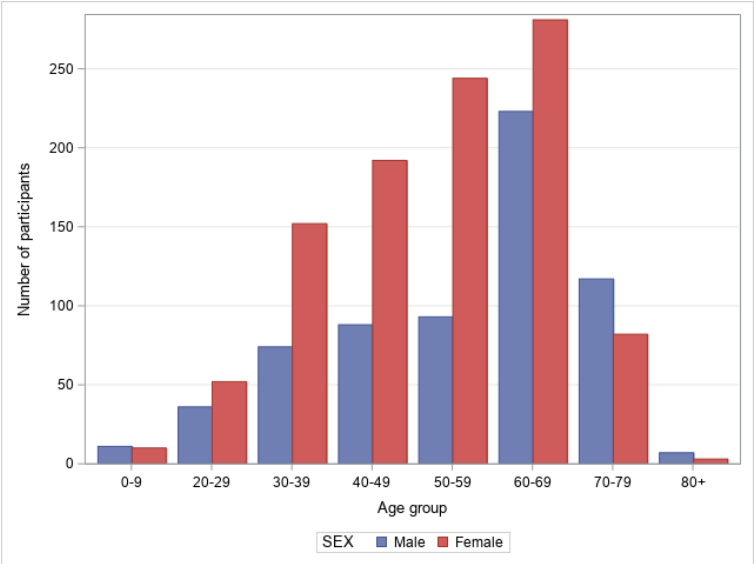

FIGURE 4.2: Distribution of participants according to age

demonstrate that Infectieradar is not representative regarding age and geographical region. Individuals from Antwerp and Limburg are highly overrepresented. In contrast, there is quite an underrepresentation of other regions. On the other hand, age groups 50-59 and 60-69 are over-represented. Concurrently, the 80+ age group is highly under-represented. As for gender, females are over-represented in the online surveillance platform as they stand for 60.45% of participants. In parallel, this statistic is only equal to 51.24% in the Flemish population.

Characteristic	N	Percentage
participants	1747	
weekly reports	37566	
mean number of reports/week	542	
participated once	130	7.4
participated ≥ 2	1550	88.72
participated ≥ 3	1463	83.74
Reported 0 COVID-19-like illness	713	40.81
Reported 1 COVID-19-like illness	338	19.34
Reported ≥ 2 COVID-19-like illness	697	39.89

TABLE 4.1: Overall descriptive statistics of the study participants

Data from the first symptom questionnaire is eliminated. Only data from participants who completed at least three weekly symptom questionnaires is included to decrease the effect of volunteers who only participated occasionally and those who took part only when they had symptoms. Figure 4.6 shows the time series of the incidence trends from both the full and filtered data; we notice that the series are mostly overlapping, indicating that most participants did not have Covid19 symptoms on their first questionnaire.

The statistical comparison between Infectieradar data and Sciensano data is executed using the full data. Identical results were obtained when using the filtered

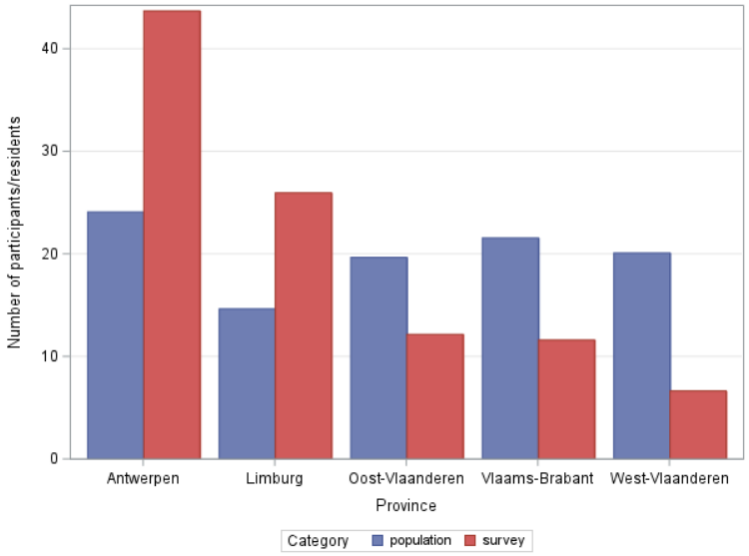

FIGURE 4.3: Representativeness of participants according to region

data. Figure 4.7 plots the weekly Covid19 cases based on symptoms from Infectieradar data against weekly Sciensano confirmed cases. For both time series, there is no clear, consistent trend (upward or downward) over the entire period. The incidence trend from Infectieradar has more variation, and its variance does not appear constant throughout the series. Meanwhile, the incidence trend from Sciensano is relatively less noisy, especially during the period from Mars 2021 until September 2021.

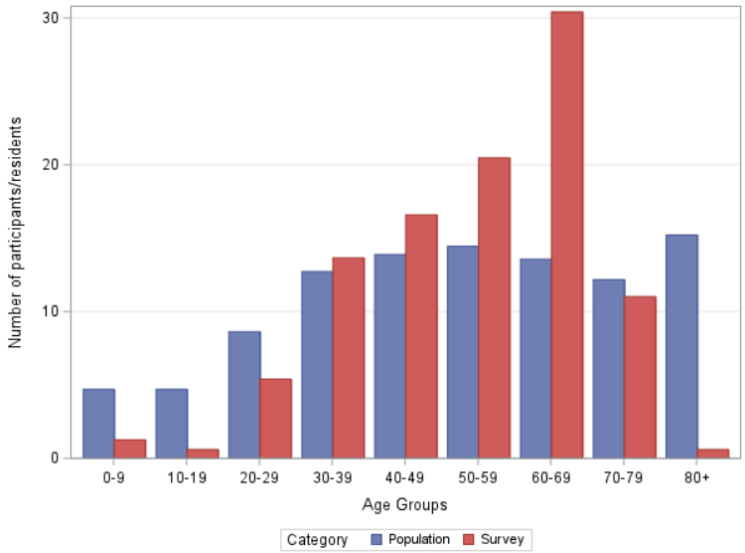

FIGURE 4.4: Representativeness of participants according to age

4.2 Time Series Analysis

As a first comparison, we calculate the raw correlation between the Infectieradar and Sciensano time series on multiple lags. We transform the Sciensano time series using a log scale to meet the normality assumption. Figure 4.8 shows the cross-correlation function, indicating a significant correlation on multiple lags.

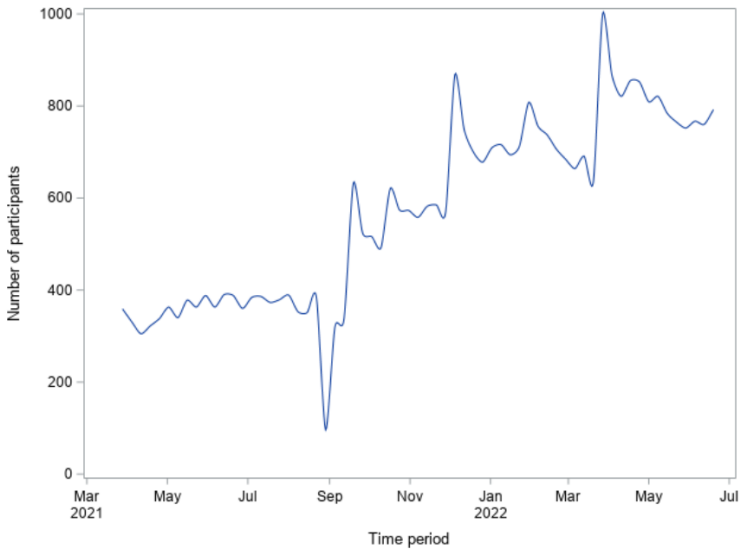

FIGURE 4.5: Participation to the survey over time

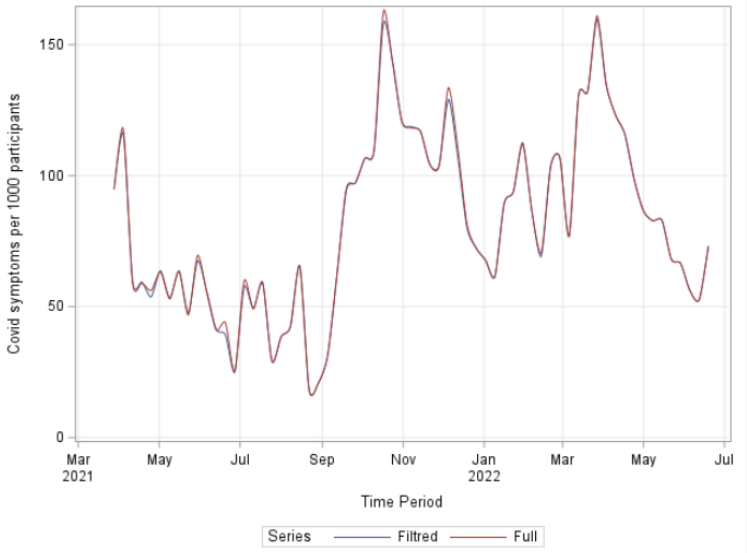

FIGURE 4.6: Graphical Display of the Filtered data against the full data

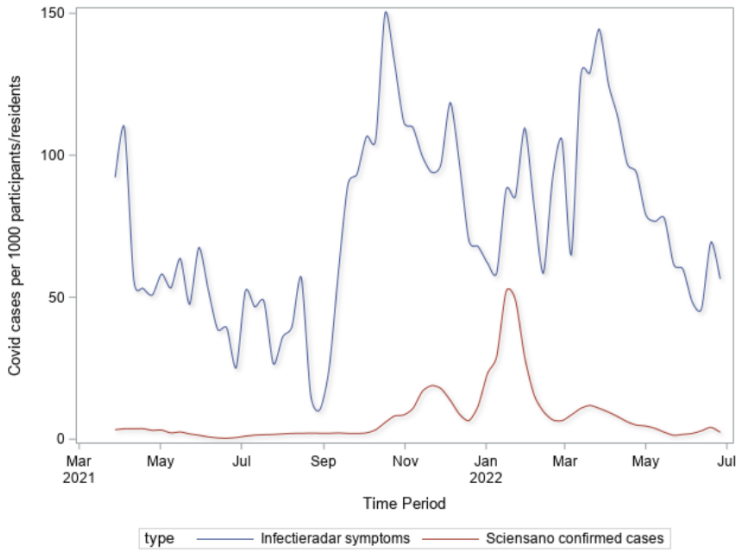

FIGURE 4.7: Time series of Infectieradar data in versus Time series of Sciensano data

Although we could not observe a clear trend from the graphical inspection, we performed formal tests to check if the time series were stationary or not. For that purpose, we used Box-Ljung test [2] and Augmented Dickey-Fuller [8] test. The ADF test examines the null hypothesis of a unit root of a univariate time series, which is equivalent to a non-stationary time series. In contrast, the Box-Ljung test checks the null hypothesis of independence in the time series, i.e., the time series is stationary. Results are shown in table 4.2. The Ljung–Box test rejected H_0 on both time series at all lags from 1 to 7, and the ADF tests failed to reject the null hypothesis on all lags for both time series. The acquired results confirm the non-stationarity for both time series. Therefore, We find it required to stabilize the series by differencing it once.

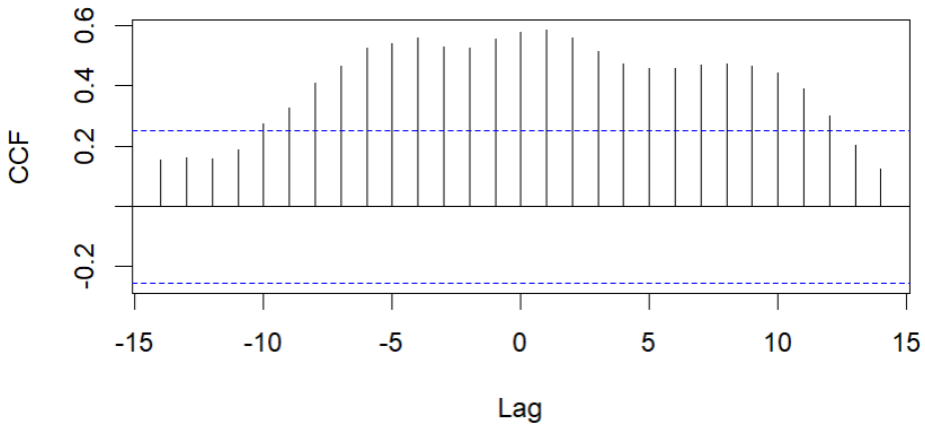

FIGURE 4.8: Cross-Correlation plot of raw series

While ACF and PACF do not directly determine the ARIMA model's order, an inspection of these plots can help understand the order of the MA and AR components

Time Series	Test	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6	lag7
Infectieradar	Ljung–Box	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
	ADF	0.2779	0.3001	0.4807	0.2526	0.5748	0.4382	0.3199
Sciensano	Ljung–Box	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
	ADF	0.4617	0.7947	0.9331	0.9717	0.9623	0.9606	0.9703

TABLE 4.2: Statistical tests for stationarity for raw time series and the corresponding P-values

and provide an idea of which model might match the time-series data well. For this end, ACF and PACF plots were obtained to depict the temporal dependency pattern in the Infectieradar series and to identify the orders of AR and MA terms in the ARIMA model (Figures 4.9 and 4.10). The PACF has an evident spike at lag one, and the ACF is slowly decaying, implying a model without an MA component and an AR component on lag 1.

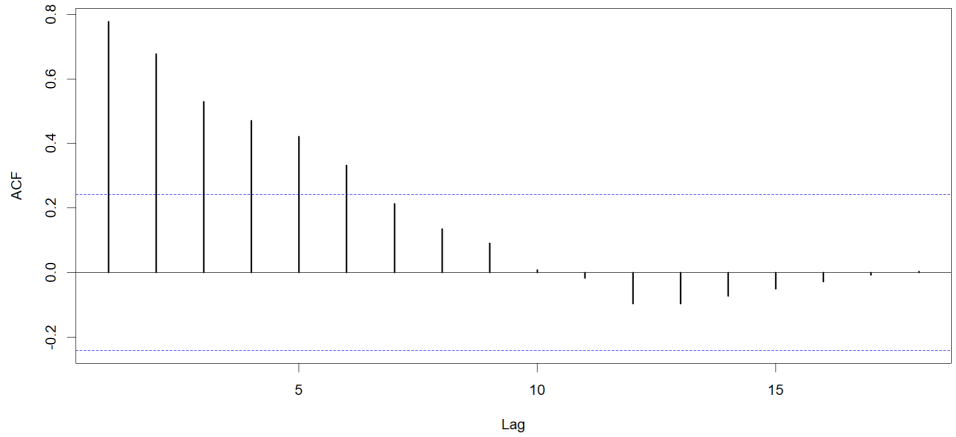

FIGURE 4.9: Autocorrelation Function of Infectieradar time series

The chosen model is ARIMA (1,1,0). In Figure 4.12, we plot a histogram of the model's residuals (Figure 4.12). We can observe that the histogram has a normal distribution with a mean of 0. Additionally, the ACF plot for residuals (Figure 4.11) shows no serial autocorrelation, which indicates that the model is a good fit.

We apply the same approach to prewhiten and detrend the time series of the incidence trend from Sciensano. Statistical tests showed the presence of autocorrelation and non-stationarity (Table 4.2). PACF displays a sharp cut-off at lag 2 (Figure 4.14), whereas the ACF gradually decreases over time (Figure 4.13). Those properties suggested ARIMA (2,1,0) (AIC 9.54) and ARIMA (2,1,1) (AIC 9.66). We chose ARIMA (2,1,0) as the final model. The histogram of the residuals (Figure 4.1) showed normality, and it was validated using the Shapiro test (P-value=0.74). Moreover, non-autocorrelation of the residuals was achieved (Figure 4.15), demonstrating that the suggested model fits the time series well.

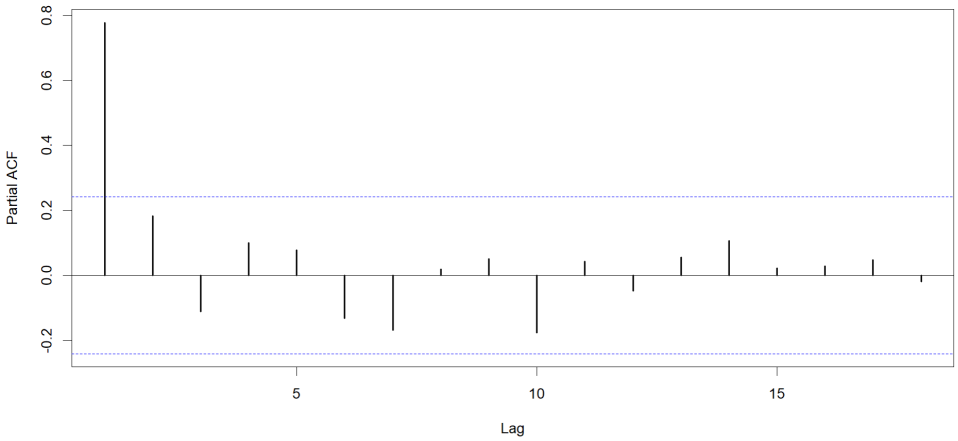

FIGURE 4.10: Partial Autocorelation Function of Ifectieradar time series

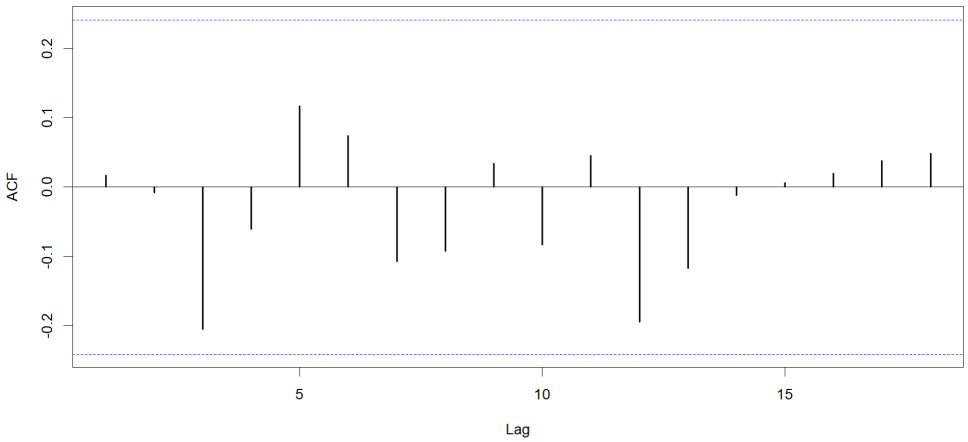

FIGURE 4.11: Autocorelation function of the residuals of the model from Infectieradar series

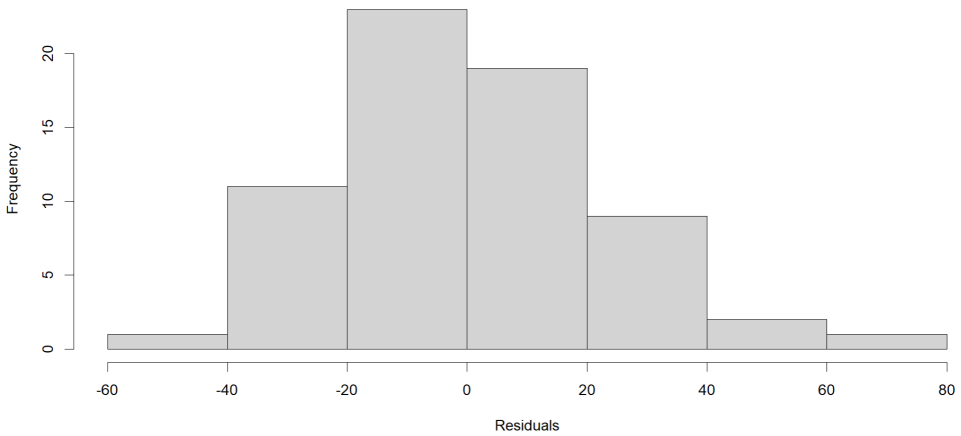

FIGURE 4.12: Histogram of the residuals of the model from Infectieradar series

Figure 4.17 displays the cross-correlation function between the incidence trend from Infectieradar against the one from Sciensano. The sign of the lag defines which series is displaced, and the lag describes how far the series is offset. The lag value with the greatest correlation coefficient indicates the two series' ideal match. The period that one series leads or lags the other is calculated by multiplying the lag by the sampling interval (1 week). We find a significant cross-correlation occurring at lag -4 with a value of 0.306. A significant correlation at a negative lag value is a correlation between Infectieradar data at a time before t and Sciensano data at time t. In other words, a higher Covid19 incidence from Infectieradar will likely lead to a higher Covid19 incidence from Sciensano four weeks later.

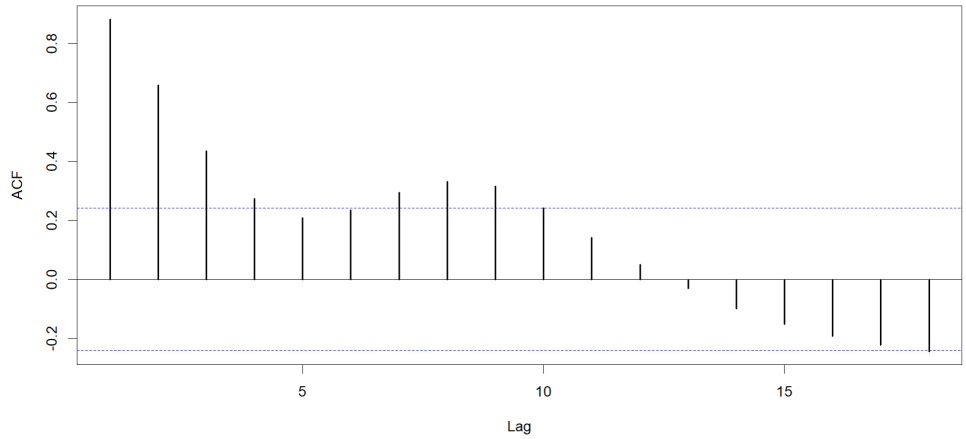

FIGURE 4.13: Autocorelation function of Sciensano time series

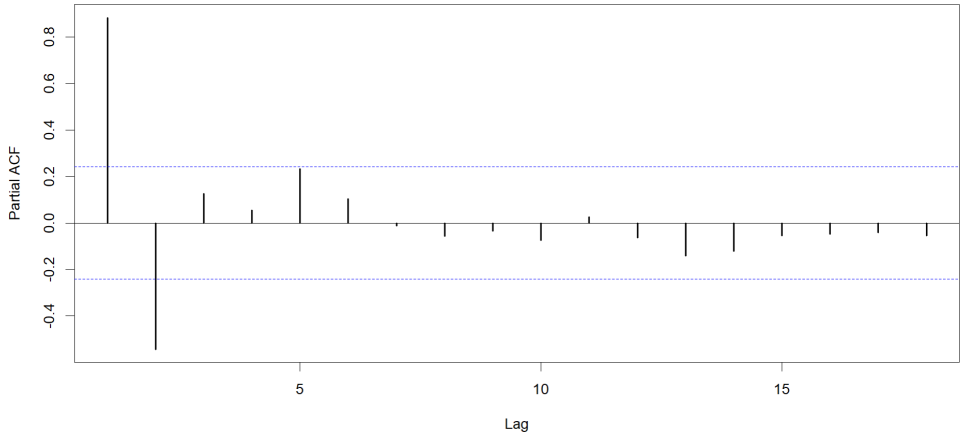

FIGURE 4.14: Partial autocorelation function of Sciensano time series

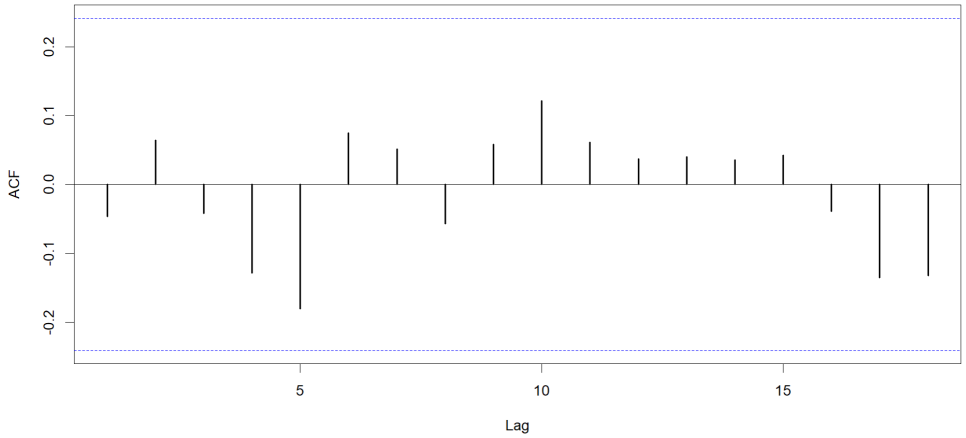

FIGURE 4.15: Autocorelation function of the residuals of the model from Sciensano series

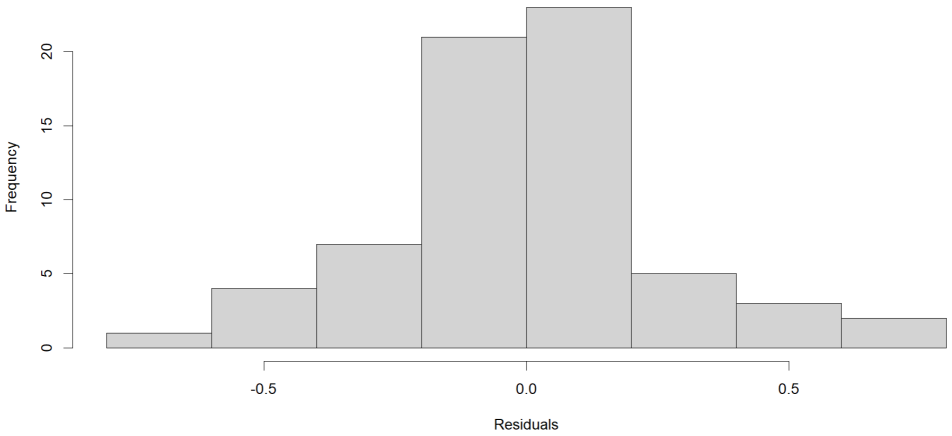

FIGURE 4.16: Histogram of the residuals of the model from Sciensano series

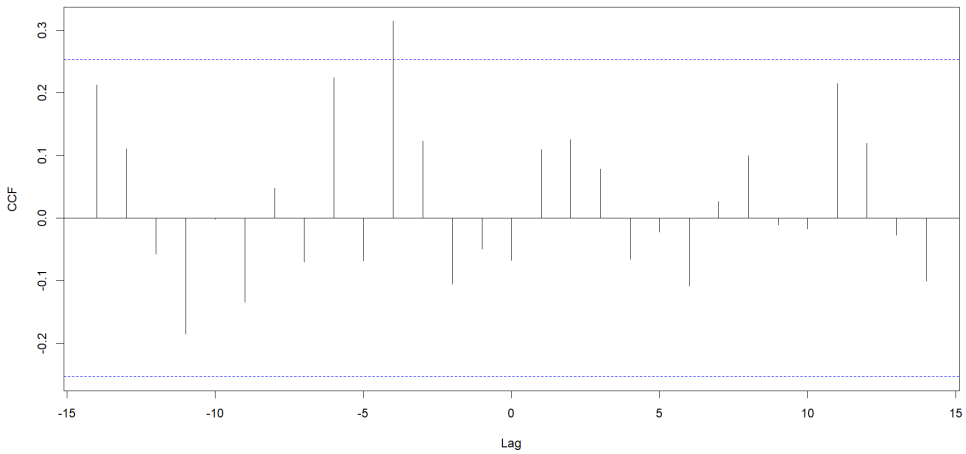

FIGURE 4.17: Cross-Correlation plot of the residuals from Infectieradar against Sciensano series

4.3 Risk factor analysis

The model was built using SAS software with the GENMOD procedure. We implement an exchangeable correlation structure to correct standard errors for repeated observations per participant. The estimated working correlation within a cluster of observations coming from the same participant was equal to 0.08. Parameter estimates for log-binomial regression are found in table 4.3. The reference categories for Gender, Age, and Method of Transportation are Male, 20-60, and Automobile, respectively. According to our model, being a female, belonging to a younger age category, suffering from a chronic disease, and living with children are linked with increased risk for having Covid19-ILI. Females have a higher risk of having Covid19-ILI by 22% more than men. While living with children causes a higher risk of 34% than living with no children. Individuals younger than 20 years old are the highest risk category with 105% more chance to have Covid19 than the reference group 20-60. Meanwhile, individuals older than 60 have a 39% less chance of having Covid19 than individuals within the 20-60 age category. People with Chronic diseases have 40% more chances of getting Covid19. On the other hand, smoking and the transportation method did not significantly affect the likelihood of having Covid19/ILI.

Variable	Category	Adjusted RR [95%]
Gender	Male	—
	Female	1.22 [1.14-1.32]
Age	< 20	2.05 [1.66-2.50]
	$20 - 60$	—
	> 60	0.61 [0.56-0.67]
Smoking	Smoker	1.03 [0.91-1.16]
	Non-smoker	—
Household Status	with children	1.34 [1.23-1.46]
	Without children	—
Method of transportation	Public Transportation	0.96 [0.82-1.12]
	Automobile	—
	walking or Biking	0.94 [0.87-1.02]
Chronic Disease	Yes	1.40 [1.28-1.53]
	No	—

TABLE 4.3: Parameter estimates of the log-binomial regression model for risk factor analysis

Chapter 5

Discussion and Conclusions

5.1 Discussion

The comparison between raw time series showed multiple significant correlations on different lags. However, those results can be misleading since both time series are autocorrelated and share a common seasonal trend. After prewhitening the time series using the best-fitted model, the cross-correlation function only shows a significant correlation on lag 4. This result is more reliable since both time series were detrended and autocorrelation was disregarded. This proclaims that the projected Covid9/ILI incidence trends based on the Infectiradar and trends from Sciensano data obtained from ARIMA models are shown to be well correlated, irrespective of the unrepresentativeness in terms of age, gender, and province. Reveals that reliable incidence trends may be generated using the InfectieRadar surveillance platform before precisely four weeks. Alternatively, the Infectieradar data is leading the Sciensano data by four weeks. Finding an intuitive explanation for the association that existed precisely four weeks before is still difficult. One view of the situation is that the prevalence of Covid is rising or falling in the general population, and it takes four weeks for such patterns to show up on the official Sciensano Covid testing base assuming the Infectieredar sample is representative of the population in Flanders which is hardly the case for this study despite weights adjustments. Although the time frame of the series is relatively short and longer series would be preferable for a more certain conclusion. Besides, the level of participation in Flanders is still low and might be seen as insufficient to provide outcomes with less noise. The Netherlands, for instance, has a far higher participation rate [7]. Appropriate steps should be taken to increase public access to the Infectieradar platform and encourage participation from all demographic groups.

Influenza literature and risk variables derived from the Infectieradar cohort are reconcilable except for smoking. According to our findings. Participants over 60 had a considerably lower incidence of ILI. This age group may also have been stricter about avoiding settings that may lead to transmission. A higher risk of ILI was seen in children and those who live with children; this may be due to children having contact with other children in schools regularly. It has also been known that women are more likely than males to contract ILI, which may be because of more frequent interaction between mothers and children. In line with observational research, serious illness is more likely to occur in patients with underlying medical conditions such as cancer, diabetes, chronic lung disease, and cardiovascular disease, which was validated according to our findings. Literature suggests that smoking is strongly associated with a higher risk of Covid19/ILI. Whereas smoking failed to show a significant effect according to the model, that might be explained by the small sample size of the Infectieradar participants. It is worth emphasizing that contact with an infected

person is the main risk factor for contracting an illness, which is not considered in this study.

5.2 Conclusion

For the period from March 2021 to June 2022, the Infectieradar syndromic surveillance system has shown to be useful for tracking the incidence of symptoms linked to COVID/ILI infection at the national level and for identifying associations between self-reported COVID-19-like illness occurrence and demographic variables, pre-existing health conditions, and other factors. Emphasis should be placed since all syndromic surveillance platforms rely heavily on active engagement in scientific communication, widespread public awareness, and a significant level of Internet accessibility. Although Infectieradar does not offer a replacement for the conventional surveillance system, it can be a crucial supplementary monitoring system that is more timely, provides data at the individual level, and directly measures the incidence of ILI in the community.

Bibliography

[1] George Box. "Box and Jenkins: time series analysis, forecasting and control". In: (2013), pp. 161–215.

[2] *Box-Ljung*. URL: https://koalatea.io/r-ljung-box-test/.

[3] Kostas Danis et al. "High impact of COVID-19 in long-term care facilities, suggestion for monitoring in the EU/EEA, May 2020". In: *Eurosurveillance* 25.22 (2020), p. 2000956.

[4] Emilia Ellsiepen. "Citizen Science for Infectious Disease Surveillance in Belgium–COVID-19". In: (2021).

[5] James Douglas Hamilton. "Time series analysis". In: (2020).

[6] Graham Kalton and Ismael Flores-Cervantes. "Weighting methods". In: *Journal of official statistics* 19.2 (2003), p. 81.

[7] Scott A McDonald et al. "Risk factors associated with the incidence of self-reported COVID-19-like illness: data from a web-based syndromic surveillance system in the Netherlands". In: *Epidemiology & Infection* 149 (2021).

[8] Rizwan Mushtaq. "Augmented dickey fuller test". In: (2011).

[9] Sander P van Noort et al. "Ten-year performance of Influenzanet: ILI time series, risks, vaccine effects, and care-seeking behaviour". In: *Epidemics* 13 (2015), pp. 28–36.

[10] Daniela Paolotti et al. "Web-based participatory surveillance of infectious diseases: the Influenzanet participatory surveillance experience". In: *Clinical Microbiology and Infection* 20.1 (2014), pp. 17–21.

[11] *Sciensano Health Institute*. URL: https://epistat.sciensano.be/covid/.

[12] Stephen D Simon. "Understanding the odds ratio and the relative risk." In: *Journal of andrology* 22.4 (2001), pp. 533–536.

[13] *statbel*. URL: https:https://statbel.fgov.be/en.

[14] *UZA*. URL: https://www.uza.be/ethics-committee-uza.

[15] Yannick Vandendijck, Christel Faes, and Niel Hens. "Eight years of the Great Influenza Survey to monitor influenza-like illness in Flanders". In: *PLoS One* 8.5 (2013), e64156.

[16] Scott L Zeger, Kung-Yee Liang, and Paul S Albert. "Models for longitudinal data: a generalized estimating equation approach". In: *Biometrics* (1988), pp. 1049–1060.