UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

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

Master's thesis

Citizen Science for Infectious Disease Surveillance in Belgium COVID-19

Emilia Ellsiepen

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

SUPERVISOR :

dr. Lisa HERMANS

2020
2021

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

Citizen Science for Infectious Disease Surveillance in Belgium COVID-19

Emilia Ellsiepen

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

SUPERVISOR :

dr. Lisa HERMANS

Abstract

In this study, data from the survey "de Grote Corona-Studie" are used to derive suspected COVID-19 cases and compare these to the official laboratory-confirmed incidence in Belgium. There are two goals to this: Firstly, to assess the general feasibility of using citizen science data to estimate incidence, and secondly, to compare the performance of different case definitions as given by national and supranational health agencies, namely Sciensano, CDC, RKI, ECDC and WHO. The comparison is performed using graphical means, raw correlations between the different time series, and correlations between the time series after pre-whitening using ARIMA models. Furthermore, different lags between the survey derived times series and the laboratory confirmed cases are tested and additional correlation analyses are performed for different gender and age groups. Lastly, sensitivity and specificity of the case definitions are compared based on a small subset of the survey data for which PCR test results were available.
As a main result, robust correlations are found that persist after removing autocorrelation when using the case definitions of ECDC, CDC, or Sciensano. For the time series based on the case definitions of WHO and RKI, correlations are vulnerable to the removal of autocorrelation. Greater correlations are observed for the female subgroup and the senior age group for all but the WHO case definition, whereas correlations are lower for the underpopulated group of teenagers and children. A zero lag is found to result in the highest correlations in the pre-whitened time series for all but the RKI case definition, which shows highest correlation for a 6-day lag and thus has a greater potential to predict cases in the future. All results are to be treated with caution when applying to future scenarios, as the anti COVID-19 measures in place in Belgium reduced the prevalence of ILI and common cold, which are easily confused with COVID-19.

1 Introduction

Surveillance of epidemics and pandemics is vital for the health sector of a country. As we have seen during the current COVID-19 pandemic, incidence data can be used to anticipate shortages in intensive care capacities as well as to inform political decision makers of the necessity as well as the effectiveness of non-pharmaceutical interventions. For influenza, this has long been acknowledged and there are multiple surveillance systems in place, including the sentinel network EISN, combining national sentinel systems from different European countries, the U.S. Influenza Surveillance System by the CDC, the Canadian FluWatch and many others. In addition to surveillance based on either laboratory confirmed cases or cases as reported by general practitioners, it is possible to use symptom questionnaires for the general public, which can be filled in online and are conducted in regular intervals. For influenza, these kinds of citizen science projects are used in many countries, including Australia, the USA, the U.K. and in many European countries as part of the Influenzanet network. For the current COVID-19 pandemic, there are also a number of citizen science surveys, which address symptom burden as well as general well-being, compliance with measures in place and many more aspects. In this study, we will use the symptom data from "De Grote Corona-studie" in Belgium and correlate the symptom burden over time to the official incidence data for COVID-19 obtained from Sciensano. In contrast to influenza, COVID-19 is only a young disease, where the clinical appearance in terms of signs and symptoms is still controversial and has undergone several amendments since its first recorded appearance in Wuhan in December 2019. As an example, the loss of smell and taste was not part of the very early descriptions (Huang et al., 2020), but has gained rapid acknowledgment as an indicative symptom in late spring 2020 (Giacomelli et al., 2020). Fever and a dry cough, in contrast, have always been considered prominent features of COVID-19, but have much less decisive potential, as they are also common for influenza and other influenza-like illnesses (ILIs). As a consequence of the still rapidly developing understanding of the disease and its most prominent symptoms, multiple clinical case definitions of COVID-19 are currently is use, issued by different authorities like the WHO, the ECDC and national agencies like Sciensano, the CDC, the RKI and many others. Our second objective in this study is to compare different clinical case definitions in terms of their correlation with the confirmed cases.

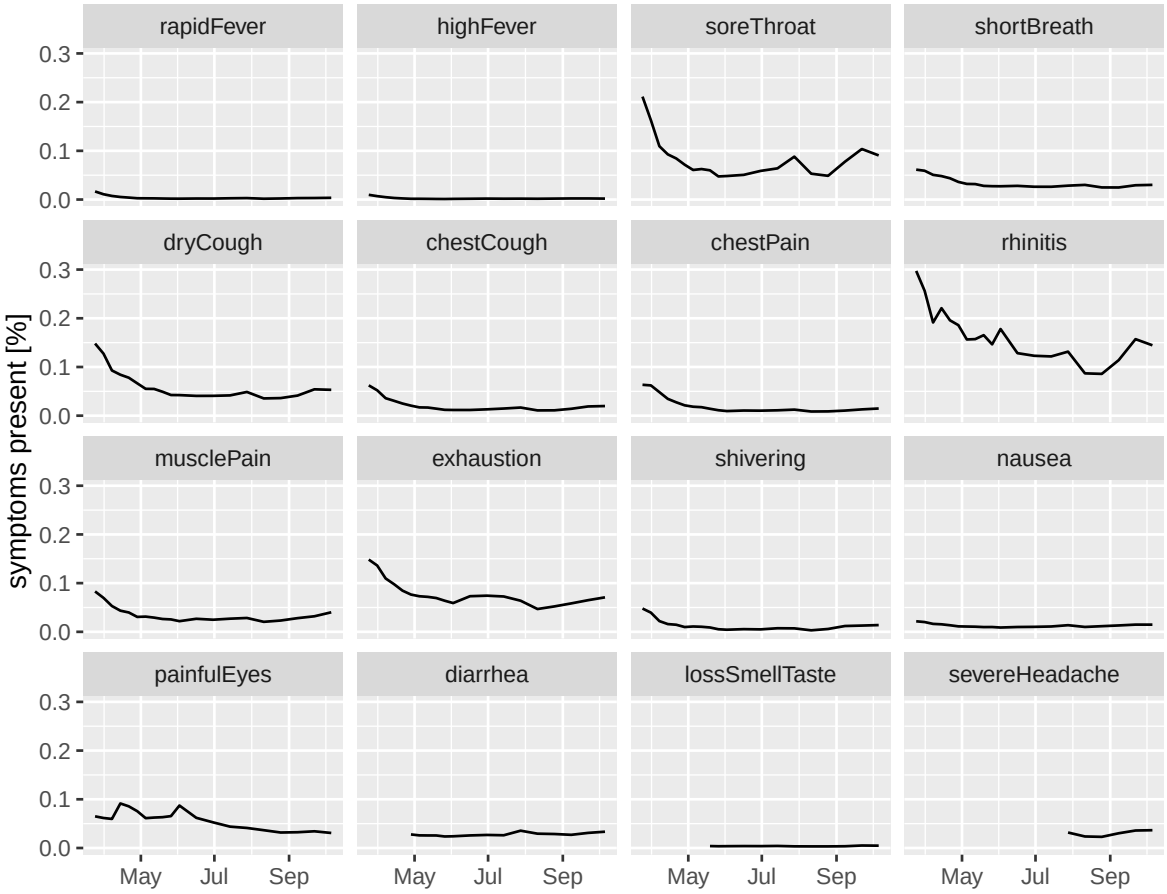

Figure 1: Evolution of symptoms over time

2 Data and summary statistics

The data analyzed here come from two main sources, namely the Grote Corona-studie (GCS), and the daily numbers of confirmed COVID-19 cases in Belgium. Two auxiliary data sources used to motivate methodological decisions are the Belgian demographic data as published by the Belgian national statistics institute statbel from the year 2019, and the number of PCR tests conducted per day, provided by the Belgian health institute Sciensano, which allows us to derive the positive test rate.

The GCS started in March 17, 2020, and is still ongoing. It is a survey which can be completed online by all residents of Belgium on a voluntary basis and is available in four languages (Dutch, French, German, English). While the questionnaire also covers the behavior and mental well-being during the pandemic, we will focus our attention on COVID-19-related symptoms and make use of the demographic data (age, gender, province, occupation, etc.). The data set also contains three sets of weights, we will use the raking weight by age, gender, and province. During its deployment, there have been several changes to the composition and timing of the questionnaire. Up to the 12th round (June 02, 2020) the survey was conducted weekly, and it switched to biweekly after that. Most importantly for the present study, symptoms were not queried after the 21st round (October 06, 2020). Therefore, we will only use the data up to that point and furthermore exclude the first round due to problems in the data collection process. Other important modifications concern the presence of specific symptoms in the symptom questionnaire, where some symptoms were added at later stages (see Figure 1).

The second main data set is the number of lab-confirmed cases provided by the Belgian population health institute (Sciensano). This data set is publicly available and contains aggregated

case numbers by age class in steps of 10, by gender and by province per day. While presently also positive rapid-antigen tests are included in this data set, this is not the case for the period we are investigating (until mid October 2020), when only positive PCR tests were counted.
Table 1 shows that the survey is not representative of the Belgian population in terms of age and gender. In particular, most rounds contain more than twice as many completed questionnaires by female participants than by male participants. Also, the younger age group (0–19) is largely underrepresented in the sample. With regard to older age groups, there seem to be some shifts over time: Up to round 5, the young adult age group (20–39) was overrepresented, while later proportions of older adults (40–59) and seniors (60+) increase. The total number in all age groups decrease over time, so this effect is likely due to the young adults dropping out with a higher rate. Tables 2 and 3 give the total numbers and percentages over the provinces with a comparison to the Belgian population. Here we also see large differences between sample and study population, in that notably the Antwerp region makes up almost half of the sample, while regions in Walloon and Brussels are underrepresented.

3 Related work

Although we find that the numbers of studies related to COVID-19 is rapidly increasing, few prove to be directly of interest in the context of our current research question. With regard to the importance of different symptoms associated with COVID-19, a recent Cochrane review article summarizes the findings of 44 studies that investigated the specificity and sensitivity of a number of symptoms in isolation and two studies that looked at symptom combinations (Struyf et al., 2021). Some of the results were high sensitivity, but low specificity for cough and fever and the reverse pattern for the loss of smell and taste. The authors conclude that no symptom in isolation exhibits high diagnostic accuracy and the rarity of studies that look at combinations of symptoms is regretted. The studies looking at combinations found relatively high specificity, but very low sensitivity ($< 30\%$). A different attempt to combine different symptoms is described by Menni et al., 2020, who analyzed data from a U.K.-based citizen science project. Using a subset of cases for which the result of a laboratory test was available, they build a logistic regression model that identifies loss of smell and taste, cough, fatigue, and anorexia as significant predictors for a positive PCR test. Subsequently, they use this model and the symptoms data from the complete data set to predict the number of COVID-19 cases among their respondents. In contrast to their approach, we will use predefined case definitions that do not quantify the influence of isolated symptoms. A further difference is that they only use data from one time frame, while we aim to account for the longitudinal evolution of cases derived from the citizen science symptom data.
Another relevant line of research comes from the study of ILI epidemics. For ILI, a larger number of surveillance systems are in place, including citizen science projects, networks of general practitioners, and national registers of laboratory-confirmed cases. A number of studies evaluate citizen science projects by comparing them to more traditional sentinel systems (Richard et al., 2020; Carlson et al., 2009; Friesema et al., 2009; Noort et al., 2015; Vandendijck, Faes, and Hens, 2013) All of these studies relate the time series derived from the citizen science survey to another time series, which is either laboratory confirmed cases, or a GP network. With regard to the methods, we identify three approaches: A purely graphical evaluation by plotting the two time series together, an evaluation based on correlations between individual time points of the two time series, and an evaluation based on correlations between the two data set after accounting for their autocorrelation structure by an ARIMA model.
For the latter, there are two distinct uses: Carlson et al., 2009 build an ARIMA model and pre-whiten both time series prior to the calculation of correlation with the goal to filter out common temporal trends, that are not causally related. As a result, their survey derived pre-whitened incidence is correlated with pre-whitened laboratory confirmed influenza cases only for the

Table 1: Demographics (age and gender) of the GCS by survey round

		2	3	4	5	6	7	8	9	10	11	
		March 24	March 31	April 07	April 14	April 21	April 28	Mai 05	Mai 12	Mai 19	Mai 26	Belgian population
	total	345966	415073	224437	197705	169876	119634	80595	82961	71592	49039	11492641
Gender	male	118561	140110	67326	58165	50753	34589	22511	23878	20336	13551	5660064
		34%	34%	30%	29%	30%	29%	28%	29%	28%	28%	49%
	female	226989	274485	156843	139323	118950	84904	57978	58991	51161	35435	5832577
		66%	66%	70%	70%	70%	71%	72%	71%	71%	72%	51%
	other	414	468	268	217	173	141	105	92	95	51	
		< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	
Age	0-19 years	12849	13496	5852	5054	3828	2629	1744	1734	1273	754	2569322
		4%	3%	3%	3%	2%	2%	2%	2%	2%	2%	22%
	20-39 years	173506	189173	98662	84741	65297	45557	29341	26643	23419	14396	2899935
		50%	46%	44%	43%	38%	38%	36%	32%	33%	29%	25%
	40-59 years	120649	155115	87459	77950	68672	47613	31771	32808	28757	19372	3095167
		35%	37%	39%	39%	40%	40%	39%	40%	40%	40%	27%
	60+ years	38960	57279	32464	29960	32079	23835	17739	21776	18143	14515	2928217
		11%	14%	14%	15%	19%	20%	22%	26%	25%	30%	26%

		12	13	14	15	16	17	18	19	20	21	
		June 02	June 16	June 30	July 14	July 27	Aug 11	Aug 25	Sept 08	Sept 22	Oct 06	Belgian population
	total	60583	35772	26688	27537	35618	29944	24947	22357	19425	19750	11492641
Gender	male	18328	10052	7830	8160	10638	8950	7390	6387	5898	6187	5660064
		30%	28%	29%	30%	30%	30%	30%	29%	30%	31%	49%
	female	42188	25671	18828	19351	24927	20950	17525	15939	13504	13540	5832577
		70%	72%	71%	70%	70%	70%	70%	71%	70%	69%	51%
	other	67	49	29	26	53	44	32	31	23	23	
		< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	
Age	0-19 years	859	376	199	144	267	218	197	134	128	153	2569322
		1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	22%
	20-39 years	18897	8480	5781	5884	9501	7008	5302	4520	3706	3756	2899935
		31%	24%	22%	21%	27%	23%	21%	20%	19%	19%	25%
	40-59 years	24201	14201	10458	11211	14364	12389	10418	9509	7996	7835	3095167
		40%	40%	39%	41%	40%	41%	42%	43%	41%	40%	27%
	60+ years	16626	12715	10248	10298	11486	10329	9030	8194	7595	8006	2928217
		27%	36%	38%	37%	32%	34%	36%	37%	39%	41%	26%

Table 2: Number of completed surveys by province, rounds 2 to 11, and Belgian population

	2	3	4	5	6	7	8	9	10	11	
	March 24	March 31	April 07	April 14	April 21	April 28	Mai 05	Mai 12	Mai 19	Mai 26	Belgian population
Antwerp	126347	148743	94854	82152	68403	52562	36264	37368	32468	23540	1869730
	37%	36%	42%	42%	40%	44%	45%	45%	45%	48%	16%
Brussels	14321	11992	4980	3985	3072	2314	1642	1464	1720	1042	1218255
	4%	3%	2%	2%	2%	2%	2%	2%	2%	2%	11%
Hainaut	5242	4378	1645	1194	809	573	428	380	379	233	1346840
	2%	1%	1%	1%	< 1%	< 1%	1%	< 1%	1%	< 1%	12%
Limburg	33211	42075	20315	18483	16590	10304	7111	7211	6158	4200	877370
	10%	10%	9%	9%	10%	9%	9%	9%	9%	9%	8%
Liège	4795	3699	1183	761	538	403	255	221	312	194	1109800
	1%	1%	1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	10%
Luxembourg	1609	882	336	322	223	181	130	93	126	76	286752
	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	3%
Namur	2963	2026	949	523	399	314	245	175	248	128	495832
	1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	4%
East Flanders	65586	80030	39512	35523	30505	20188	13064	13649	12281	8103	1525255
	19%	19%	18%	18%	18%	17%	16%	16%	17%	17%	13%
Flemish Brabant	47415	59186	30872	26555	23523	16015	10466	10901	10110	6669	1155843
	14%	14%	14%	13%	14%	13%	13%	13%	14%	14%	10%
Walloon Brabant	4408	3722	1281	1081	724	541	423	351	402	263	406019
	1%	1%	1%	1%	< 1%	< 1%	1%	< 1%	1%	1%	4%
West Flanders	35975	47287	22426	19719	18020	10709	6931	7278	6617	4109	1200945
	10%	11%	10%	10%	11%	9%	9%	9%	9%	8%	11%
Missing	4094	11053	6084	7407	7070	5530	3636	3870	771	482	
	1%	3%	3%	4%	4%	5%	5%	5%	1%	1%	

Table 3: Number of completed surveys by province, rounds 12 to 21, and Belgian population

	12 June 02	13 June 16	14 June 30	15 July 14	16 July 27	17 Aug 11	18 Aug 25	19 Sept 08	20 Sept 22	21 Oct 06	Belgian population
Antwerp	27249	16856	12630	12623	18242	14145	11904	10533	9043	9223	1869730
	45%	47%	47%	46%	51%	47%	48%	47%	47%	47 %	16%
Brussels	1193	692	503	547	566	504	481	436	406	473	1218255
	2%	2%	2%	2%	2%	2%	2%	2%	2%	2 %	11%
Hainaut	240	145	105	103	137	96	102	83	95	90	1346840
	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	12%
Limburg	5408	3267	2271	2412	2865	2577	2100	1887	1657	1601	877370
	9%	9%	9%	9%	8%	9%	8%	8%	9%	8 %	8%
Liège	185	99	62	80	101	73	60	65	83	82	1109800
	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	10%
Luxembourg	74	40	34	40	45	34	29	26	27	36	286752
	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	3%
Namur	130	69	46	44	77	41	44	41	48	53	495832
	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	4%
East Flanders	10528	5824	4395	4704	5406	5008	4081	3683	3138	3162	1525255
	17%	16%	16%	17%	15%	17%	16%	16%	16%	16%	13%
Flemish Brabant	8890	4937	3864	4016	4862	4177	3507	3191	2800	2862	1155843
	15%	14%	14%	15%	14%	14%	14%	14%	14%	14%	10%
Walloon Brabant	246	169	114	113	151	114	132	94	115	117	406019
	< 1%	< 1%	< 1%	< 1%	< 1%	< 1%	1%	< 1%	1%	1%	4%
West Flanders	5941	3382	2432	2607	2888	2931	2279	2117	1823	1823	1200945
	10%	9%	9%	9%	8%	10%	9%	9%	9%	9%	11%
Missing	499	292	232	248	278	244	228	201	190	228	
	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	

unvaccinated group, while the vaccinated group only shows a correlation in the unfiltered data sets. This is intuitively appealing, as ILIs that are not due to influenza can be expected to be correlated simply by their appearance in a similar period of the year, whereas the pre-whitened series specifically inform about departures from the general trend and the correlation shows whether these departures are related. Noort et al., 2015 use the same methodology to account for seasonal effects in their multi-seasonal data set and to identify optimal lags between the two time series. Another approach is taken by Vandendijck, Faes, and Hens, 2013. Here, a random walk model, which belongs to the class of ARIMA models, is used to model the survey derived incidence. The correlation, however, is calculated between the fitted values of the random walk model, rather than the residuals as above. The result is that the time series essentially becomes smoother and takes into account the longitudinal nature of the data. Correlations are rather expected to be higher with this approach and it does not prevent from finding correlations that are rooted in the season rather than the incidence of a specific pathogen. As they compare survey ILI incidence to clinical ILI incidence and do not differentiate between influenza and similar illnesses, this is not per se problematic. In the present study, we will focus on the first approach, as we also compare to laboratory confirmed cases. In addition, we expect COVID-19 incidence to show similar temporal patterns as ILI, and at the same time the clinical diagnosis of COVID-19 and ILI show considerable overlap. We therefore expect these two diseases to be confounded in the survey derived incidence and expect the correlation between the pre-whitened time series to be a better indicator of the ability of one case definition over the other to predict COVID-19.

4 Case definitions

The following provides an overview of different case definitions. It should be noted that the term case definition is not unambiguous and that they are formed for different purposes, most notably for either surveillance (CDC) or as a diagnostic guideline to decide which individuals should be tested (RKI). Not all definitions specify their purpose. Furthermore, most case definitions contain clinical criteria, which can be matched to the symptoms in the GCS, epidemiological criteria and laboratory criteria. We will only take into account the clinical criteria. Finally, case definitions are sometimes further divided into suspected, probable, and confirmed, with clinical symptoms either indicating suspected or probable cases. We will consider all categories alike and in the following reformulate the case definitions only considering clinical criteria. Each bullet point represents a sufficient criterion for a case, which can either consist of one symptom, two symptoms together, or two to three symptoms from a longer list of symptoms.

WHO The case definition was retrieved at `https://www.who.int/publications/i/item/WHO-2019-nCoV-Surveillance_Case_Definition-2020.2` and published on December 16, 2020.

- Acute onset of fever and cough
- Acute onset of any three or more of the following signs or symptoms:
  fever, cough, general weakness/fatigue, headache, myalgia, sore throat, coryza, dyspnoea, anorexia/nausea/vomiting, diarrhea, altered mental status
- recent onset of anosmia or ageusia

CDC The case definition was retrieved at `https://wwwn.cdc.gov/nndss/conditions/coronavirus-disease-2019-covid-19/case-definition/2020/08/05/` and published on August 08, 2020.

- at least two of the following:

Fever, chills, rigors, myalgia, headache, sore throat,
nausea/vomiting, diarrhea, fatigue, congestion or runny nose

- cough
- shortness of breath
- difficulty breathing
- new olfactory disorder
- new taste disorder

Sciensano The case definition was retrieved at `https://covid-19.sciensano.be/nl/covid-19-gevalsdefinitie-en-testing` on March 10, 2021.

- cough
- dyspnoea
- thoracic pain
- acute anosmia or dysgeusia
- at least two of the following symptoms:
  fever; muscle strain; fatigue; rhinitis; sore throat; headache; anorexia;
  watery diarrhea; acute confusion; sudden fall

ECDC The case definition was retrieved at `https://www.ecdc.europa.eu/en/covid-19/surveillance/case-definition` and published on December 3, 2020.

- cough
- fever
- shortness of breath
- sudden onset of anosmia, ageusia or dysgeusia

RKI The case definition was retrieved at `https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Falldefinition.pdf` and published on December 23, 2020.

- any respiratory complaints
- loss of taste and/or smell

For a summary and comparison, please refer to Table 4. The table shows that some symptoms are used throughout all case definitions, e.g., the loss of smell and taste is a sufficient criterion in all clinical case definitions. Other symptoms, like nausea and vomiting, appear only in a subset of case definitions and yet other symptoms, like cough and rhinitis, are sufficient criteria in some definitions while in other definitions additional symptoms have to be added to identify a case. There were also a number of symptoms present in a case definition, but not queried in the GCS, e.g., anorexia or an altered mental status.
In order to compare the case definitions, it would be interesting to impose a ranking based on the relative strictness. This is a difficult task, as the symptom sets used in the different definitions are not generally subsets of each other and because of the different status as sufficient or not. As an example, consider the ECDC definition and the WHO definition: The ECDC definition only uses symptoms, which are also used in the WHO definition, but not all of them, suggesting the ECDC definition to be stricter than WHO. At the same time, all symptoms are considered sufficient in the ECDC definition, while for a WHO case there have to be two or even three symptoms in combination, which in turn suggests the WHO definition to be stricter. Which one is ultimately triggered more often thus crucially depends on the prevalence of the symptoms in isolation as well as in combination. An easy to establish ordering is that the CDC definition is more lenient than WHO, as it includes the same set of symptoms, but counts more symptoms as sufficient to define a case. The Sciensano definition lacks nausea from the list of symptoms, but also includes more sufficient symptoms than the WHO: together with the observation that

Table 4: Table of different case definitions in relation to the symptoms queried in the Grote Corona Studie (GCS). S stands for sufficient cause, C_2 for conjunction of at least two symptoms, C_3 for conjunction of at least 3 symptoms

Symptom in GCS	WHO			CDC		ECDC	RKI	Sciensano	
	S	C_2	C_3	S	C_2	S	S	S	C_2
Rapid fever/ High fever		✓			✓	✓			✓
Sore throat			✓		✓		✓		✓
Shortness of breath			✓	✓		✓		✓	
Dry cough/Chest cough		✓		✓		✓	✓	✓	
Chest pain								✓	
Rhinitis			✓		✓		✓		✓
Muscle pain			✓		✓				✓
Fatigue			✓		✓				✓
Chills					✓				
Nausea, vomiting			✓		✓				
Painful eyes									
Diarrhea			✓		✓				✓
Loss of smell and taste	✓			✓		✓	✓	✓	
Severe headache			✓		✓				✓
Not part of GCS									
Altered mental status			✓						✓
Difficulty breathing					✓				
Headache			✓		✓				✓
Anorexia									✓
Sudden fall									✓

nausea is a rather rare symptom (see Figure 1), we also conclude that it is more lenient than WHO and very similar to the CDC case definition. The RKI case definition uses a subset of the WHO symptoms, but again all of them are sufficient criteria. In contrast to the ECDC definition, even rhinitis is considered sufficient, which we can see in Figure 1 is by far the most frequently observed symptom overall. We thus conclude that despite the low number of symptoms included, this is the most lenient one. We arrive at an ordering, where WHO is the strictest definition, followed by CDC and Sciensano and with RKI being the most lenient one, whereas it is difficult to place the ECDC definition in this ranking.

5 Methods

5.1 Data preprocessing

5.1.1 Cases derived from GCS data

The case definitions as described above were implemented in R after mapping the symptoms to their GCS counterparts (see Table 5). As noted above, some symptoms used in the case definitions were not part of the GCS and can thus not be taken into account. Other symptoms, in particular "(severe) headache", "loss of smell and taste", and "diarrhea" were only included from waves 16, 10, and 7 respectively. These symptoms were partly recovered from the free text symptom field by a simple regular expression match, i.e. each case containing the strings "oofdpijn", "maak", or "iarree", excluding the first character to match both lower and upper case, were counted as positive cases. With this technique, 9126 cases of headache, 2261 cases of loss of smell or taste and 1579 cases of diarrhea could be recovered. For "confusion", and "sudden fall" which are part of the Sciensano case definition, this method was not attempted and they were not included in the case definition coding, as they were only queried in the last wave.

Table 5: Symptom correspondence table, leaving out literal correspondence

GCS	Case definitions
Rapid fever	Fever
High fever	Fever
Sore throat	Any respiratory complaints
Shortness of breath	Dyspnoea
Dry cough	Cough, any respiratory complaints
Chest cough	Cough, any respiratory complaints
Rhinitis	Coryza, congestion or runny nose, any respiratory complaints
Muscle pain	Myalgia, muscle strain
Diarrhea	Watery diarrhea
Loss of smell and taste	Recent onset of anosmia or ageusia, new olfactory disorder, new taste disorder, acute anosmia or dysgeusia, sudden onset of anosmia, ageusia or dysgeusia
Severe headache	Headache

Cases for which all symptoms were attributed to a known allergy were set to zero as they cannot be regarded as indicative of COVID-19. This could only be done from wave 5 onward, as the question was not part of the survey in waves 2–4.
Because we have largely different numbers of completed surveys for the different waves, we will systematically use the percentage or the incidence per 100,000 participants, not the total numbers.

5.1.2 Weights

The GCS is not a representative survey, and we find that it is not balanced over the provinces (see Table 2). Since we know that the incidence of COVID-19 has a spatial component, this is expected to reduce a possible correlation between observed cases and cases derived from the GCS. For this reason, in addition to the raw percentage of clinical cases, we also look at the percentage of weighted case, where we use a raking weight including age, gender and province which was included in the data set. Raking is an iterative procedure of proportional fitting to marginal frequencies of the population. Using raking over post-stratification weights is mostly motivated by either using a larger number of variables and wanting to avoid the resulting small strata, or the sole availability of marginal frequencies for the population. Weights were used to be able to compare the two quantities (cases from the survey with confirmed cases in the general population), assuming that the laboratory confirmed cases are associated with a smaller bias, although whether a test is taken, of course, also depends on external factors and demographics. Since not all participants specified their province, cases without this information were excluded from further analysis (2.6%)

5.1.3 Confirmed cases

Although symptoms were reported either for the preceding 10 days (waves 2–12) or the preceding 14 days (waves 13–21), confirmed cases were always aggregated over the week preceding the survey date. This was done primarily to obtain a time series with equal intervals over the whole period. Moreover, it could be argued that symptoms more than a week ago may not be reported accurately. The aggregation was a simple summation over all confirmed cases in the aggregation interval.
For the analyses by gender and age group, confirmed cases were aggregated only over those subgroups. For the analyses pertaining to the optimal lag between GCS cases and confirmed

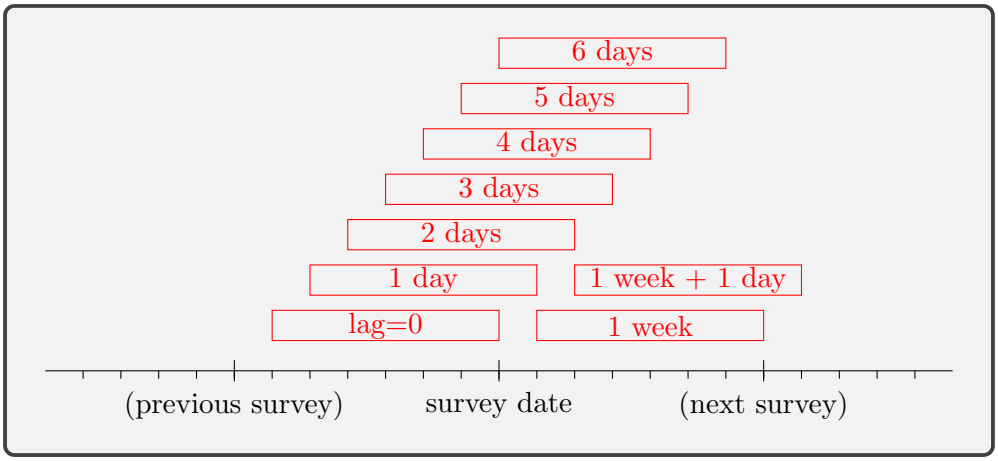

Figure 2: Illustration of aggregation windows for laboratory confirmed cases for baseline time series (lag=0) and with shifts by day. Small ticks are days, bigger ticks weeks. Days on the edge of the box are included in the aggregation interval.

cases, the aggregation window was shifted forward by 1 day up to 8 days obtaining thus 9 distinct time series, see Figure 2 for an illustration.
For the Figures, the summed confirmed cases were transformed to a 7-day incidence rate by dividing the number of cases with the size of the Belgian population (11,492,641) and multiplying by 100,000.

5.2 Statistical analysis

To validate that the cases derived from the GCS are a viable proxy for confirmed cases and to select the best fitting case definition, we proceed in three steps. Firstly, we inspect the evolution of the incidence of confirmed cases as compared to cases derived from GCS in line graphs. To facilitate means of comparison, confirmed cases are displayed on their own scale. Secondly, we calculate Pearson correlations between confirmed cases and GCS case proportions for the different case definitions. In a third step, we take into account that the two vectors are actually time series and pre-whiten the original data using an ARIMA before we calculate correlations. While we will include all available data for step 1, we will leave out waves 2–4 from the two subsequent steps, because of the potential unreliability of the confirmed cases. In particular, the ratio of positive tests was substantial in this period, namely above 25%, which we will use as the threshold here (see also Figure 3). Such a high positive test rate usually means that not enough tests were available and only those individuals got tested, where the suspicion of COVID-19 was highest. With this testing strategy, a large number of undetected cases is expected hence the observed number of cases will underestimate the true number of infections.
For all correlation analyses, proportions (GCS data) are first subjected to a logit transformation (Warton and Hui, 2011) and the incidence data is subjected to a log transformation (Benvenuto et al., 2020) in order to stabilize the mean and approach a normal distribution, which is assumed in the inferential tests for the correlation as well as in comparisons of correlations. The data is plotted without the transformation.

5.2.1 ARIMA pre-whitening

ARIMA models are used to model and predict time series data. In addition, they can be used as a filter in a pre-processing step to avoid to overestimate the correlation between two autocorrelated time series (Carlson et al., 2009; Bloom, Buckeridge, and Cheng, 2007) and to facilitate identification of a suitable lag (Helfenstein, 1991).

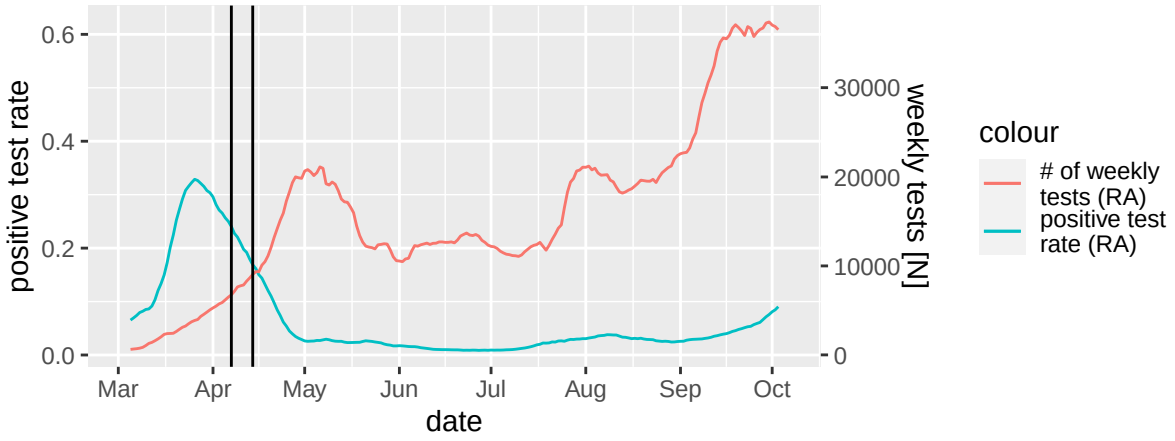

Figure 3: Rolling averages of positive test rate and weekly number of conducted tests. Vertical lines mark beginning and end of the time frame asked about in wave 5, which is the first wave that is used for correlation analyses.

In general, an ARIMA model may contain three components: an autoregressive component, a differencing component, and a moving average component. All of these components can be of different orders. A differencing component of second order, for instance, means that the data is first replaced by the differences between two consecutive data points, and on these differenced data the operation is repeated. Similarly, we can add autoregression components of second or third order and also moving average components of multiple orders. The autoregressive components enter the model equation by including previous measurements with coefficients (the ϕs in the general ARIMA model formula in equation 1). The moving average components, on the other hand, are the errors of previous measurements again multiplied by their own coefficients (the θs in equation 1). The differencing components are integrated by replacing the y_ts by the required difference.

$$y_t = \mu + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} - \theta e_{t-1} - \ldots - \theta_q e_{t-q} \tag{1}$$

The instantiation of an ARIMA (1,1,0) would then be

$$y_t - y_{t-1} = \mu + \phi_1(y_{t-1} - y_{t-2}) \text{ or equivalently } y_t = \mu + \phi_1(y_{t-1} - y_{t-2}) + y_{t-1} \tag{2}$$

In determining the ARIMA parameterization, we rely on the Box-Jenkin's method (e.g. Helfenstein, 1991), which is a stepwise procedure. Initially, the time series is subjected to transformations and differencing to improve stationarity and account for general trends in the data. The residuals are then inspected in plots of the autocorrelation function (ACF) and the partial autocorrelation function (PACF) to determine whether autocorrelation terms or moving average terms should be added and the model is refitted, if necessary. The residuals are again inspected until no apparent autocorrelation is left, which can be validated using a Ljung-Box test.
After building a suitable ARIMA model for the confirmed cases, the estimates of this model are used to pre-whiten both the confirmed cases and the case proportions as derived from the GCS, by applying the model to the data and storing the residuals. For the pre-whitened values, correlations are again calculated. The expectation is that, once we have removed common self-dependencies in the data, the correlation coefficient is less influenced by common overall trends, like for example an increase in all kinds of respiratory complaints induced by the season. Instead, more subtle correlations in the change of the signal possibly will be detected. In addition, we expect the pre-whitened series to give a better insight into which shift of the time window is most suitable.

5.2.2 Missing data

Although we do not observe missing data in the general sense, the different intervals between survey rounds (1 week or 2 weeks) can be conceptualized as a time series with a 1-week interval with observations systematically missing every second time in the later part of the series. The ARIMA model assumes an equally spaced time series, as e.g., the differencing operation is only meaningful for equal distances. We will therefore impute the missing values.
There are numerous techniques for imputation of missing data, but in general few are applicable to the univariate data of a time series. Simple techniques include last observation carried forward, nearest neighbor, and linear interpolation. Linear interpolation is generally found to perform well for short gaps, while longer gaps can be problematic (Junninen et al., 2004). In these contexts, it can even outperform more advanced methods based on ARIMA modeling and multiple machine learning techniques (Salles et al., 2015). As linear interpolation has also been used for ILI incidence before (Rasmussen et al., 2019), we are confident that the method is adequate for our data set and will use it to interpolate between surveys in the second part of the time series.

5.2.3 Influence of weights

The papers discussed above which aimed at validating ILI cases derived from citizen science projects did not use weights. We are interested to investigate, whether using weights is affecting correlations and, if so, in what direction. To this end, we will perform inferential tests on the magnitude of correlations between weighted GCS cases and confirmed cases and unweighted GCS cases and confirmed cases, for each case definition separately. We will use Williams' t statistic (Williams, 1959) for dependent correlations on each of the pairs and subject the resulting p-values to Hommel's adjustment to account for multiple testing. William's t statistic in the formulation of Steiger, 1980 is given in equation 3, where j is the common variable, whereas k and h are the different second variables.

$$T = (r_{jk} - r_{jh})\sqrt{\frac{(N-1)(1+r_{kh})}{2(\frac{N-1}{N-3})|R| + \bar{r}^2(1-r_{kh}^3)}}$$
$$\text{with } |R| = (1 - r_{jh}^2 - r_{jk}^2 - r_{hk}^2) + (2r_{jh}r_{jk}r_{hk}) \text{ and } \bar{r} = \frac{1}{2}(r_{jk} + r_{jh}) \quad (3)$$

5.2.4 Comparison of different case definitions

We will compare the different case definitions mainly on a descriptive level, i.e., using the graphs and correlations described in the previous subsections. In addition and as described above for testing for the importance of weights, we will use Williams' t test on a subset of contrasts and apply Hommel's correction to the p-values of these pairwise comparisons.
While theoretically all contrasts could arguably be of interest, we restrict ourselves here to contrasting the Belgian Sciensano case definition to all other case definitions in order not to suffer either from inflation of α or from an extreme loss in power, when applying a suitable multiple testing correction to the full set of 20 comparisons.

5.2.5 Identification of optimal lag

To test whether correlations improve or deteriorate if a lag is introduced between GCS cases and confirmed cases, we make use of the different aggregation windows described above and compare each case definition proportion to each lag.

This step will also be performed for the pre-whitened time series, where an individual ARIMA model with the same parameterization as identified for the 0-lag time series will be fitted to each of the shifted time series.

5.2.6 Effects of gender and age

COVID-19 is known to affect male and females differently on average and age also plays a prominent role in the progression and prospect of an illness. To investigate whether these differences also amount to different case definitions being more or less related to the observed incidence of this group, we calculate correlations in the subgroups male/female as well as within the age groups 0–19, 20–39, 40–59, and 60+. The reasoning behind the age groups is twofold: Theoretically and based on the literature, distinctions between children, young adults, older adults, young seniors and older seniors would be interesting to look at. However, among the respondents to the GCS, there are very few children and older seniors, therefore these categories had to be merged with adjacent ones. In addition, for the confirmed cases, we only have categorized age in steps of ten, which reduces the possibilities to split up the data. As can be seen in the demographics table above (1), this produces three more or less balanced groups and an additional underpopulated group of 0–19. For the latter group, results should be treated with caution and when interpreting their results, it is important to note that most responses came from older teenagers, with few participants below 10 and none below 6.
Correlations by gender and age are only calculated for unweighted cases, as weights include age and gender information.

5.2.7 Sensitivity and specificity for waves 20 and 21

For waves 20 and 21, there is survey information on whether a laboratory test was taken and about the result. For this subset, specificity and sensitivity will be calculated for the different case definitions. Sensitivity and Specificity for a case definition A are defined as

$$\mathsf{Sensitivity}_{\mathsf{A}} = \frac{\#\text{ of cases with positive PCR test and identified by case definition } \mathsf{A}}{\#\text{ of cases with positive PCR test}} \tag{4}$$

$$\mathsf{Specificity}_{\mathsf{A}} = \frac{\#\text{ of cases with negative PCR test and not identified by case definition } \mathsf{A}}{\#\text{ of cases with negative PCR test}} \tag{5}$$

As testing was likely conducted based on the symptoms that were reported, however, results should not be over interpreted. Normal approximation confidence intervals are reported.

5.3 Software

All statistical analyses are performed in R version 4.0.4 (R Core Team, 2021). The ARIMA analysis was conducted using the R base function arima() and the package `forecast` (Hyndman and Khandakar, 2008) and in particular the `Arima()` function to apply the derived model to the different series for pre-whitening. William's t statistic was calculated with the function `r.test()` from the package `psych`.

6 Results

6.1 Visual comparison

In Figure 4, we plot the incidence rate for each case definition derived from the GCS data and the laboratory confirmed incidence as reported by Sciensano. First focusing on the case definitions, we observe that the ECDC, CDC and Sciensano definitions largely pattern together.

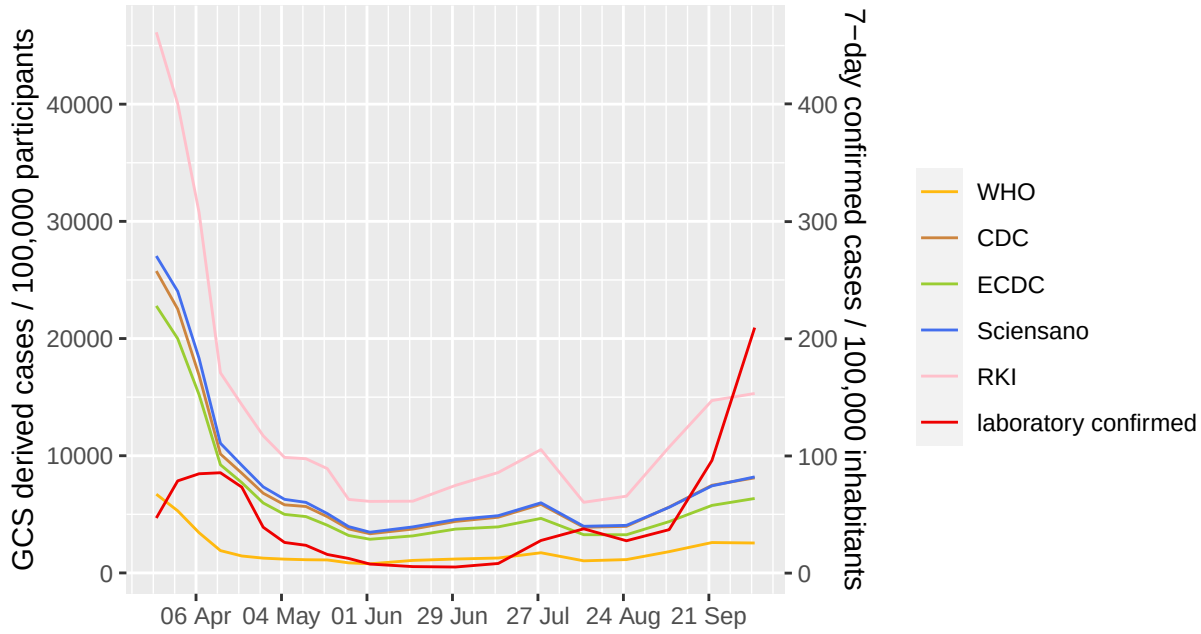

Figure 4: Incidence of COVID-19 cases by case definition as obtained from GCS data and laboratory confirmed COVID-19 incidence (see right y axis for scale), no weights

Based on the case definitions, this was expected for CDC and Sciensano, but unclear for ECDC. Cases according to the WHO are detected at a lower rate and show less of a seasonal effect: there is certainly a drop in the beginning, but the rise starting at the end of the summer and potentially signaling the beginning of the second wave is much less pronounced. The very broad RKI case definition, on the other hand, shows a much higher rate. Comparing to the laboratory confirmed incidence, we observe that the overall pattern of decreasing numbers in spring, low numbers over the summer and an increase towards autumn is reflected in all case definitions. For the confirmed cases, we additionally see a rise in cases from the beginning of the survey period (March 17) extending roughly until mid April. As discussed above (section 5.2), it is likely that the rise is, in fact, the result of the increasing number of tests and that cases are underestimated in this period. Here, the GCS derived incidence might actually be a better reflection of the overall pattern. Overall, the incidence rates differ by a factor of about 100, i.e., incidence as derived by the GCS is a hundred times higher than laboratory confirmed incidence. This may be due to a systematic underestimation of the real incidence by the number of laboratory confirmed tests, a lack of specificity of the case definitions, or a combination of both.

Figure 5 distinguishes between weighted and unweighted cases. There are some noticeable differences between the two curves. In general, it is hard to tell whether the weighted incidence resemble the time course of the laboratory confirmed incidence more closely. In the second wave, especially after September 8, the confirmed incidence rises rapidly, approximately doubling in two weeks. The rise is less pronounced in general for the GCS derived incidence, but the weighted cases show a larger increase than the unweighted cases for all case definitions, suggesting a closer fit. To compare the predictive ability of weighted and unweighted cases in a principled fashion, we refer to the next section where we look at correlations.

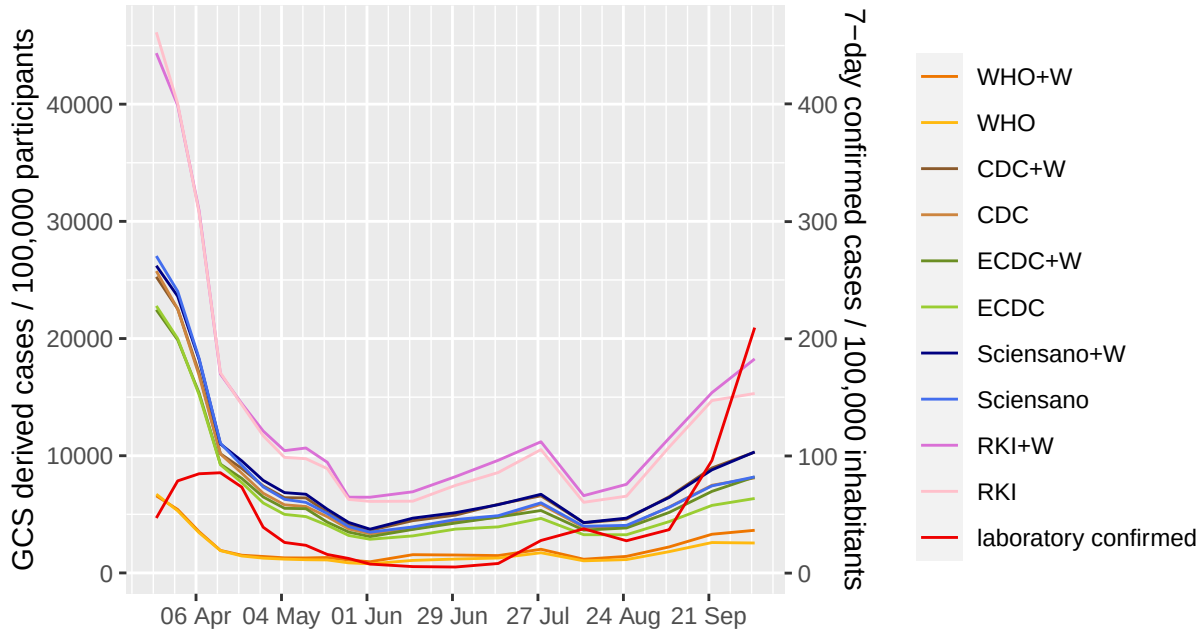

Figure 5: Incidence of COVID-19 cases by case definition and including raking weights (+W) as obtained from GCS data and laboratory confirmed COVID-19 incidence (see right y axis for scale)

6.2 Raw cross correlations

6.2.1 Complete data set without lag

Table 6 lists correlations between logit transformed case proportions derived using the different case definitions from the GCS symptom data, and log transformed laboratory confirmed cases. Only data from wave five onward is used and the table differentiates between using weighted or raw case percentages.

Table 6: Correlations between weighted and unweighted GCS derived case proportions and laboratory confirmed cases without lag, rounds 5–21

		WHO	CDC	ECDC	Sciensano	RKI
weighted	r	0.660	**0.815**	0.796	0.796	0.811
	95% CI lower bound	0.262	0.550	0.511	0.512	0.542
	95% CI upper bound	0.866	0.931	0.924	0.924	0.930
	p	0.004	0.000	0.000	0.000	0.000
unweighted	r	0.765	**0.805**	0.758	0.778	0.799
	95% CI lower bound	0.450	0.529	0.437	0.475	0.516
	95% CI upper bound	0.911	0.927	0.908	0.916	0.924
	p	0.000	0.000	0.000	0.000	0.000

In general, we observe moderate to high positive correlations between the proportion of weighted cases and the confirmed cases ranging between .660 and .815. All correlations are highly significant. The highest correlation is observed for the CDC case definition for both weighted and unweighted analyses. For the WHO case definition the correlation decreases when using weights, whereas all other correlations increase with weights. A statistical test for the difference between correlations showed no significant difference between weighted and unweighted correlations (all $p>0.05$ after applying Hommel correction). The correlation for the case definition of Sciensano was not significantly different from any correlation using the other case definitions (all $p>0.05$

after applying Hommel correction).

6.2.2 Different time lags

Table 7 lists correlations for weighted case proportions derived from the symptom data and confirmed cases aggregated over 1 week preceding the survey (lag 0) or aggregated over a shifted 7-day time interval, where shifts are performed in days. The same transformations as above were applied. We observe that for most case definitions, the optimal lag is 1 day, which includes all confirmed cases between 6 days before the survey was taken until one day after the survey was taken. This suggests that the survey data can already help to estimate confirmed cases one day ahead, which may well be related to the average time a PCR test needs to be evaluated (between 12 and 48 hours). We also observe that correlations decrease only slowly and are still in general high for the interval 1 week after the survey date (all r>0.74). This might indicate that the GCS data has a high predictive capacity, but may also be related to high auto correlation in the data. This will be resolved in section 6.3.2. A special case seems to be the WHO case definition, which shows an increasing correlation with longer lags.

Table 7: Correlations between weighted GCS derived case proportion and confirmed cases calculated for different lags, rounds 5–21

	lag 0	1 day	2 days	3 days	4 days	5 days	6 days	1 week	1 week+1
WHO	0.660	0.698	0.726	0.752	0.767	0.772	0.791	0.809	**0.822**
CDC	0.815	**0.828**	0.823	0.815	0.814	0.811	0.810	0.796	0.787
ECDC	0.796	**0.803**	0.792	0.777	0.773	0.768	0.764	0.745	0.731
Sciensano	0.796	**0.805**	0.796	0.784	0.781	0.776	0.773	0.756	0.743
RKI	0.811	**0.829**	0.825	0.820	0.821	0.818	0.824	0.812	0.803

6.2.3 Analyses by gender and age

Table 8 shows correlations between unweighted GCS derived case proportions and laboratory confirmed cases by gender. Except for the WHO case definition, we observe higher correlations for females than for males, especially for the Sciensano and ECDC case definitions. This suggests that these case definitions are better able to capture the prototypical clinical appearance of COVID-19 in women. For the WHO case definition, the correlation is higher for males, but the difference is also smaller than for the other case definitions, which points to the case definition being more general. When comparing the case definitions, we see that the WHO case definition exhibits the highest correlation in males, while in females again the CDC case definition shows the highest correlation with confirmed cases. All correlations are highly significant.

Table 8: Correlations between unweighted GCS derived case proportion and confirmed cases by gender, transformed, rounds 5–21

	male r	males 95% CI	female r	female 95% CI
WHO	**0.766**	[0.452 ; 0.911]	0.728	[0.380 ; 0.895]
CDC	0.646	[0.240 ; 0.860]	**0.848**	[0.620 ; 0.944]
ECDC	0.598	[0.165 ; 0.838]	0.816	[0.552 ; 0.931]
Sciensano	0.608	[0.180 ; 0.842]	0.833	[0.587 ; 0.938]
RKI	0.712	[0.352 ; 0.889]	0.814	[0.547 ; 0.930]

Table 9 gives correlations between unweighted GCS derived case proportions and laboratory confirmed cases by age group. We first note that correlations are in general much lower for teenagers and children, with the ECDC case definition not differing significantly from zero. As pointed out in section 5.2.6, this might be due to the sparsity of the data, other possible

explanations will be explored in the discussion. Similar to the general population, the CDC case definition results in the highest correlation for teenagers and children. In young adults, the correlations are similar as in the general population, with the highest correlation found for the CDC case definition. In older adults, correlations are also comparable to the general population, and here we find the highest correlation for the WHO case definition. For the senior group, finally, we see that except for the WHO case definition, correlations are very high (all $r > 0.8$), with the RKI case definition resulting in the highest correlation. The WHO case definition, by contrast, shows only a moderate correlation of 0.375, which is not significantly different from zero.

Table 9: Correlations between unweighted GCS derived case proportion and confirmed cases by age group, transformed, rounds 5–21

Age		r	95% CI
teenagers	WHO	0.595	[0.141 ; 0.842]
and children	CDC	**0.597**	[0.143 ; 0.843]
0–19	ECDC	0.443	[-0.067 ; 0.770]
	Sciensano	0.586	[0.127 ; 0.838]
	RKI	0.500	[0.006 ; 0.798]
young adults	WHO	0.715	[0.357 ; 0.890]
20–39	CDC	**0.834**	[0.589 ; 0.938]
	ECDC	0.792	[0.503 ; 0.922]
	Sciensano	0.809	[0.538 ; 0.929]
	RKI	0.776	[0.471 ; 0.915]
older adults	WHO	**0.801**	[0.522 ; 0.926]
40–59	CDC	0.776	[0.471 ; 0.915]
	ECDC	0.716	[0.360 ; 0.890]
	Sciensano	0.743	[0.407 ; 0.901]
	RKI	0.794	[0.507 ; 0.923]
seniors	WHO	0.375	[-0.129 ; 0.725]
60+	CDC	0.867	[0.661 ; 0.951]
	ECDC	0.858	[0.643 ; 0.948]
	Sciensano	0.875	[0.681 ; 0.954]
	RKI	**0.895**	[0.728 ; 0.962]

6.3 Pre-whitened cross correlations

6.3.1 Complete data set without lag

To approach stationarity in the time series of confirmed cases, the series was first log transformed and differenced to account for the time trend in the data. Inspection of PACF and ACF on the transformed and differenced data showed a significant partial autocorrelation at lag 1 and a decreasing pattern of autocorrelations with increasing lag for the ACF, therefore an autocorrelation term of first order was included and estimated at 0.736 (0.143). Model diagnostics of the resulting ARIMA of the form (1,1,0) are displayed in Figure 6 and show that the time series was successfully detrended. This was confirmed by a Ljung-Box test on the residuals ($p = 0.865$). Alternative ARIMA models (0,1,0) and (0,2,0) were fitted in addition. The random walk model (0,1,0) displays significant autocorrelation in the residuals and a significant Ljung-Box test and is hence not considered adequate. The random walk model of second order (0,2,0) shows similarly good diagnostics as the selected model and a slightly smaller AIC value. If applied to the GCS data, however, it produces as an artifact a zero residual for all interpolated values, which distorts the general picture considerably and hence this model is also

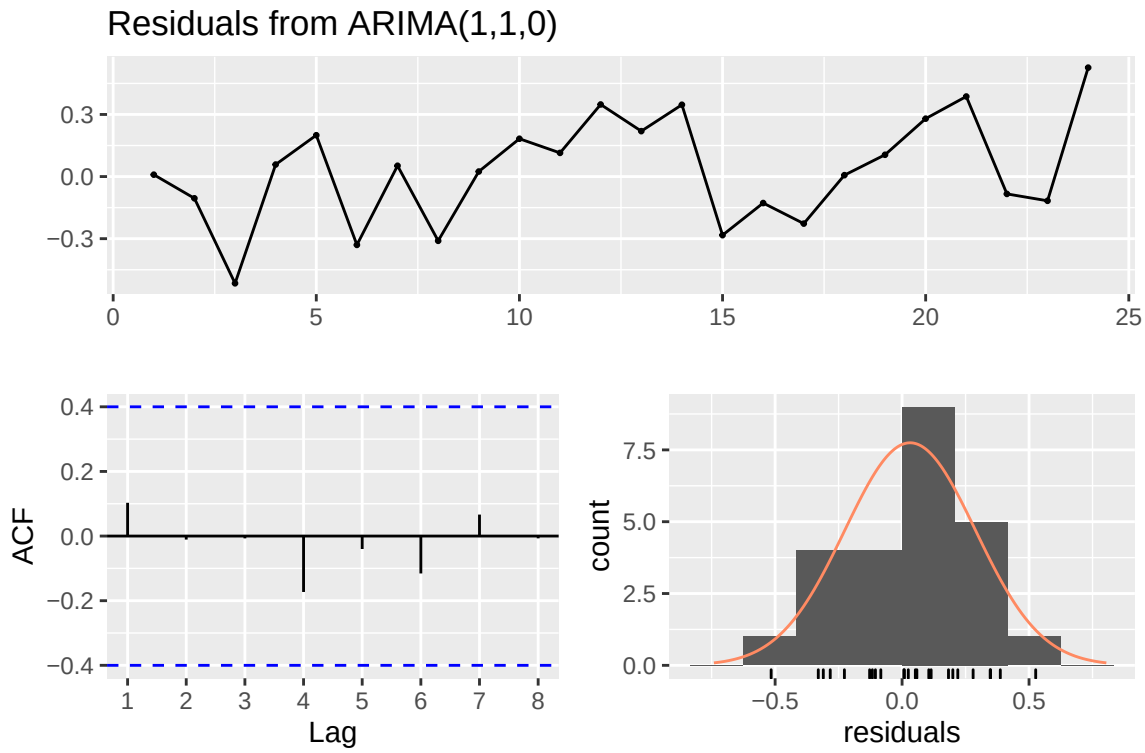

Figure 6: Model diagnostics of ARIMA model for confirmed cases time series.

not used.
Figure 7 shows the residuals of the selected ARIMA model along with the pre-whitened GCS derived case definition time series. After detrending, we notice that graphical inspection is not very informative and gives little intuition as to which case definition produces the greatest correlation. We therefore turn to Table 10, which lists correlation coefficient, 95% confidence intervals and p-values for the pre-whitened series. As expected, correlations are in general much reduced because the common trend is removed from the data and only reach small to moderate effect sizes (between 0.12 and 0.48). We further observe that some correlations do not reach significance any longer, in particular the WHO and RKI case definitions. The highest correlation is observed for the ECDC case definition for both weighted and unweighted cases and closely followed by the CDC and Sciensano case definitions. For these three, the weighted cases result in slightly stronger correlations, whereas for RKI and WHO case definitions, the unweighted cases show a larger correlation coefficient. For the RKI and WHO case definitions, we can thus not conclude that they correlate to confirmed cases more than expected by a common temporal trend. For the other three, the correlations are only moderate after accounting for the common trend, but still significantly different from zero.

6.3.2 Different time lags

For each of the time series obtained by shifting the aggregation window, the same procedure as above was applied to obtain an appropriate ARIMA model. For all lags, the ARIMA (1,1,0) was found to be an adequate model with Table 11 giving the autocorrelation estimate and standard error along with the AIC and the Ljung-Box p-value.
Table 12 shows correlation coefficients for the weighted and pre-whitened time series derived from the GCS and the pre-whitened confirmed cases aggregated over different time windows with lag 0 starting 6 days before the survey was filled in and extending to the survey date and the subsequent windows each shifted by one day to the right (see Figure 2). For CDC, ECDC,

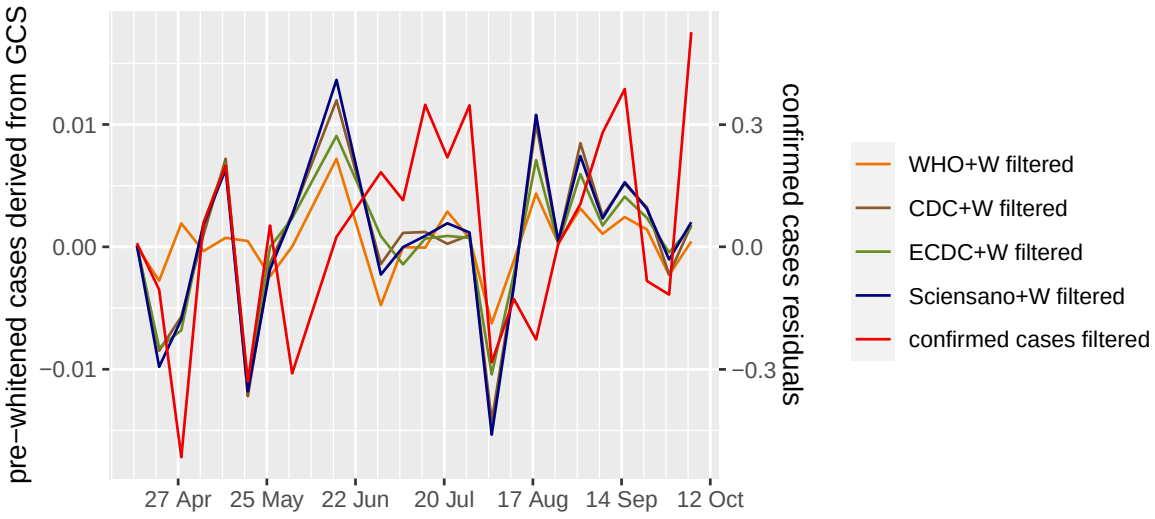

Figure 7: Time series pre-whitened by ARIMA (1,1,0)

Table 10: Correlations between ARIMA pre-whitened weighted and unweighted GCS derived case proportions and pre-whitened laboratory confirmed cases without lag, rounds 5–21

		WHO	CDC	ECDC	Sciensano	RKI
weighted	r	0.117	0.427	**0.479**	0.412	0.238
	95% CI lower bound	-0.301	0.028	0.093	0.010	-0.183
	95% CI upper bound	0.497	0.708	0.739	0.699	0.586
	p	0.587	0.038	0.018	0.046	0.262
unweighted	r	0.143	0.426	**0.453**	0.410	0.271
	95% CI lower bound	-0.277	0.027	0.061	0.008	-0.148
	95% CI upper bound	0.516	0.707	0.724	0.698	0.608
	p	0.506	0.038	0.026	0.046	0.200

and Sciensano, we see highest correlations for the 0 shift, which is closest to the time window for which symptoms were reported. In contrast to our observations in section 6.2.2, correlations decrease quite quickly when moving to the future, with no significant correlations after lag 0 for the CDC, ECDC, and Sciensano case definitions. This indicates that the high correlations observed for longer shifts for the unfiltered time series were largely due to the autocorrelation in the data. Even shifting one day into the post-survey period results in non-significant correlations for all case definitions. For the RKI case definition, on the other hand, correlations fluctuate between 0.175 and 0.410 with no apparent pattern and with only the correlation at the 6 day lag reaching significance. While this pattern was not expected, it is in line with a virtually non-decreasing correlation for the raw time series with the lags studied here.

Table 11: Autocorrelation estimates and diagnostics of ARIMA (1,1,0) model at different shifts.

	lag 0	1 day	2 days	3 days	4 days	5 days	6 days	1 week	1 week+1
AR1	0.736	0.805	0.806	0.749	0.737	0.752	0.714	0.770	0.822
(SE)	0.143	0.128	0.129	0.145	0.149	0.147	0.160	0.150	0.130
AIC	8.01	3.69	3.82	8.43	9.82	9.14	13.60	12.05	7.00
Ljung-Box p	0.865	0.547	0.522	0.708	0.911	0.927	0.468	0.929	0.614

Table 12: Correlations between ARIMA pre-whitened weighted and unweighted GCS derived case proportions and pre-whitened laboratory confirmed cases at different lags, rounds 5–21

	lag 0	1 day	2 days	3 days	4 days	5 days	6 days	1 week	1 week+1
WHO	**0.117**	0.096	0.065	0.054	0.044	0.013	-0.044	-0.009	-0.109
CDC	**0.427***	0.329	0.339	0.256	0.243	0.192	0.131	0.030	0.098
ECDC	**0.479***	0.373	0.396	0.294	0.279	0.228	0.140	0.024	0.094
Sciensano	**0.412***	0.325	0.340	0.279	0.266	0.217	0.140	0.050	0.085
RKI	0.238	0.356	0.201	0.129	0.193	0.175	**0.410***	0.325	0.297

Table 13: Cross tabulation of lab-confirmed cases and cases as derived from different case definitions for the subset of participants who reported a PCR test from rounds 20 and 21 of the GCS

		pos PCR	
		1	0
WHO	1	28	306
	0	15	1637

		pos PCR	
		1	0
CDC	1	36	592
	0	7	1351

		pos PCR	
		1	0
ECDC	1	35	475
	0	8	1468

		pos PCR	
		1	0
Sciensano	1	36	580
	0	7	1363

		pos PCR	
		1	0
RKI	1	35	754
	0	8	1189

6.4 Specificity and sensitivity of different case definitions

Table 13 gives the frequencies of true positives, false positives, false negatives and true negatives of GCS-derived cases using the different case definitions when compared to the result of a PCR test. From the subtables, specificity and sensitivity measures are derived and given in Table 14 and plotted in Figure 8. Unsurprisingly, the strict definition of WHO results in the lowest sensitivity, but highest specificity. CDC, ECDC and Sciensano again pattern very closely together, with relatively high sensitivity of above 80% and medium specificity of around 70%. Somewhat surprisingly, the RKI case definition does not result in higher sensitivity than these, with the exact same number of cases covered as the ECDC definition and even one less than Sciensano and CDC. Specificity, on the other hand, is lowest for the RKI definition. We can conclude that the use of this very broad definition only increases false positives, but not true positives when compared to ECDC, Sciensano and CDC case definitions and is hence strictly worse, at least in the sample we investigate here. Of course, there are only 43 positive cases in total, so potentially more subtle differences might not become apparent.

Table 14: Sensitivity and specificity of different case definitions with confidence interal, calculated on tested subset of rounds 20 and 21.

	Sensitivity	95% CI	Specificity	95% CI
WHO	0.651	[0.509; 0.794]	0.843	[0.859; 0.827]
CDC	0.837	[0.727; 0.947]	0.695	[0.675; 0.716]
ECDC	0.814	[0.698; 0.930]	0.756	[0.737; 0.775]
Sciensano	0.837	[0.727; 0.947]	0.701	[0.681; 0.721]
RKI	0.814	[0.698; 0.930]	0.612	[0.590; 0.634]

7 General Discussion

In general, the presented results show that longitudinal data from the citizen science project "De Grote Corona-Studie" can be used as an indicator of the ongoing COVID-19 epidemic in

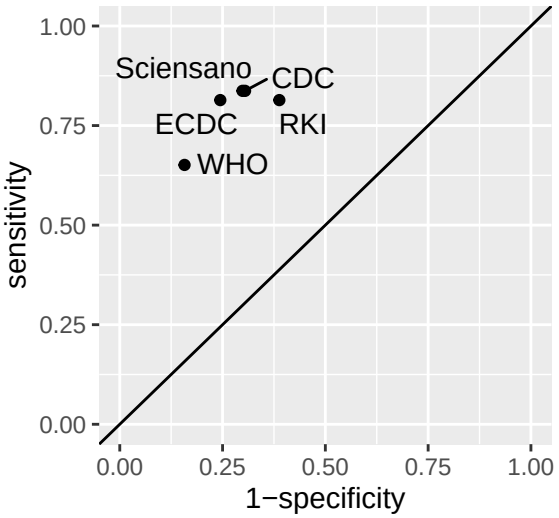

Figure 8: Sensitivity vs 1-Specificity for different case definitions for tested individuals in rounds 20 and 21

Belgium. In particular, strong correlations are observed between the percentage of cases derived with different case definitions and the summed laboratory confirmed cases in the week preceding the survey. After accounting for autocorrelation in the data, correlations decrease, but remain significant for three out of five case definitions. Correlations show different strengths for different gender and age groups and restricted ability to predict incidence rather than reproducing the current occurrence of infections. In the following, we will discuss the findings for the different case definitions in detail and compare the results again, before discussing possible shortcomings and limitations of the current analysis.

7.1 Sciensano, ECDC, and CDC

Since the results for the cases derived using the definitions by Sciensano, ECDC and CDC are so similar, we discuss these three together. As noted above, for Sciensano and CDC the case definitions themselves are very similar, such that the similar patterns were expected, while for ECDC this was less clear a priori. Sensitivity and specificity values are very similar and in general impressive, compared to the results for single symptoms and symptom combinations investigated in previous research (Struyf et al., 2021). Possibly, this is due to testing being largely dependent on the presence of symptoms. Given the values observed, we would likely conclude that the ECDC case definition is the best diagnostic test for COVID-19. Our aim, however, is not to diagnose patients, but to find a correlate for laboratory confirmed cases. For this task, the same case definition does not necessarily perform best. If we had a small data set, high sensitivity might be more important, while in a very large data set, a very specific test with low sensitivity might also be able to successfully detect trends.
In the correlation analyses, we see that all three case definitions lead to significant correlations to laboratory confirmed cases in the week preceding the survey. This is true for raw percentages, but also after accounting for common autocorrelation and irrespective of the use of weights. Overall, these three case definitions are thus a viable proxy for true COVID-19 cases and can be used to estimate concurrent incidence. We also see, however, that the correlation is reduced considerably for the male subgroup, where both the broader definition of RKI and the stricter definition of WHO perform better. It would be interesting to see whether specificity and sensitivity are also dependent on gender, but unfortunately we cannot investigate this further as our subset of laboratory tested individuals is fairly small. As it is, it is difficult to say whether the difference is observed because of chance and the male population being

smaller and hence providing less accurate estimates, or whether it is due to the difference of the clinical appearance of COVID-19 in males and females. Indeed, a relatively small study on differences between symptom burden between males and females showed that females show a higher prevalence for virtually all symptoms and most strongly for anosmia (Biadsee et al., 2020). If symptom burden is on average higher in women for COVID-19, we would indeed expect to correctly identify more cases with all case definitions for women compared to men. In this respect, however, it is also important to consider whether symptom burden for ILIs and other respiratory illnesses that tend to be confused with COVID-19 is also higher for women or not, because more false positives would weaken the correlation in turn.

For the different age groups, we see stable and high correlations for young adults, older adults, and seniors for all three definitions. In seniors, the correlation is highest, reaching up to 0.88 for the Sciensano definition. In teenagers and children, on the other hand, the correlations are considerably smaller and not reaching significance for the ECDC case definition. We have noted above already that this might be due to the small sample we have from this group, with only around 200 responses per survey round after round 12. Given that the 7-day incidence per 100,000 ranged between 5 and 200 in the period we investigate, we would therefore not necessarily expect any actual COVID-19 cases in our sample in this age group. An alternative line of thought could start with the observation that the case definitions are better suited to diagnosing adults than children, as children are known to remain more often asymptomatic than adults, but to show gastrointestinal symptoms like vomiting and diarrhea more often (Zare-Zardini et al., 2020). These symptoms are only considered in combination with other symptoms, if at all. Although this might well be the case, it is probably not an explanation for the low correlations we observed here, as even in the 0-19 age group, most participants were at least 17 and would thus not be counted as children in most studies.

Potentially due to the differences between age and gender groups, but possibly rather due to spatially varying incidence, the use of raking weights was found to be beneficial for correlations both for the raw time series, as well as for the pre-whitened time series, although the differences did not reach significance. This result is expected to carry over to other citizen science projects, whenever there is a severe imbalance of participants from different age and gender groups or from different regions.

With regard to the best lag to match confirmed cases to GCS-derived cases, our results are not entirely consistent. In the raw time series, we see higher correlation with confirmed cases in the one day ahead shifted time window, i.e., including five days before the survey date and one day after the survey date. Together with only slightly decreasing correlations for higher lags, it seemed that the GCS data can be predictive of laboratory-confirmed cases. The analysis of the pre-whitened time series, however, refutes this conclusion by only showing significant correlation for the zero-lag time window, which stretches from 6 days prior to the survey date until the survey date itself. Looking in some more detail at the ARIMA models used for pre-whitening, it becomes clear why the correlations for lags up to 8 days remain so high: the fitted ARIMA model shows an autocorrelation estimate of roughly 0.75. We can therefore conclude that the high correlations for lags of several days are not to be taken as indicative of the predictive nature of the GCS data, but rather as the result of high autocorrelation — here, the original confirmed cases data could equally well be used to predict cases in the future. The one point that remains unclear is why correlations are actually higher for a one-day shift when looking at the raw time series. As the difference is not big, we conjecture that this difference is not systematic and leave this point for further research.

In summary, the three case definitions presented in this section all exhibit the ability to estimate concurrent incidence, with only subtle differences between the different definitions. If a choice has to be made, it will be the ECDC definition, which has highest specificity and sensitivity and which shows the highest correlation with confirmed cases after accounting for autocorrelation.

7.2 WHO

In comparison to the three case definitions discussed in the previous section, the WHO case definition is stricter in that only the loss of smell or taste is a sufficient criterion, whereas cough, fever and other more general symptoms have to be encountered in combination. As a result, we see much lower sensitivity, but considerably higher specificity, as could be expected.
In the correlation analysis, we see in general good correlations when looking at the raw time series, while the correlation drops severely to a non-significant level when removing autocorrelation. Also, for seniors, as well as for teenagers and children, this case definition is not correlating highly. With regard to gender, on the other hand, we do not see the distinctive pattern described above, but indeed see a slightly higher correlation for males. We can only speculate why this is the case - possibly the more severe cases with multiple symptoms present are distributed equally across gender. The low performance in seniors further suggests that this age group tends to exhibit fewer symptoms in combination, although this seems counterintuitive given that seniors show more severe progressions in general.
While for the original time series correlations increased with increasing lags, this pattern did not persist after removing autocorrelation and hence has to be considered a spurious effect.
In conclusion, the WHO case definition seems to be less suited to estimate concurrent incidence compared to the definitions by ECDC, CDC, and Sciensano.

7.3 RKI

From all case definitions investigated here, the RKI case definition can be argued to be the widest one, including all respiratory symptoms in isolation and resulting in the highest rate of detected cases. On the other hand, more general symptoms like fever, headache, and fatigue are missing from this case definition. This combination leads to a considerable drop in specificity compared to the other case definitions, but also no higher sensitivity - at least not in the sample at hand. This already suggests that it is less suited as an indicator of incidence.
Looking at the raw time series, however, this prediction is not borne out, as it is on par with the set of three case definitions discussed above and even surpasses them at a one-day lag and when looking at the male subset as well as for the elderly. Only after correcting for autocorrelation, the correlation with the confirmed cases drops to a non-significant level. We can understand this finding, by thinking about the false positives produced by the RKI case definition: these will mostly be cases of common cold or ILI. In general, these illnesses show a strong seasonal effect and in particular a rise in early autumn, which resembles the emergence of the second wave for COVID-19. Here the importance of correcting for autocorrelation becomes apparent, only after removing these general trends, the specific ability of the RKI case definition becomes tractable and is found to be very low.
A possible advantage of the RKI definition, however, might be in predicting future incidence. In both the raw time series as well as the pre-whitened time series, we see comparatively high correlation for lags around 6 days, which is significant even in the pre-whitened data set. While this could, of course, also be a chance finding (note that we do not adjust the p-values for multiple testing), it is not unlikely that the more lenient case definition is better able to capture COVID-19 cases in the earliest stage, i.e., before individuals feel so sick they will consult a physician and potentially well before being administered a PCR test. Whether this advantage outweighs the disadvantages depends on how stable this effect is and what the goal of using the citizen science data is exactly.
In summary, the RKI case definition is too broad to be a good estimate of incidence, but might have a stronger predictive component than the other case definitions investigated here.

7.4 Limitations

While the results discussed above suggest that data from a citizen science project can be used to estimate the incidence of COVID-19 when using existing case definitions, there are some limitations that we acknowledge. Most importantly, it is not clear how our results carry over to future COVID-19 waves. In the time period we analyzed, there were various and changing non-pharmaceutical interventions in place, including school closures, closures of nonessential shops, but also social distancing measures and enhanced hygiene exercised by the general public. These measures and the increased alertness in the population reduced the occurrence of the common cold as well as influenza to a considerable degree. This has implications for the estimates of specificity as well as to correlations between GCS derived cases and confirmed cases. Since common cold and ILI, which are easily confused with COVID-19 based on the clinical profile, had a lower than usual prevalence, less false positives are expected to have occurred. In the situation where ILI and common cold rates would be similar to preceding years, a lot more false positives would be expected, hence reducing the specificity of all our case definitions, but potentially more for the more lenient ones. The correlations would then be expected to be much lower and susceptible to influenza outbreaks that are asynchronous to COVID-19 outbreaks. Whether any of our case definitions would be capable to successfully differentiate between COVID-19 and ILI, is an open question. Our qualitative result preferring the case definitions of medium rigor (ECDC, CDC or Sciensano) might also not hold in a post-pandemic state: Here more specific case definitions, as the WHO one, might fare much better, depending again on how well it differentiates between COVID-19 and other respiratory illnesses.
If we think one step further, however, anticipating the next pandemic caused by a novel pathogen, our insights might still be useful. Especially in the early phase of this pandemic, there were a lot of restrictive measures, reducing other illnesses, but at the same time there were not enough tests available to obtain realistic estimates of incidence. In the population, the willingness to invest time and participate in citizen science projects was enormous, providing a good and stable data basis. We suspect that in this early stage, results from citizen science projects could have been used and might have been more reliable than the official numbers of confirmed cases.
The analysis itself also has some drawbacks. Our main correlation analyses using pre-whitened time series rely partly on interpolated values. This was a compromise in order to use all available data, but also to respect the requirement of equal spacing when using ARIMA models. An alternative would have been to construct a consistently biweekly time series by discarding every second survey round from the first phase. With this method, however, we would not only have lost available data, but also shortened the time series so severely that ARIMA modeling would not have been appropriate any longer. This problem, of course, only occurred, because the survey was conceived and realized as a fast reaction to the evolving pandemic. If instead existing citizen science projects targeted at ILI incidence and instantiated in a regular manner would have been used, these problems would not have been present. The consequences of this choice are not so clear. On the one hand, using linear interpolation and a combination of differencing and autocorrelation for pre-whitening resulted in artificially low residuals, as the interpolation can be expected to be closer to the model than a real and potentially noisy data point. This might result in an increased correlation, but can equally well result in the opposite—depending on the missing value itself. Fortunately, this problem does not hold for the raw time series, where we only took the available data into account.

8 Conclusion

The present work aimed to investigate whether citizen science symptom burden data can be used to estimate incidence of COVID-19 and to evaluate which case definition from a set of

five is best suited for the task. We found robust correlations that persisted after removing autocorrelation and conclude that indeed this data source can be used to estimate concurrent COVID-19 incidence. The best performing case definitions were those of ECDC,CDC and Sciensano, while the stricter WHO case definition was less correlated and vulnerable to the removal of the autocorrelation. The case definition from RKI was also susceptible to removal of the autocorrelation, but was found to have a greater potential to predict incidence in the future. All results are to be treated with caution when applying to future scenarios, as the anti COVID-19 measures in place in Belgium reduced the prevalence of ILI and common cold, which are easily confused with COVID-19.

References

Benvenuto, Domenico, Marta Giovanetti, Lazzaro Vassallo, Silvia Angeletti, and Massimo Ciccozzi (2020). "Application of the ARIMA model on the COVID-2019 epidemic dataset". en. In: *Data in Brief* 29, p. 105340. URL: https://linkinghub.elsevier.com/retrieve/pii/S2352340920302341 (visited on 04/28/2021).

Biadsee, Ameen, Ameer Biadsee, Firas Kassem, Or Dagan, Shchada Masarwa, and Zeev Ormianer (2020). "Olfactory and Oral Manifestations of COVID-19: Sex-Related Symptoms—A Potential Pathway to Early Diagnosis". en. In: *Otolaryngology–Head and Neck Surgery* 163.4, pp. 722–728. URL: http://journals.sagepub.com/doi/10.1177/0194599820934380 (visited on 05/27/2021).

Bloom, R. M., D. L. Buckeridge, and K. E. Cheng (2007). "Finding Leading Indicators for Disease Outbreaks: Filtering, Cross-correlation, and Caveats". en. In: *Journal of the American Medical Informatics Association* 14.1, pp. 76–85. URL: https://academic.oup.com/jamia/article-lookup/doi/10.1197/jamia.M2178 (visited on 04/06/2021).

Carlson, Sandra J, Craig B Dalton, Frank A Tuyl, David N Durrheim, John Fejsa, David J Muscatello, JLynn Francis, and Edouard Tursan d'Espaignet (2009). "Flutracking surveillance: comparing 2007 New South Wales results with laboratory confirmed influenza notifications". In: *Communicable diseases intelligence quarterly report* 33.3, pp. 323–327.

Friesema, I.H.M., C.E. Koppeschaar, G.A. Donker, F. Dijkstra, S.P. van Noort, R. Smallenburg, W. van der Hoek, and M.A.B. van der Sande (2009). "Internet-based monitoring of influenza-like illness in the general population: Experience of five influenza seasons in the Netherlands". en. In: *Vaccine* 27.45, pp. 6353–6357. URL: https://linkinghub.elsevier.com/retrieve/pii/S0264410X09007464 (visited on 03/17/2021).

Giacomelli, Andrea, Laura Pezzati, Federico Conti, Dario Bernacchia, Matteo Siano, Letizia Oreni, Stefano Rusconi, Cristina Gervasoni, Anna Lisa Ridolfo, Giuliano Rizzardini, Spinello Antinori, and Massimo Galli (2020). "Self-reported Olfactory and Taste Disorders in Patients With Severe Acute Respiratory Coronavirus 2 Infection: A Cross-sectional Study". en. In: *Clinical Infectious Diseases* 71.15, pp. 889–890. URL: https://academic.oup.com/cid/article/71/15/889/5811989 (visited on 06/01/2021).

Helfenstein, Ulrich (1991). "The Use of Transfer Function Models, Intervention Analysis and Related Time Series Methods in Epidemiology". en. In: *International Journal of Epidemiology* 20.3, pp. 808–815. URL: https://academic.oup.com/ije/article-lookup/doi/10.1093/ije/20.3.808 (visited on 04/22/2021).

Huang, Chaolin et al. (2020). "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China". en. In: *The Lancet* 395.10223, pp. 497–506. URL: https://linkinghub.elsevier.com/retrieve/pii/S0140673620301835 (visited on 06/01/2021).

Hyndman, Rob J and Yeasmin Khandakar (2008). "Automatic time series forecasting: the forecast package for R". In: *Journal of Statistical Software* 26.3, pp. 1–22. URL: https://www.jstatsoft.org/article/view/v027i03.

Junninen, Heikki, Harri Niska, Kari Tuppurainen, Juhani Ruuskanen, and Mikko Kolehmainen (2004). "Methods for imputation of missing values in air quality data sets". en. In: *Atmospheric Environment* 38.18, pp. 2895–2907. URL: https://linkinghub.elsevier.com/retrieve/pii/S1352231004001815 (visited on 05/06/2021).

Menni, Cristina et al. (2020). "Real-time tracking of self-reported symptoms to predict potential COVID-19". en. In: *Nature Medicine* 26.7, pp. 1037–1040. URL: http://www.nature.com/articles/s41591-020-0916-2 (visited on 04/28/2021).

Noort, Sander P. van, Cláudia T. Codeço, Carl E. Koppeschaar, Marc van Ranst, Daniela Paolotti, and M. Gabriela M. Gomes (2015). "Ten-year performance of Influenzanet: ILI time series, risks, vaccine effects, and care-seeking behaviour". en. In: *Epidemics* 13, pp. 28–36. (Visited on 03/17/2021).

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rasmussen, I. S., L. H. Mortensen, T. G. Krause, and A-M. Nybo Andersen (2019). "The association between seasonal influenza-like illness cases and foetal death: a time series analysis". en. In: *Epidemiology and Infection* 147, e61. URL: https://www.cambridge.org/core/product/identifier/S0950268818003254/type/journal_article (visited on 05/05/2021).

Richard, Aude, Laura Müller, Ania Wisniak, Amaury Thiabaud, Thibaut Merle, Damien Dietrich, Daniela Paolotti, Emilien Jeannot, and Antoine Flahault (2020). "Grippenet: A New Tool for the Monitoring, Risk-Factor and Vaccination Coverage Analysis of Influenza-Like Illness in Switzerland". en. In: *Vaccines* 8.3, p. 343. URL: https://www.mdpi.com/2076-393X/8/3/343 (visited on 03/17/2021).

Salles, Rebecca, Eduardo Bezerra, Jorge Soares, and Eduardo Ogasawara (2015). "Evaluating Linear Models as a Baseline for Time Series Imputation". en. In: p. 6.

Steiger, James H (1980). "Tests for comparing elements of a correlation matrix." In: *Psychological bulletin* 87.2, p. 245.

Struyf, Thomas et al. (2021). "Signs and symptoms to determine if a patient presenting in primary care or hospital outpatient settings has COVID-19". en. In: *Cochrane Database of Systematic Reviews*. Ed. by Cochrane Infectious Diseases Group. URL: http://doi.wiley.com/10.1002/14651858.CD013665.pub2 (visited on 03/22/2021).

Vandendijck, Yannick, Christel Faes, and Niel Hens (2013). "Eight Years of the Great Influenza Survey to Monitor Influenza-Like Illness in Flanders". en. In: *PLoS ONE* 8.5. Ed. by Jodie McVernon, e64156. URL: https://dx.plos.org/10.1371/journal.pone.0064156 (visited on 05/04/2021).

Warton, David I. and Francis K. C. Hui (2011). "The arcsine is asinine: the analysis of proportions in ecology". en. In: *Ecology* 92.1, pp. 3–10. URL: http://doi.wiley.com/10.1890/10-0340.1 (visited on 05/09/2021).

Williams, Evan J (1959). "The comparison of regression variables". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 21.2, pp. 396–399.

Zare-Zardini, Hadi, Hossein Soltaninejad, Farzad Ferdosian, Amir Ali Hamidieh, and Mina Memarpoor-Yazdi (2020). "Coronavirus Disease 2019 (COVID-19) in Children: Prevalence, Diagnosis, Clinical Symptoms, and Treatment". en. In: *International Journal of General Medicine* Volume 13, pp. 477–482. URL: https://www.dovepress.com/coronavirus-disease-2019-covid-19-in-children-prevalence-diagnosis-cli-peer-reviewed-article-IJGM (visited on 05/27/2021).