Received: 17 June 2021 | Revised: 6 May 2022 | Accepted: 21 May 2022

DOI: 10.1002/bimj.202100186

Biometrical Journal

RESEARCH ARTICLE

A spatial model to jointly analyze self-reported survey data of COVID-19 symptoms and official COVID-19 incidence data

Maren Vranckx[1] | **Christel Faes**[1] | **Geert Molenberghs**[1,2] | **Niel Hens**[1,3] | **Philippe Beutels**[3] | **Pierre Van Damme**[3] | **Jan Aerts**[1] | **Oana Petrof**[1] | **Koen Pepermans**[4] | **Thomas Neyens**[1,2]

[1]I-BioStat, Data Science Institute, Hasselt University, Hasselt, Belgium

[2]L-BioStat, Department of Public Health and Primary Care, Faculty of Medicine, KU Leuven, Leuven, Belgium

[3]Center for Health Economics Research and Modeling Infectious Diseases, Vaccine and Infectious Disease Institute, University of Antwerp, Antwerp, Belgium

[4]Faculty of Social Sciences, University of Antwerp, Antwerp, Belgium

Correspondence
Hasselt University, Inter-university Institute for Biostatistics and Statistical Bioinformatics, Data Science Institute, Martelarenlaan 42, 3500 Hasselt, Belgium; KU Leuven, Leuven Biostatistics and Statistical Bioinformatics Centre, Kapucijnenvoer 35, 3000 Leuven, Belgium.
Email: thomas.neyens@uhasselt.be, thomas.neyens@kuleuven.be

Funding information
Horizon 2020 Framework Programme, Grant/Award Number: 101003688; Fonds Wetenschappelijk Onderzoek, Grant/Award Numbers: G0G1920N, G0G9820N; Internal Funds KU Leuven, Grant/Award Number: 3M190682; Vlaamse regering, Grant/Award Number: AI Research Program

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract

This work presents a joint spatial modeling framework to improve estimation of the spatial distribution of the latent COVID-19 incidence in Belgium, based on test-confirmed COVID-19 cases and crowd-sourced symptoms data as reported in a large-scale online survey. Correction is envisioned for stochastic dependence between the survey's response rate and spatial COVID-19 incidence, commonly known as preferential sampling, but not found significant. Results show that an online survey can provide valuable auxiliary data to optimize spatial COVID-19 incidence estimation based on confirmed cases in situations with limited testing capacity. Furthermore, it is shown that an online survey on COVID-19 symptoms with a sufficiently large sample size per spatial entity is capable of pinpointing the same locations that appear as test-confirmed clusters, approximately 1 week earlier. We conclude that a large-scale online study provides an inexpensive and flexible method to collect timely information of an epidemic during its early phase, which can be used by policy makers in an early phase of an epidemic and in conjunction with other monitoring systems.

KEYWORDS
bivariate conditional autoregressive random effect, COVID-19, disease mapping, preferential sampling, survey data

1 | INTRODUCTION

COVID-19, the respiratory disease caused by the betacoronavirus SARS-CoV-2, was first observed in Wuhan, the capital of the Hubei province in the China, in December 2019. It has quickly spread across continents and has been declared a global pandemic on March 11, 2020 (WHO, 2020). The first confirmed COVID-19 patient in Belgium was a Belgian national who was repatriated from the Hubei province in early February. As the person was quarantined immediately after repatriation, he very likely did not infect other individuals within Belgian borders. The second known case was confirmed on March 1, 2020, the date that is commonly referred to as the start of the epidemic in Belgium. As COVID-19 incidences increased, a number of federal measures were taken to curtail community transmission, including the implementation of a semilockdown of the country on March 18, 2020.

Limited available resources for testing, in terms of both chemical supplies and manpower, posed a difficulty during the early phase of the Belgian outbreak. As a result, only patients with severe symptoms of COVID-19 and health care workers were tested. From May 2020, the testing strategy was adapted, in which all symptomatic cases were tested to enhance contact tracing (Desson et al., 2020). As a result, the limitations in the testing strategy during the initial phase of the outbreak have hampered the understanding of the spatial spread of the virus. In addition, the reported number of confirmed cases is an underestimation of the number of infected cases, given that many patients are asymptomatic and paucisymptomatic patients who may not seek health care.

In order to spatially model COVID-19 incidence during the March–April 2020 wave, Neyens et al. (2020) applied standard spatial statistical methods to aid the analysis of the suboptimal test-confirmed cases' data by using auxiliary self-reported data on COVID-19 symptoms that were collected during the third round of the Big Corona Study (henceforth, BCS). The BCS (Corona study 2020) is an open online survey and an initiative of the University of Antwerp, in collaboration with Hasselt University, KU Leuven, and the Free University of Brussels. Since March 17, 2020, the study has collected weekly (biweekly from June 2, 2020, onward) data on COVID-19-related themes through self-reporting by the general Belgian public, including the incidence of COVID-19 symptoms within communities. The response rates have been high, with approximately 5% of the Belgian population (537, 172 individuals) participating during its first round on March 17, 2020.

Neyens et al. (2020) used data of all 397, 131 respondents (approximately 3.5% of the Belgian population) during the third round of the BCS (March 31, 2020) to estimate symptoms incidence via a spatially discrete binomial Leroux model (Leroux et al., 1999). These estimated incidences were subsequently used as predictors in a Poisson Leroux model of confirmed COVID-19 cases, as reported by the Belgian government between April 7 and 9, 2020, that is, approximately 7–9 days later than the BCS's third round. The investigators reported, among other things, that the model-based spatial symptoms' incidences have a significant, but limited, predictive performance, when used as an explanatory variable in a spatial analysis of test-confirmed COVID-19 cases diagnosed 7–9 days later.

The limited prediction capabilities of the model presented in Neyens et al. (2020) are possibly explained by data misalignment between the test-confirmed and symptoms data: (i) the test-confirmed data might overrepresent local outbreaks in nursing homes. These could have distorted the epidemiological signal of the general population, since limited test capacity has led to primarily administering COVID-19 tests to those with a severe pathology, which occurs more frequently among elderly; (ii) the BCS underrepresents elderly and nonadults. This is not necessarily a problem in a comparative study, but it might lead to biased results when only very little signals can be obtained from these age groups. An efficient and straightforward solution is to restrict the analysis to the working population.

Although the method of Neyens et al. (2020) provides a straightforward strategy to pool information from the spatial incidence of COVID-19-like symptoms to improve the estimation of the spatial distribution of test-confirmed incidence, an important disadvantage of this method is that it does not correct for uncertainty in the model-based predictions of symptoms' prevalences. An extension is to model both outcomes simultaneously, for example, by using a random-effects structure that allows both processes to be correlated. The joint analysis of both processes can improve the predictive inference of COVID-19 incidence risk. We refer to Neyens et al. (2016) for an overview of bivariate modeling approaches in a disease mapping context.

As the BCS is a voluntary survey, response rates vary among citizens with different ages, genders, and socioeconomic backgrounds. In particular, the data might be preferentially sampled Diggle et al. (2010), a phenomenon that reflects processes that generate spatial response rates that are stochastically dependent on the spatial process under investigation. In the BCS, the geographical distribution of public engagement to participate in the BCS might reflect a similar mechanism. Geographical tendencies in response rates can depend on several unmeasured processes that are associated with increased COVID-19 symptoms' incidence. When left unaccounted for, preferential sampling can invalidate predictive inferences.

Diggle et al. (2010) propose a joint modeling approach where a spatial model of the outcome of interest shares a spatial random effect with a point process model, which corrects for the choice of sampling locations.

In this study, we propose a joint modeling framework to improve the estimation of the spatial distribution of the latent COVID-19 incidence. In essence, a model is proposed that simultaneously analyzes three spatial processes: (i) the reported incidence of COVID-19, based on test-confirmed cases, the process of main interest; (ii) the incidence of COVID-19-like symptoms, as obtained by self-reporting by Belgian citizens; and (iii) the geographical distribution of participants of the BCS. We apply Bayesian estimation to fit this model. Processes (i) and (ii) are linked via a bivariate random-effects structure. Processes (ii) and (iii) share a conditional autoregressive (CAR) random-effects term. To the best of our knowledge, this is the first study that proposes the extension of a joint model that accounts for preferential sampling toward the context where two correlated disease outcomes are simultaneously analyzed. We extend the modeling framework of Neyens et al. (2020) considerably: (i) by correcting the causal pathway via modeling incidences of symptoms and test-confirmed cases simultaneously in a joint modeling framework; (ii) by accounting for sampling bias due to preferential sampling in self-reported survey data; and (iii) by optimizing the alignment between the data of the self-reported symptoms and test-confirmed cases by focusing on the working-age population.

2 | METHODOLOGY

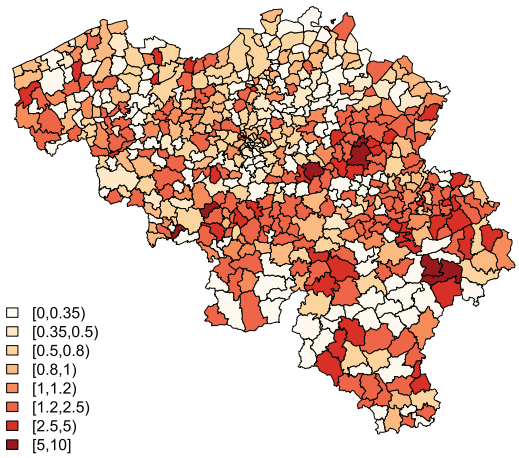
Two data sources collected during the initial phase of the epidemic, March and April, 2020, are used: (i) reported test-confirmed cases in Belgium and (ii) self-reported symptoms data from the BCS, an online survey.

The first data set is provided by the Belgian public health institute (Sciensano). It contains 2124 confirmed COVID-19 cases between the age of 25 and 64 diagnosed between April 7 and April 9, 2020. The data include for each diagnosed case the residential information (municipality), age, and gender. The number of confirmed COVID-19 cases in municipality i ($i = 1, \ldots, 589$) is denoted as Y_{1i}. We perform an internal age-gender standardization (Waller & Gotway, 2004) to calculate the municipality-specific expected number of cases, E_i. This internal age-gender standardization uses an age/gender-specific incidence rate (p_g) for a standard population (here, the Belgian population), computed as the total observed number of cases in the age/gender-group g divided by the age/gender-specific standard population number. The expected number is then calculated as the sum over the different age/gender-strata of the age/gender-specific incidence rate multiplied with the population number in municipality i ($n_{g,i}$), that is, $E_i = \sum_g p_g n_{g,i}$. The standardization procedure uses age intervals that allow for optimal balance in the BCS data, as detailed in the following paragraph, namely, those containing 25–44, and 45–64, respectively. The upper left panel of Figure 1 depicts then the standardized incidence rates, $SIR_i = Y_{1i}/E_i$.

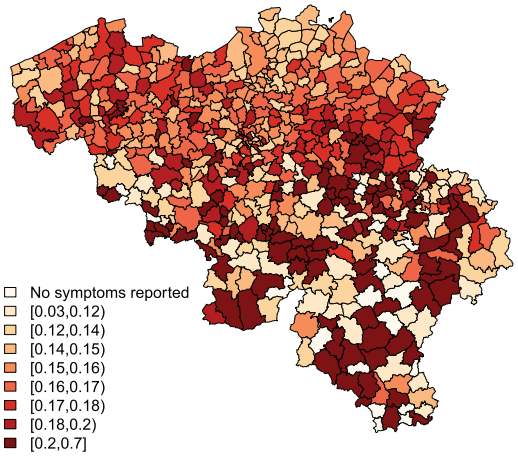
The second data set contains crowd-sourced COVID-19 symptoms, self-reported during the third round of the BCS (March 31, 2020). This survey is designed to collect data on epidemiological, socioeconomic, behavioral, and psychological trends in the Belgian population during the SARS-CoV-2 outbreak of 2020. The majority of the respondents comes from Flanders, the northern Dutch-speaking part of Belgium (Figure 1, bottom left). Data of the third round of the survey are investigated, since we are interested in the spatial distribution of the COVID-19 incidence in the early phase of the Belgian epidemic. Furthermore, the number of participants between the age of 25 and 64 of the BCS in the third round is 320,463.

Although the survey was also available in French, English, and German, the BCS, being an initiative of the University of Antwerp, was most intensively promoted in Dutch speaking Flanders, and in particular in the province of Antwerp, which explains the largest response rates in the Antwerp region. All participants were asked to indicate which of the following COVID-19-like symptoms they experienced during the week preceding the third round of the survey (March 24–30, 2020), if any: (i) a rapidly increasing fever, (ii) a high fever, (iii) a dry cough, (iv) shortness of breath, (v) chest pain, (vi) muscle pain, (vii) exhaustion, (viii) chills, (ix) nausea, (x) painful eyes, (xi) a sore throat, (xii) a rattling cough, and/or (xiii) a running nose. We denote $Y_{2ij} = 1$, when person $j = 1, \ldots, n_i$ in municipality i experienced at least one of the most common symptoms (according to WHO, 2020; Yang et al., 2020), which we define as symptoms (i)–(iv); otherwise, $Y_{2ij} = 0$. The observed proportion of participants reporting at least one of these key symptoms is depicted in Figure 1 (top right).

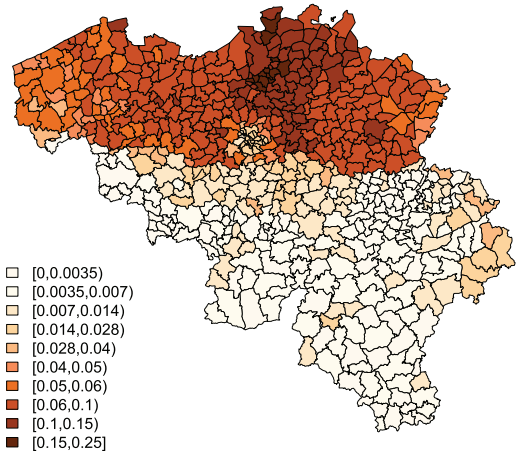
The choice of a time difference between the symptoms' prevalence (situation before March 31, 2020) and COVID-19 cases (diagnosis on April 7, 8, or 9) reflects the delay between symptoms and testing. Neyens et al. (2020) investigated multiple delay periods and came to the conclusion that delay times of 7–9 days provided the best predictive results, while not being drastically better than slightly shorter or longer delay periods. Due to computational limitations related to our modeling approach (Section 2.1), we have opted to focus on this particular recommended delay period.



(a) Standardized incidence rates (SIR)



(b) Proportion of respondents reporting symptoms



(c) Proportion respondents

	Proportion	Number
Minimum	0.09	1
Mean	4.67	544
Median	3.97	240
Maximum	22.35	31,081

(d) Summary statistics of respondents per municipality

FIGURE 1 Top left: SIR map of the number of confirmed COVID-19 cases in the age interval 25–64 between April 7 and April 9, 2020, per municipality, as reported by the Belgian population health institute. Top right: the proportion of respondents reporting at least one COVID-19 like symptoms, as reported by 320,463 respondents between the age of 25 and 64 during the third round of the Big Corona Study on March 31, 2020. Bottom left: the proportion of the population between the age of 25 and 64 per municipality taking the survey on March 31, 2020. Bottom right: summary statistics of the proportion (%) and the total number of the respondents between the age of 25 and 64 taking the survey on March 31, 2020

2.1 | Statistical model

A joint model is fitted that consists of three submodels that simultaneously analyze test-confirmed cases, crowd-sourced symptoms based on the BCS, and the response rates in the BCS. We define the model as follows:

$$
\begin{aligned}
Y_{1i} &\sim \text{Poisson}(E_i\theta_i),\\
Y_{2ij} &\sim \text{Bernoulli}(\pi_{ij}),\\
n_i &\sim \text{Poisson}(E_{n,i}\psi_i),\\
\theta_i &= \exp(\alpha_0 + \upsilon_{1i} + u_{1i}),\\
\pi_{ij} &= \text{expit}(\beta_0 + \beta_1 A_j + \beta_2 G_j + \beta_3 A_j G_j + \upsilon_{2i} + u_{2i} + \delta u_{p,i}),\\
\psi_i &= \exp(\gamma_0 + \gamma_1 R_{1i} + \gamma_2 R_{2i} + \upsilon_{n,i} + u_{p,i}),
\end{aligned}
\tag{1}
$$

where θ_i denotes the relative risk of confirmed COVID-19 cases for municipality i and $E_{n,i}$ represents the expected number of respondents in municipality i, which is calculated via the same internal age-gender standardization as was applied to obtain E_i. We will refer to ψ_i as the *response factor*, a multiplicative parameter representing the surplus or reduced number of respondents, according to what one would expect based on the general response rate in the whole population. Predictions of π_{ij} are based on the total Belgian population and not on the survey population. Parameters α_0, β_0, and γ_0 represent process-specific intercepts. $A_j = 1$ for people in the age interval 45–64 years, 0 for the age interval 25–44 years; $G_j = 1$ for males, 0 for females. The dummy variables R_1 and R_2 represent whether the BCS respondents reside in, respectively, the Brussel Capital Region or the Walloon Region, with the Flemish Region as the reference group.

The response factor is modeled using two random effects. One is the area-specific parameter $v_{n,i}$, $v_{n,i} \sim N(0, \sigma_n^2)$, denoting small-scale variation that is specific to the process that generates the response rate. The second area-specific term $u_{p,i}$ corrects for possible preferential sampling, which is shared between the linear predictor of the symptoms' incidence and that of the response rate. We parameterize it as an intrinsic CAR random-effects term, such as introduced by Besag and Kooperberg (1995):

$$\begin{aligned} u_{p,i} | \boldsymbol{u}_{\boldsymbol{p},-\boldsymbol{i}} &\sim N\left(\bar{\mu}_{p,i}, \tau_{p,i}^2\right), \\ \bar{\mu}_{p,i} &= \frac{1}{m_i} \sum_{k \sim i} u_{p,k}, \\ \tau_{p,i}^2 &= \frac{\tau_p^2}{m_i}, \end{aligned} \tag{2}$$

where $k \sim i$ indicates two adjacent areas and m_i are the number of neighbors of area i. The parameter δ estimates the amount of preferential sampling, based on recommendations by Pati et al. (2011), who extended the preferential sampling model of Diggle et al. (2010).

To quantify the stochastic correlation between the COVID-19 confirmed cases and the crowd-sourced symptoms, we use correlated random effects $\boldsymbol{v_i}=(v_{1i}, v_{2i})^T$ and/or $\boldsymbol{u_i}=(u_{1i}, u_{2i})^T$. Three model parameterizations were considered.

Option 1 assumes spatial correlation on a multivariate scale. In this case, a bivariate intrinsic CAR random effect (Jin et al., 2007; Martinez-Beneito & Botella-Rocamora, 2019) is defined as

$$\boldsymbol{u_i} \mid \boldsymbol{u}_{-\boldsymbol{i}}, \Omega \sim \mathcal{N}_2\left(\frac{1}{m_i} \sum_{k \sim i} \boldsymbol{u_k^T}, \frac{1}{m_i}\Omega\right), \tag{3}$$

where

$$\Omega = \begin{pmatrix} \tau_1^2 & \rho_s \tau_1 \tau_2 \\ \rho_s \tau_1 \tau_2 & \tau_2^2 \end{pmatrix}. \tag{4}$$

The diagonal terms of the covariance matrix Ω represent the conditional dependence between neighboring areas for either the confirmed cases or the COVID-19-like symptoms, while the off-diagonal terms represent the spatial conditional dependence between the confirmed cases and symptoms. In addition, v_{1i} and v_{2i} capture small-scale spatial variation and are univariately normally distributed, $v_{ji} \sim \mathcal{N}(0, \sigma_j^2)$, $j = 1, 2$.

Option 2 shares a bivariate normally distributed small-scale heterogeneity term, defined as

$$\boldsymbol{v_i} \sim \mathcal{N}_2(0, \Sigma), \tag{5}$$

where

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho_u \sigma_1 \sigma_2 \\ \rho_u \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix}. \tag{6}$$

In addition, the processes generating the confirmed cases and the symptoms include process-specific intrinsic CAR random effects, u_{1i} and u_{2i}, defined similarly as (2).

Option 3 uses a multivariate CAR convolution model, which combines a bivariate intrinsic CAR random effect, as defined in (3), and a bivariate normally distributed heterogeneity term, given by (5).

Option 4 models $\boldsymbol{u}$ as a bivariate intrinsic CAR random effect (given by Equation 3) and v_{2i} as univariate normally distribution, similar as option 1. However, the small-scale heterogeneity of the confirmed COVID-19 cases, which is represented by the parameter v_{1i}, is modeled via a gamma distribution instead of a lognormal distribution. The relative risk of the confirmed cases is then defined as

$$\theta_i = v_{1i} \exp(\alpha_0 + u_{1i}), \tag{7}$$

with $v_{1i} \sim \text{gamma}(a, a)$.

A gamma distribution for the confirmed cases is opted for since flexible alternatives to the Gaussian distribution are frequently preferred when modeling COVID-19 incidence data. Schumacher et al. (2020) uses, for example, mixed-effect models for COVID-19 deaths and Mamode Khan et al. (2020) considers Conway–Maxwell Poisson terms. Option 4, therefore, leads to an extension of the combined model, which was introduced by Molenberghs et al. (2010) and adapted to a spatial setting in Neyens et al. (2012).

2.2 | Analysis procedure

R 3.6.0 (R Core Team, 2019) was used to perform the analysis, via the R package NIMBLE (Numerical Inference for statistical Models for Bayesian and Likelihood Estimation, NIMBLE Development Team, 2019; de Valpine et al., 2017). NIMBLE applies Markov chain Monte Carlo (MCMC) techniques. Lawson (2020) gives NIMBLE code for Bayesian disease mapping. We provide the codes to fit all models as supplementary material. Additional information regarding fitting the combined model can be found in Neyens et al. (2012).

An $N(0, 10^3)$ prior is used for the process-specific intercepts and covariate coefficients. For the variance parameters of the random-effects terms, a gamma(0.01, 0.01) hyperprior is assigned. For the hyperparameter a of option 4, an exp(1) distribution is taken. A Wishart prior with 2 degrees of freedom is used for the covariance matrix of the bivariate random-effects structure. The scaled matrix was defined as $\begin{pmatrix} 0.1 & 0 \\ 0 & 0.1 \end{pmatrix}$. A sensitivity analysis of the prior choices has been added as Appendix B. We denote covariate effects as *significant*, when their associated 95% credible interval does not include 0.

Note that extensive modeling indicated that results of this three-way joint model can be moderately sensitive to the initial parameter values. Therefore, parameter estimates of three separate model fits on each process, that is, without correlated or shared random effects, are used as starting values for the MCMC sampling of the covariates. For convergence purposes, two MCMC chains are generated, with a burn-in of 15,000 iterations and 35,000 additional iterations per chain.

3 | RESULTS

Model selection was conducted to select the correlated random effects structure between the confirmed cases' and the crowd-sourced symptoms' processes, using WAIC (Watanabe–Akaike Information Criterion, Watanabe, 2010). The WAIC estimate was the lowest for option 4 (WAIC 17,401.38), as compared to option 1 (WAIC 17,414.68), option 2 (WAIC 17,426.42), and option 3 (WAIC 17,428.63). Moreover, we investigated which random-effects structure performs better when prediction is the objective using the LMPL (log marginal predictive likelihood; Geisser & Eddy, 1979). Only very small differences between the LMPL estimates of the different options were noticed, which were too small to prefer a single structure (Vranckx et al., 2021). Hence, based on the WAIC, this section gives the results of the model where the processes generating the confirmed cases and symptoms are linked via a bivariate spatial correlation, in addition to a univariate normal uncorrelated heterogeneity term for the symptoms' processes and a gamma uncorrelated heterogeneity term for the confirmed cases' process. An overview of the results for options 1, 2, and 3 is given in Appendix A.

We find significant stochastic dependency between the COVID-19 confirmed cases' and crowd-sourced symptoms' processes (Table 1, $\hat{\rho}_s = 0.5421$). Furthermore, both the crowd-sourced symptoms' process and the test-confirmed process have area-specific variance estimates that are considerably large as compared to the unstructured heterogeneity. No

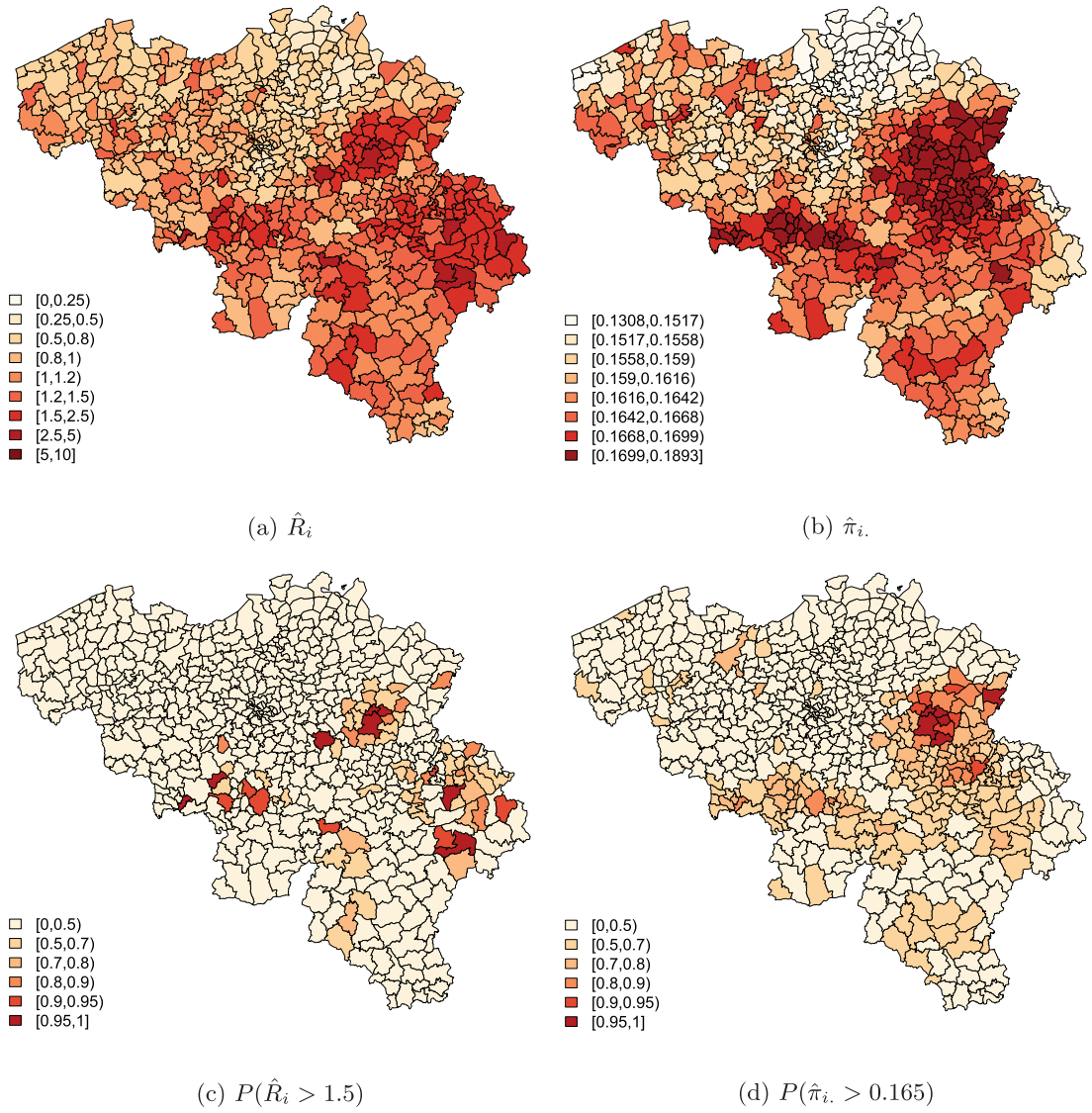
TABLE 1 Fit of the joint model with a bivariate spatial random effect, and a univariate gamma and normal small-scale heterogeneity terms for the combined test-confirmed cases, respectively, crowd-sourced symptoms processes (option 4)

Effect	Parameter	Estimate (posterior median)	95% equal tail credible interval
Test-confirmed COVID-19 process			
Intercept	α_0	−0.0190	(−0.0828, 0.0431)
Area-specific variance	τ_1^2	0.3147	(0.1810, 0.5161)
Unstructured variance	σ_1^2	0.1268	(0.0848, 0.1867)
Test-confirmed COVID-19 and crowd-sourced symptoms processes			
Spatial correlation	ρ_s	0.5421	(0.2206, 0.7561)
Crowd-sourced symptoms process			
Intercept	β_0	−1.4786	(−1.4985, −1.4543)
agecat	β_1	−0.3162	(−0.3385, −0.2911)
male	β_2	−0.0339	−0.0588, −0.0084)
agecat ∗ male	β_3	0.0070	(−0.0393, 0.0453)
Area-specific variance	τ_2^2	0.0108	(0.0069, 0.0170)
Unstructured variance	σ_2^2	0.0019	(0.0010, 0.0033)
Preferential sampling			
Preferential sampling	δ	−0.0319	(−0.0771, 0.0116)
Area-specific variance	τ_p^2	0.4839	(0.4052, 0.5688)
Responses rate process			
Intercept	γ_0	−0.0284	(−0.0945, 0.0439)
regioncat_1	γ_1	−1.0641	(−1.3903, −0.7898)
regioncat_2	γ_2	−1.7345	(−1.8962, −1.5892)
Unstructured variance	σ_n^2	0.0117	(0.0047, 0.0248)

evidence of preferential sampling is found, as indicated by a nonsignificant parameter δ. Therefore, a joint model of the COVID-19 confirmed cases' process and crowd-sourced symptoms' process will not lead to biased predictions due to informative sampling. Note that, although we do not find stochastic association between the response rates and the symptoms' incidence, individuals residing in Brussels or in the Walloon region are significantly associated with a lower response rate in the BCS, as compared to Flemish citizens. This is arguably due to the fact that the survey was mostly carried out in Flanders, as we see in Figure 1.

The variables age and gender have significant effects, while the interaction effect between age and gender is not significant. Leaving out the interaction term gives similar model results which shows robustness in terms of model fit, and a very small difference in WAIC estimates is observed, leading to no clear support for the model with or without the interaction term (Vranckx et al., 2021). Therefore, we here present the model with the interaction term between age and gender due to the primary purpose of using the age and gender variable as confounding variables in this epidemiological process. Women have the highest probability of experiencing at least one typical COVID-19 symptom; this probability is also increased in the age interval between 25 and 44 years. This can be explained by age-dependent variation in social distancing behavior, as was suggested by analyses of the adherence to measures against disease transmission, based on the BCS.

Figure 2 (top left) shows the relative risk of test-confirmed cases, the spatial distribution of the probability of experiencing at least one typical COVID-19 symptom (top right), as well as the exceedance probabilities (bottom). For the test-confirmed cases, the exceedance probabilities are defined as $P(\hat{\theta}_i > 1.5)$. The threshold 1.5, which represents an important increased risk, is chosen in line with choices in Neyens et al. (2020), for consistency purposes. However, the method is flexible to change this threshold when necessary. For the BCS-based symptoms' incidences, the exceedance probabilities are set as $P(\hat{\pi}_{i.} > 0.165)$, where $\hat{\pi}_{i.}$ represents the symptom's incidence in the general population of each municipality. The threshold value 0.165 is the mean proportion of respondents reporting at least one COVID-19-like symptom. Based on both the relative risks of confirmed COVID-19 cases and the predictions of the probability of reporting at least one of the typical COVID-19 symptoms in the BCS, a main cluster of increased COVID-19 incidence is observed around the area of



(a) $\hat{R}_i$ (b) $\hat{\pi}_{i.}$

(c) $P(\hat{R}_i > 1.5)$ (d) $P(\hat{\pi}_{i.} > 0.165)$

FIGURE 2 Top left: the relative risk predictions of confirmed COVID-19 cases. Top right: the predicted probability of having at least one COVID-19 symptoms. Bottom left: $P(\hat{R}_i > 1.5)$; bottom right: $P(\hat{\pi}_{i.} > 0.165)$. These maps are all based on the joint model with a bivariate spatial random effect, and a univariate gamma and normal small-scale heterogeneity terms for the combined test-confirmed cases, respectively, crowd-sourced symptoms processes (option 4)

Sint-Truiden, which is situated in the central-east part of Belgium. Note that this region is well-known for its large COVID-19 outbreak during March and April 2020. Compared to Neyens et al. (2020), fewer local outbreaks appear in the relative risk predictions based on the test-confirmed COVID-19 process. Furthermore, in the Walloon region, that is, the southern part of Belgium, there is more discrepancy between the maps of the confirmed COVID-19 cases and crowd-sourced symptoms processes, most likely due increased uncertainty as a result of generally low response rates in the BCS throughout that region.

4 | DISCUSSION

In this manuscript, a joint model for the spatial analysis of COVID-19 incidence in the working-age population that accounts for preferential sampling is presented. Different bivariate random-effects structures were considered to link the confirmed COVID-19 cases process to the crowd-sourced symptoms' incidence process. Model-based symptoms' incidence predictions that are obtained in the BCS are moderate significantly associated with the heterogeneity in the relative risks

of the confirmed cases that are reported approximately 1 week to 9 days after reporting ongoing symptoms. In addition, we find little evidence of preferential sampling in reporting symptoms in the BCS.

Our results highlight, to a substantially larger extent than the results in Neyens et al. (2020), the opportunities of a large-scale online survey to signal incidence trends with a lead time of a week or longer. This is illustrated by the overlap between the test-confirmed risk map and the symptoms-based incidence map (Figure 2), where considerably less anomalies are observed in the relative risk predictions based on the test-confirmed data, as compared to the analyses of Neyens et al. (2020). This can be explained by the difference in age category considered; Neyens et al. (2020) studied the whole population, while in this manuscript only the working-age population is considered. By not including elderly, local outbreaks in nursing homes are not accounted for in the confirmed COVID-19 cases' process, which provides improved insights in the spatial dynamics of the epidemic. A number of conclusions are similar to those obtained by Neyens et al. (2020). First of all, our analysis yields similar insights in the demographic heterogeneity in key COVID-19 symptoms. Furthermore, we find a modest, but significant, stochastic dependency between the relative risks of the COVID-19 confirmed cases and the incidences of the crowd-sourced symptoms, which is in line with the moderately significant covariate effects of predicted symptoms' incidences on test-confirmed relative risks. Both analyses also pinpoint the same important cluster of elevated COVID-19 risk around the city of Sint-Truiden, in the central-east part of Belgium.

From a methodological perspective, the statistical model considered in this manuscript improves the analysis of Neyens et al. (2020) in two aspects: (i) the two-step approach of Neyens et al. (2020), who used the predictions of the crowd-sourced symptoms as a covariate for the confirmed COVID-19 cases, is not optimal, since it does not account for uncertainty in the symptoms predictions. By jointly analyzing these processes, we overcome this problem; (ii) a joint model yields an estimate of the correlated and the processes-specific variability on a spatial and nonspatial level. In addition, the model considered in this manuscript also corrects for possible bias due to preferential sampling. Although no significant preferential sampling was found, Diggle et al. (2010) argue that for each spatial analysis whereby the spatial process can be stochastic, it is important to take this into account.

This study presented a spatial analysis that shows the opportunities of crowd-sourced data to mitigate an epidemic during its outbreak phase. Future work will extend this modeling framework to a spatiotemporal setting to investigate the spatial and temporal differences in infections. These models can be extensions of the CAR convolution models where temporal correlation can be corrected for via a random-walk parameterization. In addition, the CAR and random-walk effects can interact, which allows for purely spatiotemporal trends. One difficulty will be to correct for time-dependent test capacity and case definitions. Besides extending the model to a spatiotemporal context, it can be further refined by the addition of other prediction covariates, for example, information of governmental measurements, which is likely to affect individual mobility and, through this, obstruct possible transmission routes. Furthermore, the time delay between experiencing symptoms and being diagnosed with COVID-19 was based on results from Neyens et al. (2020), who assumed a fixed delay. However, the delay time can be treated as a stochastic phenomenon, which has not been corrected for, due to current computational hurdles in the current modeling framework. Future research will investigate the possibility to map the symptoms experienced to the confirmed cases (or vice versa), using delay distributions. This allows to take into account the uncertainty of the period between the times of symptom reporting and diagnosis. Note that the confirmed COVID-19 cases are restricted to a specific time slot as aggregated data over a specific time, while reporting symptoms in the BCS does not necessarily reflect the date of symptom onset, leading to uncertainty regarding the exact start of the symptoms. Another assumption of the analysis in this manuscript is that there are no geographical differences in test strategies across the considered regions. Discrepancies between these strategies could lead to effort-based differences in observed incidences. Accounting for this would however entail the introduction of an additional model process that corrects for sampling bias in the confirmed cases. We chose not to pursue this extension in this study, since this would push our model beyond the limits of model complexity.

Note as well that we have presented analyses using a bivariate intrinsic CAR random-effects term and/or a bivariate normally distributed heterogeneity term. Since a CAR convolution model can suffer from identifiability problems, other random-effect structures were investigated as well, such as the multivariate Leroux prior. The random effect of a multivariate Leroux model is defined as the Kronecker product of the univariate CAR prior proposed by Leroux et al. (1999) and a between-process covariance matrix. It is characterized by a single random effect that contains a spatial mixing parameter to represent a range of weak to strong spatial correlation. However, this analysis was hampered by substantial convergence problems and was therefore not included in the manuscript. It is important to acknowledge again that the complexity of a three-way joint model leads us to the upper limit of what can be achieved in this data analysis. We therefore strongly advise others to thoroughly monitor posterior convergence when undertaking a similar analysis.

ACKNOWLEDGMENTS

The resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government. The data used in this manuscript were provided by the Belgian public health institute (Sciensano). This research received funding from the Flemish Government (AI Research Program). Authors Beutels, Faes, and Hens acknowledge funding from the European Union's Horizon 2020 research and innovation programme - project EpiPose (No. 101003688). Authors Beutels, Hens, Neyens, and Van Damme acknowledge funding from the Research Foundation Flanders (No. G0G1920N). Authors Faes and Neyens acknowledge funding from the Research Foundation Flanders (No. G0G9820N). Author Neyens acknowledges funding by the Internal Funds KU Leuven (project number 3M190682).

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

DATA AVAILABILITY STATEMENT

The covid data that support the findings of this study are provided by the Belgian public health institute (Sciensano) and are confidential. They are available in an aggregated format on https://epistat.wiv-isp.be/covid/. The crowd-sourced symptoms data are confidential, but are available after approval by the Corona Study steering committee. The full code to reproduce this study on a synthetic data set is provided as supplementary material.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

Maren Vranckx https://orcid.org/0000-0002-6509-7777
Oana Petrof https://orcid.org/0000-0002-1802-9640
Koen Pepermans https://orcid.org/0000-0001-7294-9491
Thomas Neyens https://orcid.org/0000-0003-2364-7555

REFERENCES

Besag, J., & Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, *82*, 733–746.

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Temple Lang, D., & Bodik, R. (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, *26*, 403–413.

Desson, Z., Weller, E., Mc Meekin, P., & Ammi, M. (2020). An analysis of the policy responses to the COVID-19 pandemic in France, Belgium, and Canada. *Spatial and Spatio-Temporal Epidemiology*, *9*, 430–446.

Diggle, P. J., Menezes, R., & Su, T. (2010). Geostatistical inference under preferential sampling. *Applied Statistics*, *59*, 191–232.

Geisser, S., & Eddy, W. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, *74*, 153–160.

Jin, X., Banerjee, S., & Carlin, B. P. (2007). Order-free coregionalized areal data models with application to multiple disease mapping. *Journal of the Royal Statistical Society B*, *69*, 817–883.

Lawson A. B,. (2020). NIMBLE for Bayesian disease mapping. *Spatial and Spatio-temporal Epidemiology*, *33*, 100323.

Leroux, B., Lei, X., & Breslow, N. (1999). Estimation of disease rates in small areas: A new mixed model for spatial dependence. In *Statistical models in epidemiology, the environment and clinical trials* (pp. 135–178). The IMA Volumes in Mathematics and Its Applications, No. 116. Springer.

Mamode Khan, N., Soobhug, A. D., & Heenaye-Mamode Khan, M. (2020). Studying the trend of the novel coronavirus series in Mauritius and its implications. *PLoS One*, *15*(7), e0235730.

Martinez-Beneito, M. A., & Botella-Rocamora, P. (2019). *Disease mapping from foundations to multidimensional modeling*. Chapman and Hall/CRC.

Neyens, T., Faes, C., & Molenberghs, G. (2012). A generalized Poisson-gamma model for spatially overdispersed data. *Spatial and Spatio-temporal Epidemiology*, *3*(3), 185–194.

Molenberghs, G., Verbeke, G., Demétrio, C. G. B., & Vieira, A. M. C. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science*, *25*(3), 325–347.

Neyens, T., Faes, C., Vranckx, M., Pepermans, K., Hens, N., Van Damme, P., Molenberghs, G., Aerts, J., & Beutels, P. (2020). Can COVID-19 symptoms as reported in a large-scale online survey be used to optimise spatial predictions of COVID-19 incidence risk in Belgium? *Spatial and Spatio-Temporal Epidemiology*, *35*, 100379.

Neyens, T., Lawson, A. B., Kirby, R. S., & Faes, C. (2016). The bivariate combined model for spatial data analysis. *Statistics in Medicine*, *35*, 3189–3202.

NIMBLE Development Team. (2019). NIMBLE: MCMC, particle filtering, and programmable hierarchical modeling. R Package Version 0.7.0. https://cran.r-project.org/package=nimble

Pati, D., Reich, B. J., & Dunson, D. B. (2011). Bayesian geostatistical modelling with informative sampling locations. *Biometrika*, *98*, 35–48.

R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Schumacher, F. L., Ferreira, C. S., Prates, M. O., Lachos, A., & Lachos, V. H. (2020). A robust nonlinear mixed-effects model for COVID-19 death data. *Statistics and Its Interface*, *14*(1), 49–57.

Vranckx, M., Neyens, T., & Faes, C. (2021). The (in) stability of Bayesian model selection criteria in disease mapping. *Spatial Statistics*, *43*, 100502.

Waller, L. A., & Gotway, C. A. (2004). *Applied spatial statistics for public health data*. Wiley.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571–3594.

World Health Organization (WHO). (2020). WHO Director-General's opening remarks at the media briefing on COVID-19. March 11. https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

Yang, W., Cao, Q., Qin, L., Wang, X., Cheng, Z., Pan, A., Dai, J., Sun, Q., Zhao, F., Qu, J., & Yan, F. (2020). Clinical characteristics and imaging manifestations of the 2019 novel coronavirus disease (COVID-19): A multi-center study in Wenzhou city, Zhejiang, China. *Journal of Infection*, *80*, 388–393.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Vranckx, M., Faes, C., Molenberghs, G., Hens, N., Beutels, P., Van Damme, P., Aerts, J., Petrof, O., Pepermans, K., & Neyens, T. (2022). A spatial model to jointly analyze self-reported survey data of COVID-19 symptoms and official COVID-19 incidence data. *Biometrical Journal*, 1–11. https://doi.org/10.1002/bimj.202100186