# Synthetic Reproduction and Augmentation of COVID-19 Case Reporting Data by Agent-Based Simulation

PRACTICE PAPER

NIKOLAS POPPER
MELANIE ZECHMEISTER
DOMINIK BRUNMEIR
CLAIRE RIPPINGER
NADINE WEIBRECHT
CHRISTOPH URACH
MARTIN BICHER
GÜNTER SCHNECKENREITHER
ANDREAS RAUBER

*Author affiliations can be found in the back matter of this article*

## ABSTRACT

We generate synthetic data documenting COVID-19 cases in Austria by the means of an agent-based simulation model. The model simulates the transmission of the SARS-CoV-2 virus in a statistical replica of the population and reproduces typical patient pathways on an individual basis while simultaneously integrating historical data on the implementation and expiration of population-wide countermeasures. The resulting data semantically and statistically aligns with an official epidemiological case reporting data set and provides an easily accessible, consistent and augmented alternative. Our synthetic data set provides additional insight into the spread of the epidemic by synthesizing information that cannot be recorded in reality.

CORRESPONDING AUTHOR:
**Günter Schneckenreither**
Institute of Information Systems Engineering, TU Wien, Vienna, Austria
*guenter.schneckenreither@tuwien.ac.at*

KEYWORDS:
COVID-19; synthetic data; agent-based simulation; data augmentation

TO CITE THIS ARTICLE:
Popper, N, Zechmeister, M, Brunmeir, D, Rippinger, C, Weibrecht, N, Urach, C, Bicher, M, Schneckenreither, G and Rauber, A. 2021. Synthetic Reproduction and Augmentation of COVID-19 Case Reporting Data by Agent-Based Simulation. *Data Science Journal*, 20: 16, pp. 1–13. DOI: *https://doi.org/10.5334/dsj-2021-016*

## 1 INTRODUCTION

Vaccines against SARS-CoV-2 and cures for the induced coronavirus disease (COVID-19) are not yet available. Authorities are challenged with imposing measures against the spread of the virus that can affect all areas of society and economy such as curfews, mandatory face-protection, social distancing, etc. It is clear that a comprehensive and timely overview on the spread and circulation of the virus is crucial for implementing the correct countermeasures, thus mitigating the impact on our health and lives. Informed decisions in the fight against the pandemic depend on reliable data. However, collected epidemiological data is prone to errors, artifacts and missing values, may be inaccessible or lack critical information.

The reasons partially stem from the nature of the disease including but not limited to a long latent period and contagiousness in asymptomatic cases. Additionally, due to the novelty of the virus, epidemiological and medical insight increases gradually with the pandemic progressing at the same time. For instance, the attribution of symptoms changed over time and still varies across reporting systems and geographic regions (Grant et al. 2020). Hence, we recognize that the interpretation of collected data can change for different records depending on their date of entry and for different data sets. Because reliable and wide-spread testing was not at hand until the second half of 2020, screening programs can lead to a post-hoc increase in the number of asymptomatic cases. For the same reasons, we observe in reported cases that the ratio between infections with 'severe' and 'mild' symptoms decreased as testing was conducted on a broader basis. Scientific and medical insight was also gained gradually about the duration of the latent and the infectious periods in patients, reinfection, the development of antibodies as well as the duration of their protective effect and the mutation rate of the virus. For example, an early survey published in (Backer, Klinkenberg, and Wallinga 2020) states that the incubation period is distributed with a mean of 6.8 days whereas more recent studies suggest rather 5 to 6 days (Wei et al. 2020). As a consequence, we are confronted with the temporal transformation of recorded and *perceived* patient pathways, which can require structural changes in databases and under circumstances lead to inconsistencies, missing entries or conversion errors in previously collected data.

On a technical level, implemented epidemiological (data) procedures and systems suffer from a long period of neglect and were initially not designed for quick response. Inconsistent and fractured (data) management of health care institutions and authorities can hinder the assessment of available and occupied resources. Data may be collected and published independently by regional authorities and health care institutions with different electronic systems, semantics and interpretation. For example, in Austria, public hospitals are operated by nine federal states; up to the current epidemic, there was no nation-wide reporting on the occupancy of intensive care units (ICU). Furthermore, decision makers and authorities only slowly or insufficiently implement new technical methods for surveillance including detailed case documentation, registration of individual quarantine orders and identifying persons at risk of infection via contact tracing. For example, the time periods between indication of COVID-19, testing and test result, which are crucial for assessing the effectiveness of the testing-tracing-isolating (TTI) strategy, are often not recorded. Especially if reporting systems are implemented in a hurry, data may not be annotated or documented. Access to the data for researchers is often complicated by administrative procedures. Or, even worse, data may only be available through the media, in non-machine-readable formats (Marivate and Combrink 2020) or held back completely due to political reasons or unresolved data privacy issues.

Data pre-processing, elaborate statistical methods and interleaving or the fusion of multiple data sources can be used to increase the insight into the characteristics, spread and circulation of the virus. Assessment of the quality of data is mandatory for informed decisions and transparent administrative processes. Even more so, the prediction of the future course of the epidemic is not possible without reliable data sources and careful statistical evaluation. A supplementary approach for increasing the insight into complex epidemiological processes, predicting the short-term development and evaluating possible future scenarios is by computer simulation. The technical approaches applied in this context range form statistical extrapolation of key figures to dynamic models that integrate a large number of parameters. The most prominent dynamic modeling approach describes the temporal evolution of the

size of different population compartments (susceptible, infected and recovered persons). This approach is usually implemented as susceptible-infected-recovered (SIR) differential equation models, which date back to Kermack and McKendrick (1927), or in the System Dynamics context (Homer and Hirsch 2006). Individual- or agent-based models, on the other hand, describe the temporal evolution of each single individual of a population from an ego-centered perspective. Such models require a larger number of parameters and a deeper understanding or hypotheses about individual behavior (e.g. contact behavior and social structuring). Theoretically, more complex models should be able to display effects that cannot be reproduced in aggregate models. Agent-based models with varying complexity have been developed and successfully employed recently for simulating the spread of COVID-19 (Chang et al. 2020; Cuevas 2020; Karatayev, Anand, and Bauch 2020; Mahmood et al. 2020; Silva et al. 2020).

Here, we aim to synthetically reproduce, pad and augment a data set of reported SARS-CoV-2 cases in Austria by means of agent-based simulation. To that end, we start this paper with the outline of an existing agent-based and event-driven simulation model (section 2) that was developed to reproduce and predict the course of the SARS-CoV-2 epidemic in Austria (M. R. Bicher et al. 2020). Our model is focused on the correct representation of the statistical configuration of the Austrian population, on the mapping of the reported incidence and on the actual history of the implementation and suspension of countermeasures. We furthermore take a patient-centered perspective by aiming for an accurate reproduction of patient pathways taking into account their variability and gradual transformation. We then discuss the structure and format of a data-excerpt (subsection 3.1), which was obtained from an infectious disease reporting system implemented by the Austrian authorities. We contrast the original data with an analogous synthetic case data set generated with our simulator (subsection 3.2). The synthetic data is characterized by reduced sparsity, logical consistency (i.e. no data-errors) and information included that is not observable in reality (e.g. the moment of infection). In section 4 we present preliminary comparative evaluations. We show that our synthetic data reflects the original data set on a longitudinal and aggregated scale. Furthermore, we show that our augmented and synthetic data displays effects that are known to exist in reality but cannot be recorded on a population scale.

The use of synthetic data gains increasing interest in various research areas (Wang, Myles, and Tucker 2019); initiatives for using synthetic data in COVID-19 research have been developed recently (UK Medicines and Healthcare products Regulatory Agency 2020). With our approach, we can provide synthetic data to researchers that is semantically analogous to official data, available without restrictions (no privacy issues, no restricted access), provides a higher resolution and is augmented with information that is not observable in reality.

## 2 INDIVIDUAL-BASED SIMULATION MODEL

We use an agent-based model that was previously developed to simulate the spread of COVID-19 in Austria (M. R. Bicher et al. 2020; dwh GmbH 2020). The technical implementation of our model aligns with the general approach in agent-based infection models found in literature (Chang et al. 2020; Cuevas 2020; Karatayev, Anand, and Bauch 2020; Mahmood et al. 2020; Meyer 2015; Miksch et al. 2014; Schneckenreither and Popper 2017; Silva et al. 2020) and is based on the combination of a population model with a model for face-to-face encounters and disease transmission.

### 2.1 MODEL DESIGN AND PARAMETERIZATION

The population of a country (i.e. Austria; approximately nine million inhabitants) is mapped by individual agents that correctly reproduce the demographic properties of the country (M. Bicher, Urach, and Popper 2018). The transmission of a contagious disease (i.e. SARS-CoV-2) between agents results from simulated face-to-face encounters which occur on a daily basis within virtual groups that represent workplaces, schools, households and leisure-time (Schneckenreither and Popper 2017). This *social structuring* is obtained (parameterized) from publicly available data on demography (Statistik Austria 2019a), commuting distances (Statistik Austria 2019c), geographic population density (Florczyk et al. 2015; Perlot 2017) and the distribution of workplace and school class sizes (Statistik Austria 2004, 2019b).

Temporal progression in the simulator is based on dynamic scheduling and processing of discrete events. On the one hand, such events are for instance the occurrence of death and childbirth which are stochastically generated for each agent depending on various social attributes (age, gender, administrative region, etc.) in order to align with the statistical distributions reported in official census data. On the other hand, physical contact among agents is randomly generated depending on social attributes, social structuring and according to a daily schedule that approximates the contact patterns found in large survey studies (Mossong et al. 2008). Simulated close proximity interaction permits the transmission of a disease with a certain likelihood, which in turn depends on the attributes of the interacting agents and on a global set of policies (currently implemented NPIs).

Once infected, an agent traverses different stages of the disease and treatment. The exposure of 'mild' or 'severe' symptoms, the occurrence of a medical test, hospitalization, isolation (self-quarantine) etc. introduce a number of possible branching points in the pathways of patients (***Figure 1***). In our model, the transition times between different disease stages (e.g. latent period or infectious period) are configured to reproduce available medical data (Federal Ministry of Social Affairs, Health, Care and Consumer Protection 2020a; Hellewell et al. 2020; Lauer et al. 2020; Pollán et al. 2020; Robert Koch Institut 2020). The transition times, branching behavior and transmission likelihoods, however, also depend on the currently implemented public policies (quarantine orders, face protection, etc.).

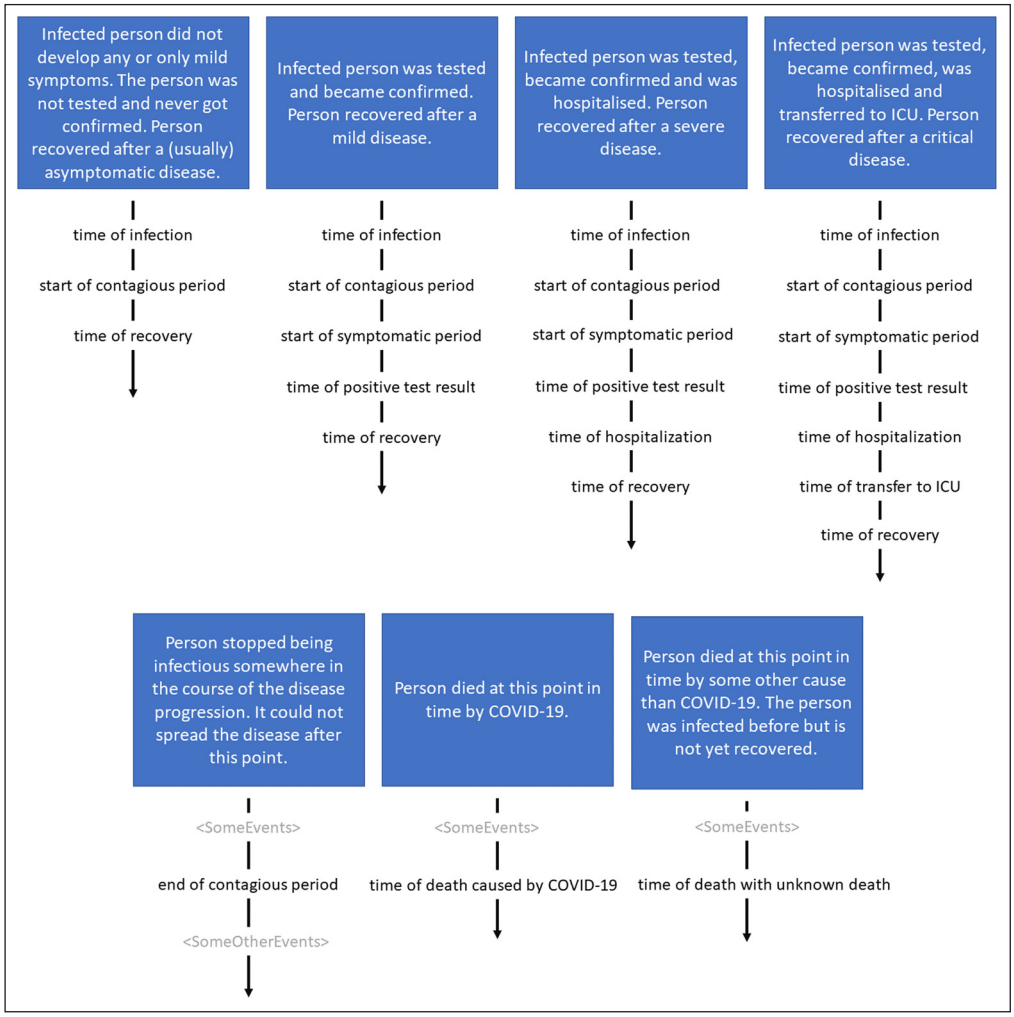


**Figure 1** Illustration of different disease and treatment pathways from an agent or patient perspective.

As countermeasures and public policies influence the typical disease pathways of individual patients and also the contact behavior of the population, we use a manually recorded timeline of the implementation and expiration of administrative policies (e.g. curfews) to modulate the behavior of agents (Rippinger et al. 2020). Besides clearly defined policies like school closures or restriction of public venues, this also includes reduced infection likelihoods, which result from increased awareness and hygienic measures, or preemptive reduction of contacts due to psychological or social effects that cannot be easily quantified and in turn require heuristic modeling.

DOI: 10.5334/dsj-2021-016

## 2.2 TECHNICAL IMPLEMENTATION AND USABILITY

Our simulator was implemented in the Java programming framework and is composed of different *logic components* such as population dynamics, disease pathways and interaction. These modules are responsible for implementing specific aspects of the agent behavior. Each module registers a set of event handlers and provides the necessary routines to schedule new events and to modify the state of agents (e.g. disease) depending on global and agent states. For instance, if a physical contact event is processed by the disease transmission module, the states of both interacting agents are checked for infection. If the transmission of the infection is decided to occur, an event is scheduled for the end of the latent period and for the start of the infectious period in the newly infected agent. This basic algorithmic concept is rather common in event-driven agent-based simulation models and frameworks (Meyer 2015). However, specialities of the underlying conceptual model require some features that are difficult to achieve with existing frameworks.

The (relatively) large number of agents and the resulting complexity of the event-handling approach require specific optimizations that cannot be provided by general purpose agent-based simulation frameworks due to the broadness of their intended scope of application. A large number of existing frameworks are designed to be operated via a graphical user-interface, which adds enormous overhead with a large number of agents. A directly programmed simulator grants more flexibility in the implementation of logic components or modules and allows to integrate arbitrary data (parameterization) without being restricted to predefined interfaces. Furthermore, fine-grained data collection and the controlled generation of random numbers adds additional benefits in automated Monte Carlo simulations. For further documentation on our simulation framework we kindly refer the reader to (M. R. Bicher et al. 2020; dwh GmbH 2020).

A crucial prerequisite for simulating epidemic spread with our model is data pre-processing, which includes acquisition, validation and transformation of a large number of data sets. Typically, a simulation experiment consists in the initial parameterization of all model components, an iterative calibration and simulation process (see the following section), data collection and final evaluation.

## 2.3 MODEL CALIBRATION

We calibrate our simulations to quantitatively and qualitatively match the first epidemic wave in Austria (March 11–July 1) by varying the infection probability during face-to-face contacts and the impact of countermeasures implemented by authorities on the contact behavior of agents. In other words, we iteratively compare our aggregated simulation results with data to find the correct parameter values for disease transmission and the impact of certain countermeasures on the same. The reference data consists of the cumulative confirmed cases as stated by the official epidemiological reporting system ('Epidemiologisches Meldesystem', Federal Ministry of Social Affairs, Health, Care and Consumer Protection (2020a)) to which we fit the corresponding output variables of the model. A detailed specification of the calibrated parameter values used for generating the data set is found in (Rippinger et al. 2020). A specification of the calibration method is presented in (dwh GmbH 2020).

Besides the data we process for parameterizing and calibrating the model, we use additional data to validate the model dynamics by comparing the corresponding outcomes of the model with the numbers reported in this *independent* data. This data includes time-series of the hospitalization, severity and fatality in cases (Federal Ministry of Internal Affairs 2020; Federal Ministry of Social Affairs, Health, Care and Consumer Protection 2020a) as well as the age distribution in confirmed cases (Federal Ministry of Social Affairs, Health, Care and Consumer Protection 2020a) and mobility data provided by mobile phone companies (Heiler et al. 2020). Each of these data sets is used to validate separate sub-aspects (modules) of our model. However, most of this data is not publicly available due to data privacy issues.

Although some model parameters are calibrated solely with regard to the number of confirmed cases, the resulting parameters are similar to values proposed in literature. For example, the infection probability during face-to-face contact has been calibrated to 5%, resembling a value of 4.4% reported by Böhmer et al. (2020). Similarly, the obtained value of 78% reduction of leisure time contacts and the closure of 50% of workplaces during the lock-down phase align with 87% mobility reduction for retail and recreation and 51% mobility reduction for workplaces evaluated by Google (2020) in late March.

If the model parameters and the assumptions on disease pathways are correct or at least reasonably realistic, we can expect that our simulations can provide valid insight into otherwise unobserved dynamics in the spread of the virus.

# 3 DATA

## 3.1 REPORTED CASE DATA

As a real world example we refer to a data set that was extracted from the official epidemiological reporting system documenting cases of notifiable diseases in Austria (Federal Ministry of Social Affairs Health Care and Consumer Protection 2020b). The system is available to government agencies and decision makers, provides technical reporting interfaces for health care facilities and diagnosis laboratories and displays various information about individual patient pathways. Diagnoses of COVID-19 and SARS-CoV-2 infections are registered in the system starting with January 26, 2020.

Beginning with June, a platform was created to provide researchers with data on the SARS-CoV-2 epidemic (Federal Ministry of Social Affairs Health Care and Consumer Protection 2020a). The data is made available via a subsidiary of the ministry ('Gesundheit Österreich GmbH'). We obtained a data-excerpt via a research contract at the end of July. The data was provided as a file of comma separated pathway data, each record reporting on a specific confirmed COVID-19 case, totaling 20,797 cases. In this data set, patient information was aggregated and anonymized and only a limited number of data fields is available (***Table 1***). Our excerpt only contains cases diagnosed with COVID-19 between calendar weeks 9 and 31 and is not regularly updated.

**Table 1** Data fields contained in an excerpt of a case reporting database.

| COLUMN NAME | DESCRIPTION |
|---|---|
| patient ID | Random index or identifier. |
| age group | Age of the patient aggregated in six groups (<20, 20–34, 35–49, 50–64, 65–79, >79). |
| gender | Gender of the patient. |
| province | Regional attribution of the patient. |
| region code | Additional regional attribution of the patient as 3-digit code. The used regional structuring does not align with administrative structuring but was developed by Austrian health care institutions to fit their specific needs Federal Ministry of Social Affairs Health Care and Consumer Protection 2020c). |
| region name | Additional regional attribution of the patient as name. See notes above. |
| time of diagnosis | Week of the year when the patient was officially diagnosed with COVID-19 by a certified health care facility. This field is available for all recorded patients. |
| time of death | Date when the patient deceased. A patient is registered as deceased only if the passing was officially related to COVID-19 by an authorized institution. Otherwise this field is empty. |
| time of recovery | Date when the patient was considered recovered. A patient is considered recovered either if sufficient negative test-results were obtained or if the patient was in quarantine for two weeks after the initial diagnosis. The latter is relevant in particular for patients with mild symptoms. Either a recovery date or a deceased date must be present. |

We also note that the data was obtained without documentation or proper annotation and that various interesting information (e.g. on hospitalization) is excluded from our contract. We recognize isolated data errors, which is to be expected in collected data. For instance, the data set contains records of patients with a COVID-19 diagnosis that are never registered as either deceased or recovered within a reasonable time period. Concerning the registration of fatalities, we recognize that the affiliation with COVID-19 is not always unambiguous and varies across reporting systems. In particular, 'infection fatalities' are usually underestimated in contrast to 'case fatalities'; non-COVID-related deaths (e.g. caused by accidents or other diseases) of persons that were positively tested, are rarely reported to epidemiological surveillance systems (compare (Henriques 2020)). Hence, we treat this number as a crude estimate that reflects the qualitative development of actual fatalities over time.

### 3.2 SYNTHETIC CASE DATA

We collect the events processed by our simulator and extract the disease trajectories of individual agents to obtain a display analogous to the data described in subsection 3.1. The resulting data includes fields that are not available in the research data set due to privacy considerations. Since simulated individuals are merely statistical representatives, all personal (but artificial) information can be included in the highest detail. Furthermore, in simulated patients we have access to otherwise inaccessible information, like the exact (but simulated) time of infection, or to information that is not recorded in reality. As a consequence, we can provide the complete history or pathway (compare ***Figure 1***) for every infected (virtual) person.

We provide access to our data via a public repository (Rippinger et al. 2020) without restrictions. The data is organized according to ***Table 2*** and contains simulated infections in the period from February 12 to July 1, which corresponds to a number of 98,342 virtual patients. The data only contains relevant events in the timeline of agents that get infected with the virus (i.e. does not include additional information, events or agents that never get infected).

**Table 2** Fields in a synthetic data set generated with an agent-based simulation model. The data is available in (Rippinger et al. 2020). Additional information on the data fields and their interpretation is available in the same reference and in the following sections of this paper.

| COLUMN NAME | DESCRIPTION |
|---|---|
| patient ID | Agent identifier. |
| date of birth | Date when the virtual person was born. |
| gender | Gender of the virtual patient. |
| time of infection | Date of the patient (agent) getting infected. This timestamp is available for every patient contained in the synthetic case-reporting data set. This information is not observable in reality. |
| start of contagious period | Point in time when the patient starts being contagious. Corresponds to the end of the latent period. This information is not observable in reality. |
| end of contagious period | Point in time when the patient stops being contagious. |
| start of symptomatic period | Corresponds to the timestamp when the patient is due to get tested. Hence, we implicitly assume that all persons experiencing symptoms are getting tested. Vice versa, most often the cause for initiating a test is the patient experiencing symptoms. However, we do not differentiate between the motives for initiating the testing process (e.g. being traced as a contact partner, being screened randomly, or actively suspecting a possible infection due to prior contacts). If present, this date is always two days after the agent became infectious, which corresponds to the average pre-symptomatic phase (incubation period) as reported in studies (Robert Koch Institut 2020). |
| time of positive test result | Timestamp when a positive test result is obtained. |
| time of hospitalization | Timestamp of hospitalization of the patient. This event to occur requires previous testing. |
| time of transfer to ICU | Timestamp of transfer to intensive care unit. This event to occur requires previous hospitalization. |
| time of recovery | Date when the virtual patient recovers from COVID-19. A recovery event implies that symptoms and infectiousness stop. Analogous to original data set. |
| time of death caused by COVID-19 | Date when the virtual patient dies of an infection with the SARS-CoV-2 virus. Due to model limitations, this event and timestamp is determined retrospectively. |
| time of death with unknown cause | Time of death that is not caused by COVID-19 but implied by the population dynamics. Only one of recovery, death by COVID-19 and death by unknown cause can apply. |

## 4 RESULTS AND EVALUATION

We demonstrate possible application scenarios and statistical evaluation of our synthetic COVID-19 case data and compare some key figures in the synthetic and in real data. According to the limitations of both data sets, we restrict the following evaluations to the time period between March 15 to June 26.

Our model was calibrated on the detected prevalence of COVID-19. As a consequence it is clear that the number of reported cases corresponds in both data sets (***Figure 2***, left). For the individual age compartments 0–20, 20–34, 35–49, 50–64, 65–79 and 80+ we obtain the mean square errors (MSE) 61, 63, 75, 95, 121 and 78 between the timeseries of real and synthetic reported case

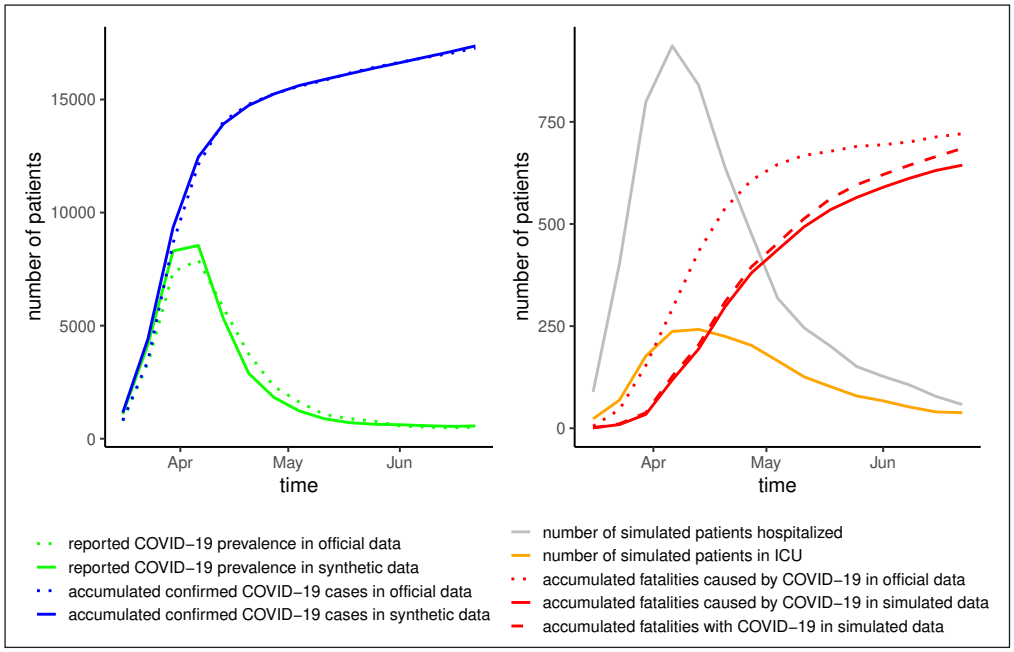


**Figure 2** Comparison of real and synthetic data. On the left: The simulation model is calibrated to correctly reproduce the reported prevalence and accumulated number of COVID-19 cases. Right-hand side: The number of reported fatalities caused by COVID-19 is closely approximated in dynamic simulations. The synthetic data provides additional figures on hospitalization that are not included in the original data set.

numbers. Since the model simulates the contact (and contagion) characteristics of individual persons, it is possible to infer additional latent figures such as the number of unreported cases. This is mirrored by the number of records (i.e. reported cases) in the original data (20,797) and the number of records (including unreported cases) in the synthetic data set (98,342).

With the inclusion of additional knowledge on individual patient trajectories (disease pathways), we can roughly reproduce the total number of fatalities as reported in real data. However, due to the limitations of our model (subsection 3.2) and the limitations in reliably determining the true cause of mortality in actual fatalities (subsection 3.1), we can expect certain discrepancies. Because in the model the decease of a patient is determined retrospectively at the expected time of recovery, we statistically overestimate the infection period in fatal cases. In ***Figure 2*** (right) we show that the total number of fatalities attributed to COVID-19 are qualitatively reflected in our synthetic data, yet the curve displays a significant time delay. Nevertheless, we argue that synthetic data (in general) could allow better estimation of the case fatality ratio (CFR) because severe cases are over-represented when only confirmed infections are observed. For instance, the first assessments of the CFR in Austria from March to June were about 4%, whereas only considering cases from July onward, a CFR of about 0.4% can be observed (AGES – Austrian Agency for Health and Food Safety 2020).

Furthermore, the synthetic data contains additional information on patient pathways that is not included in the original data such as severe (hospitalization required) and critical (intensive care unit) disease progressions (***Figure 2***, right). In reality, the estimation of these figures is crucial for evaluating the expected load on the health care system. We compare our results for the first day of April, May and June 2020 (hospitalized: 842, 330, 116; in ICU: 152, 120, 43) with public data (AGES – Austrian Agency for Health and Food Safety 2020) (hospitalized: 856, 348, 68; in ICU: 215, 124, 26) and obtain a maximum deviation of about 60 (average 27).

We observe in our simulations a peak number in weekly (March 15–21, 2020) aggregated COVID-19 infections during the first epidemic wave of 27,656, which is composed of 5,141 confirmed and 22,515 undetected cases. Hence, our model suggests that less than 20% percent of the infections were detected in this initial phase of the epidemic in Austria. Antibody studies (Medizinische Universität Innsbruck 2020) estimate that the detection ratio was roughly 15% during this period. Findings for other countries support these numbers (Wu et al. 2020). Especially when specific testing strategies are applied (e.g. screening of school children or in nursery homes) or testing capacities are limited, which also leads to a strong pre-selection of the tested population, the age-distribution of confirmed cases in official numbers is fundamentally biased. Mapping those testing strategies in the simulation model and calibrating on official numbers allows to reduce this bias and to observe a more realistic age-distribution of COVID-19 cases in the synthetic data. ***Figure 3*** shows clearly the under-reporting of COVID-19 prevalence in younger age-groups especially within the first months of the pandemic. During this period

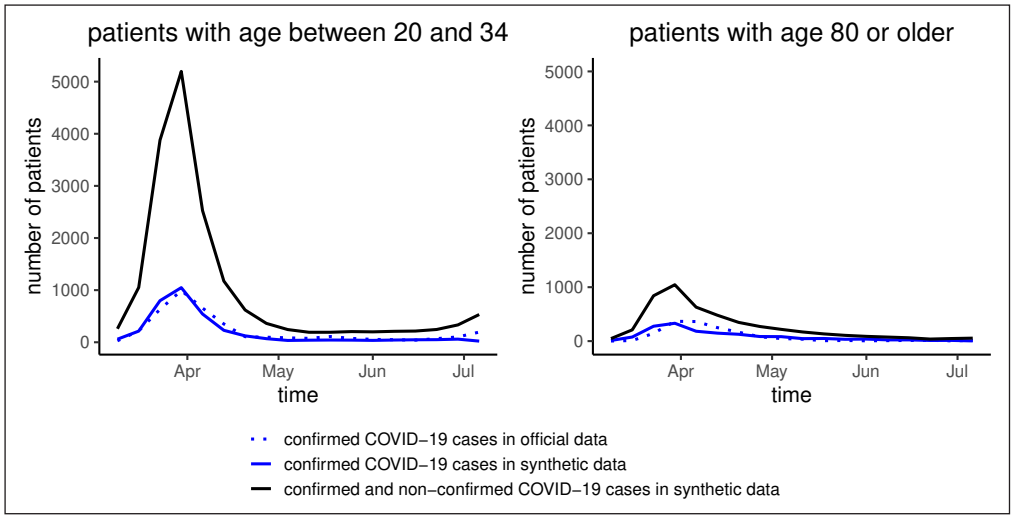


**Figure 3** Number of confirmed and unconfirmed cases in different risk groups according to age (low risk: 20–34, high risk: 80+). In contrast to real data, the simulation model also provides the number of unconfirmed cases. We observe that the relative number of unconfirmed cases is higher in the low risk group.

(March 15–21, 2020), the detection rate obtained was 20% in the age group 20–34 and 32% in the high risk group 80+ (compared to a total detection rate of about 20%). Note, that the total population numbers are approximately 1,730,000 in the age group 20–34 and 470,000 in the age group 80+ (Statistik Austria 2019a), corresponding to a ratio of 3.5; the number of confirmed cases in the younger low-risk group is greater by only a factor of 2.

## 5 DISCUSSION

We provide a synthetic data set on documented COVID-19 cases in Austria (Rippinger et al. 2020). Our data statistically reflects the figures in official infectious disease reporting and is further augmented with additional information on SARS-CoV-2 infections that are not detected in reality. The synthetic data is generated from an agent-based simulation model (M. R. Bicher et al. 2020; M. Bicher, Urach, and Popper 2018; dwh GmbH 2020) that is carefully parameterized and calibrated based on additional data sources and expert knowledge. Hence, in this context we regard agent-based simulation as a dynamic method for the fusion, imputation and augmentation of data.

The rationale of gradually growing medical insight (e.g. duration of incubation and infectious periods) as well as the availability and implementation of countermeasures (vaccination, treatment, public policies) leads us to the notion of *perceived patient trajectories*. By this term we refer to the currently established model of (typical) patient trajectories. A primary premise of our approach is to adequately reproduce these agent- or patient-centric disease pathways while simultaneously adhering to the available historical data. Whereas it is clear that the numbers on which the model is calibrated are reproduced correctly, modeling of dynamic effects and processes on the population level (based on historical data) and in individual patient trajectories allows to reproduce numbers and figures that are not directly evident in the data that was used for parameterization and calibration. Because in our model, events that simulate the traversal of individual disease pathways are generated according to stochastic parameters, our data must be regarded as a statistical sample taken from the parameterized model. Hence, it is clear that our data only reflects the official numbers in an aggregate and statistical fashion and that the model input data cannot be reconstructed from the result data.

Our synthetic data is intended as a corrected and augmented analogous to official case documentation (subsection 3.1) that is accessible for researchers and data scientists on a broader basis and can be used, for instance, in the development and testing of data procedures and visualization tools. To that end, our synthetic data maintains the same semantics and structure as the original data as much as possible. Due to the elimination of privacy issues and access formalities, our data is available without restrictions. Furthermore, we provide extensive documentation and annotation with our data set. However, since parameterization of our simulation model relies on a large number of data sets and requires additional pre-processing, the model cannot be easily adapted to another country or another disease. Also the modification or inclusion of additional model components is not straight forward but usually involves additional effort for data acquisition, data processing, programming and validation. Our model was not tested for other regions than Austria, but the basic concept of generating synthetic or augmented data with the help of agent-based simulation models could be transferred to other countries and simulation models.

With this approach it should further be possible to generate artificial case reporting data that simulates the progression of the epidemic in hypothetical scenarios. In the end, this should allow to qualitatively but systematically evaluate countermeasures and prevention strategies in a virtual environment, such as vaccination and lock-down policies (Abdollahi et al. 2020). By successive improvement in the simulation of physical contact behavior and disease transmission, we hope to increase the insight into actual transmission paths and in the occurrence of infection clusters (Leclerc et al. 2020). In particular, by interlacing of statistical information on social relations and social structuring (Schneckenreither and Popper 2017) and on geographic mobility patterns (Heiler et al. 2020) we infer and dynamically reproduce transmission trajectories as observed or anticipated in reality. We intend to provide synthetic data on transmission clusters including super-spreading events in combination with geographic and socio-structural information in the future.

## ACKNOWLEDGEMENTS

This work was supported by the Austrian Federal Ministry for Social Affairs, Health, Care and Consumer Protection and 'Gesundheit Österreich GmbH'.

The authors acknowledge the 'TU Wien Bibliothek' for financial support through its Open Access Funding Program.

## FUNDING INFORMATION

Austrian Research Promotion Agency (FFG) COVID-19 Emergency Call, Vienna Science and Technology Fund WWTF-COVID-19 Rapid Response Funding, Medical Scientific Fund of the Mayor of the City of Vienna, Society for Medical Decision Making (SMDM) COVID-19 Decision Modeling Initiative. The funding institutions had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## COMPETING INTERESTS

The authors have no competing interests to declare.

## AUTHOR CONTRIBUTIONS

Study conception: NP MZ CU GS AR. Model implementation: DB CR CU MB. Model parameterization and calibration: NP MZ DB CR CU MB. Data curation and processes: NP MZ NW AR. Results evaluation: NP MZ DB CR NW MB. Statistical evaluation: MZ CR. Writing: NP MZ DB CR NW CU MB GS AR. Figures: MZ. Final writing and manuscript: GS.

## AUTHOR AFFILIATIONS

**Nikolas Popper** *orcid.org/0000-0003-4615-2774*
Institute of Information Systems Engineering, TU Wien, Vienna, Austria

**Melanie Zechmeister**
DEXHELPP – Decision Support for Health Policy and Planning, Vienna, Austria

**Dominik Brunmeir** *orcid.org/0000-0002-9005-4066*
Dwh simulation services, Vienna, Austria

**Claire Rippinger** *orcid.org/0000-0001-9916-562X*
Dwh simulation services, Vienna, Austria

**Nadine Weibrecht**
DEXHELPP – Decision Support for Health Policy and Planning, Vienna, Austria

**Christoph Urach** *orcid.org/0000-0002-2480-5140*
Dwh simulation services, Vienna, Austria

**Martin Bicher** *orcid.org/0000-0002-1362-6868*
Institute of Information Systems Engineering, TU Wien, Vienna, Austria

**Günter Schneckenreither** *orcid.org/0000-0002-9217-9399*
Institute of Information Systems Engineering, TU Wien, Vienna, Austria

**Andreas Rauber**
Institute of Information Systems Engineering, TU Wien, Vienna, Austria

## REFERENCES

**Abdollahi, E,** et al. Dec. 2020. Simulating the effect of school closure during COVID-19 outbreaks in Ontario, Canada. en. In: *BMC Medicine*, 18(1): 230. ISSN: 1741-7015. URL: *https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-020-01705-8* (visited on 10/12/2020). DOI: *https://doi.org/10.1186/s12916-020-01705-8*

**AGES – Austrian Agency for Health and Food Safety.** 2020. *Dashboard COVID19*. URL: *https://covid19-dashboard.ages.at/dashboard.html?area=10* (visited on 10/14/2020).

**Backer, JA, Klinkenberg, D** and **Wallinga, J.** Feb. 2020. Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020. en. In: *Eurosurveillance*, 25(5). ISSN: 1560-7917. URL: *https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2020.25.5.2000062* (visited on 10/14/2020). DOI: *https://doi.org/10.2807/1560-7917.ES.2020.25.5.2000062*

**Bicher, M, Urach, C** and **Popper, N.** 2018. GEPOC ABM: A Generic Agent-Based Population Model for Austria. In: *Proceedings of the 2018 Winter Simulation Conference*. Gothenburg, Sweden: IEEE, pp. 2656–2667. DOI: *https://doi.org/10.1109/WSC.2018.8632170*

**Bicher, MR,** et al. May 2020. *Agent-Based Simulation for Evaluation of Contact-Tracing Policies Against the Spread of SARS-CoV-2*. en. preprint. Epidemiology. URL: *http://medrxiv.org/lookup/doi/10.1101/2020.05.12.20098970* (visited on 09/23/2020). DOI: *https://doi.org/10.1101/2020.05.12.20098970*

**Böhmer, MM,** et al. Aug. 2020. Investigation of a COVID-19 outbreak in Germany resulting from a single travel-associated primary case: a case series. en. In: *The Lancet Infectious Diseases*, 20(8): 920–928. ISSN: 14733099. URL: *https://linkinghub.elsevier.com/retrieve/pii/S1473309920303145* (visited on 09/30/2020). DOI: *https://doi.org/10.1016/S1473-3099(20)30314-5*

**Chang, SL,** et al. Dec. 2020. Modelling transmission and control of the COVID-19 pandemic in Australia. en. In: *Nature Communications*, 11(1): 5710. ISSN: 2041-1723. URL: *http://www.nature.com/articles/s41467-020-19393-6* (visited on 02/24/2021). DOI: *https://doi.org/10.1038/s41467-020-19393-6*

**Cuevas, E.** June 2020. An agent-based model to evaluate the COVID-19 transmission risks in facilities. en. In: *Computers in Biology and Medicine*, 121: 103827. ISSN: 00104825. URL: *https://linkinghub.elsevier.com/retrieve/pii/S001048252030192X* (visited on 10/12/2020). DOI: *https://doi.org/10.1016/j.compbiomed.2020.103827*

**dwh GmbH.** Sept. 2020. *Technical Model Specification of the dwh and TU Wien Agent-Based Covid-19 Model*. Tech. rep. URL: *https://www.dwh.at/projects/covid-19/Covid19_Model_20200817.pdf*.

**Federal Ministry of Internal Affairs.** Aug. 2020. *COVID-19 data collected by the Federal Ministry of Internal Affairs*. URL: *https://bmi.gv.at/news.aspx?id=4A7171477A51625143334D3D* (visited on 09/15/2020).

**Federal Ministry of Social Affairs, Health, Care and Consumer Protection.** Sept. 2020a. *Datenplattfrom COVID-19*. URL: *https://datenplattform-covid.goeg.at/* (visited on 09/15/2020).

**Federal Ministry of Social Affairs, Health, Care and Consumer Protection.** Sept. 2020b. *Rechtliche Grundlagen und Meldung übertragbarer Krankheiten*. URL: *https://www.sozialministerium.at/Themen/Gesundheit/Uebertragbare-Krankheiten/Rechtliches.html* (visited on 09/14/2020).

**Federal Ministry of Social Affairs, Health, Care and Consumer Protection.** Sept. 2020c. *Regionale Gliederung Stand 2019*. URL: *https://www.sozialministerium.at/dam/jcr:f5dbe811-fd95-489b-a7e5-6d141979f011/regionale_gliederung,_stand_2019.xlsx* (visited on 09/23/2020).

**Florczyk, AJ,** et al. 2015. A new European settlement map from optical remotely sensed data. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9.5. Publisher: IEEE, pp. 1978–1992. DOI: *https://doi.org/10.1109/JSTARS.2015.2485662*

**Google.** Mar. 2020. *COVID-19 Community Mobility Report – Austria*. Tech. rep. URL: *https://support.google.com/covid19-mobility*.

**Grant, MC,** et al. June 2020. The prevalence of symptoms in 24,410 adults infected by the novel coronavirus (SARS-CoV-2; COVID-19): A systematic review and meta-analysis of 148 studies from 9 countries. en. In: Hirst, JA (ed.), *PLOS ONE*, 15(6): e0234765. ISSN: 1932-6203. (visited on 10/16/2020). DOI: *https://doi.org/10.1371/journal.pone.0234765*

**Heiler, G,** et al. 2020. Country-wide mobility changes observed using mobile phone data during COVID-19 pandemic. In: *arXiv preprint arXiv:2008.10064*. DOI: *https://doi.org/10.1109/BigData50022.2020.9378374*

**Hellewell, J,** et al. Apr. 2020. Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts. en. In: *The Lancet Global Health*, 8(4): e488–e496. ISSN: 2214109X. URL: *https://linkinghub.elsevier.com/retrieve/pii/S2214109X20300747* (visited on 10/02/2020). DOI: *https://doi.org/10.1016/S2214-109X(20)30074-7*

**Henriques, M.** 2020. *Coronavirus: Why death and mortality rates differ*. URL: *https://www.bbc.com/future/article/20200401-coronavirus-why-death-and-mortality-rates-differ*.

**Homer, JB** and **Hirsch, GB.** Mar. 2006. System Dynamics Modeling for Public Health: Background and Opportunities. en. In: *American Journal of Public Health*, 96(3): 452–458. ISSN: 0090-0036, 1541-0048. URL: *http://ajph.aphapublications.org/doi/10.2105/AJPH.2005.062059* (visited on 02/26/2021). DOI: *https://doi.org/10.2105/AJPH.2005.062059*

**Karatayev, VA, Anand, M** and **Bauch, CT.** Sept. 2020. Local lockdowns outperform global lockdown on the far side of the COVID-19 epidemic curve. In: *Proceedings of the National Academy of Sciences*, 117(39): 24575–24580. ISSN: 0027-8424, 1091-6490. URL: *http://www.pnas.org/lookup/doi/10.1073/pnas.2014385117* (visited on 10/12/2020). DOI: *https://doi.org/10.1073/pnas.2014385117*

**Kermack, WO** and **McKendrick, AG.** Aug. 1927. A contribution to the mathematical theory of epidemics. en. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115(772): 700–721. ISSN: 0950-1207, 2053-9150. URL: *https://royalsocietypublishing.org/doi/10.1098/rspa.1927.0118* (visited on 02/26/2021). DOI: *https://doi.org/10.1098/rspa.1927.0118*

**Lauer, SA,** et al. May 2020. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. en. In: *Annals of Internal Medicine*, 172(9): 577–582. ISSN: 0003-4819, 1539-3704. URL: *https://www.acpjournals.org/doi/10.7326/M20-0504* (visited on 10/14/2020). DOI: *https://doi.org/10.7326/M20-0504*

**Leclerc, QJ,** et al. June 2020. What settings have been linked to SARS-CoV-2 transmission clusters? en. In: *Wellcome Open Research*, 5: 83. ISSN: 2398-502X. URL: *https://wellcomeopenresearch.org/articles/5-83/v2(visited* on 10/12/2020). DOI: *https://doi.org/10.12688/wellcomeopenres.15889.1*

**Mahmood, I,** et al. Aug. 2020. FACS: a geospatial agent-based simulator for analyzing COVID-19 spread and public health measures on local regions. en. In: *Journal of Simulation*. pp. 1–19. ISSN: 1747-7778, 1747-7786. URL: *https://www.tandfonline.com/doi/full/10.1080/17477778.2020.1800422* (visited on 10/12/2020). DOI: *https://doi.org/10.1080/17477778.2020.1800422*

**Marivate, V** and **Combrink, HM.** May 2020. Use of Available Data To Inform The COVID-19 Outbreak in South Africa: A Case Study. en. In: *Data Science Journal*, 19: 19. ISSN: 1683-1470. URL: *http://datascience.codata.org/articles/10.5334/dsj-2020-019/* (visited on 09/14/2020). DOI: *https://doi.org/10.5334/dsj-2020-019*

**Medizinische Universität Innsbruck.** June 2020. *Ischgl-Studie: 42,4 Prozent sind Antikörper-positiv*. URL: *https://www.i-med.ac.at/mypoint/news/746359.html*.

**Meyer, R.** 2015. Event-Driven Multi-agent Simulation. In: *Multi-Agent-Based Simulation XV*. Ed. by Francisco Grimaldo and Emma Norling. Vol. 9002. Cham: Springer International Publishing. pp. 3–16. ISBN: 9783319146263. URL: *http://link.springer.com/10.1007/978-3-319-14627-0_1* (visited on 02/24/2021). DOI: *https://doi.org/10.1007/978-3-319-14627-0_1*

**Miksch, F,** et al. 2014. A Flexible Agent-Based Framework for Infectious Disease Modeling. In: *Information and Communication Technology*. Ed. by Linawati et al. Vol. 8407. Berlin, Heidelberg: Springer Berlin Heidelberg. pp. 36–45. ISBN: 9783642550317. URL: *http://link.springer.com/10.1007/978-3-642-55032-4_4* (visited on 02/24/2021). DOI: *https://doi.org/10.1007/978-3-642-55032-4_4*

**Mossong, J,** et al. Mar. 2008. Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases. en. In: Riley, S (ed.), *PLoS Medicine*, 5(3): e74. ISSN: 1549-1676. (visited on 02/24/2021). DOI: *https://doi.org/10.1371/journal.pmed.0050074*

**Perlot, F.** 2017. *Geo- and TopoJSON files of municipalities, districts and states in Austria*. URL: *https://github.com/ginseng666/GeoJSON-TopoJSON-Austria* (visited on 05/12/2019).

**Pollán, M,** et al. Aug. 2020. Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study. en. In: *The Lancet*, 396(10250): 535–544. ISSN: 01406736. URL: *https://linkinghub.elsevier.com/retrieve/pii/S0140673620314835* (visited on 10/02/2020). DOI: *https://doi.org/10.1016/S0140-6736(20)31483-5*

**Rippinger, C,** et al. Oct. 2020. *Synthetic COVID-19 Case Reporting Data Generated from an Agent-Based Simulation Model*. en. type: dataset. URL: *https://zenodo.org/record/4055943* (visited on 10/01/2020). DOI: *https://doi.org/10.5281/ZENODO.4055943*

**Robert Koch Institut.** 2020. SARS-CoV-2 *Steckbrief zur Coronavirus-Krankheit-2019 (COVID-19)*. URL: *https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Steckbrief.html#doc13776792bodyText7*.

**Schneckenreither, G** and **Popper, N.** Dec. 2017. Dynamic multiplex social network models on multiple time scales for simulating contact formation and patterns in epidemic spread. In: *2017 Winter Simulation Conference (WSC)*. Las Vegas, NV: IEEE. pp. 4324–4336. ISBN: 9781538634288. URL: *http://ieeexplore.ieee.org/document/8248138/* (visited on 10/12/2020). DOI: *https://doi.org/10.1109/WSC.2017.8248138*

**Silva, PCL,** et al. Oct. 2020. COVID-ABS: An agent-based model of COVID-19 epidemic to simulate health and economic effects of social distancing interventions. en. In: *Chaos, Solitons & Fractals*, 139: 110088. ISSN: 09600779. URL: *https://linkinghub.elsevier.com/retrieve/pii/S0960077920304859* (visited on 10/12/2020). DOI: *https://doi.org/10.1016/j.chaos.2020.110088*

**Statistik Austria.** 2004. *Arbeitsstättenzählung 2001*. Verlag Österreich.

**Statistik Austria.** 2019a. *Bevölkerungsstand und Bevölkerungsveränderung*. URL: *http://www.statistik.at/web_de/statistiken/bevoelkerung/bevoelkerungsstand_und_veraenderung/index.html*.

**Statistik Austria.** 2019b. *Bildung – Bundesanstalt Statistik Österreich*. URL: *https://statistik.at/web_de/statistiken/menschen_und_gesellschaft/bildung/index.html*.

**Statistik Austria.** 2019c. *Pendlerinnen und Pendler – Bundesanstalt Statistik Österreich*. URL: *https://www.statistik.at/web_de/statistiken/menschen_und_gesellschaft/bevoelkerung/volkszaehlungen_registerzaehlungen_abgestimmte_erwerbsstatistik/pendlerinnen_und_pendler/index.html*.

**UK Medicines and Healthcare products Regulatory Agency.** July 2020. *New synthetic datasets to assist COVID-19 and cardiovascular research*. URL: *https://www.gov.uk/government/news/new-synthetic-datasets-to-assist-covid-19-and-cardiovascular-research*.

**Wang, Z, Myles, P** and **Tucker, A.** June 2019. Generating and Evaluating Synthetic UK Primary Care Data: Preserving Data Utility & Patient Privacy. In: *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*. Cordoba, Spain: IEEE. pp. 126–131. ISBN: 9781728122861. URL: *https://ieeexplore.ieee.org/document/8787436/* (visited on 10/02/2020). DOI: *https://doi.org/10.1109/CBMS.2019.00036*

**Wei, Y,** et al. June 2020. *A systematic review and meta-analysis reveals long and dispersive incubation period of COVID-19*. en. preprint. Infectious Diseases (except HIV/AIDS). URL: *http://medrxiv.org/lookup/doi/10.1101/2020.06.20.20134387* (visited on 10/14/2020). DOI: *https://doi.org/10.1101/2020.06.20.20134387*

**Wu, SL,** et al. Dec. 2020. Substantial underestimation of SARS-CoV-2 infection in the United States. en. In: *Nature Communications*, 11(1): 4507. ISSN: 2041-1723. URL: *http://www.nature.com/articles/s41467-020-18272-4(visited* on 03/02/2021). DOI: *https://doi.org/10.1038/s41467-020-18272-4*

TO CITE THIS ARTICLE:
Popper, N, Zechmeister, M, Brunmeir, D, Rippinger, C, Weibrecht, N, Urach, C, Bicher, M, Schneckenreither, G and Rauber, A. 2021. Synthetic Reproduction and Augmentation of COVID-19 Case Reporting Data by Agent-Based Simulation. *Data Science Journal*, 20: 16, pp. 1–13. DOI: *https://doi.org/10.5334/dsj-2021-016*

Submitted: 10 November 2020
Accepted: 13 March 2021
Published: 27 April 2021



*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.