La Medicina del Lavoro

Original article

Med. Lav. 2024; 115 (1): e2024007
DOI: 10.23749/mdl.v115i1.14690

Determinants of COVID-19 Infection Among Employees of an Italian Financial Institution

Roberta De Vito[1], Martina Menzio[2], Pierluigi Lacqua[2], Stefano Castellari[2], Alberto Colognese[2], Giulia Collatuzzo[3], Dario Russignaga[4], Paolo Boffetta[3,5,*]

[1]Department of Biostatistics and Data Science Institute, Brown University, Providence, RI, USA
[2]Direzione Centrale Data Office, Data Science & Artificial Intelligence, Intesa Sanpaolo, Italy
[3]Department of Medical and Surgical Sciences, University of Bologna, Bologna, Italy
[4]Tutela Aziendale, Intesa Sanpaolo, Italy
[5]Stony Brook Cancer Center, Stony Brook University, Stony Brook, NY, USA

Keywords: COVID-19; Transition Probabilities; COVID-19 Infection Trends; Multi-State Model; COVID-19 Transmission; Risk Assessment

Abstract

Background: *Understanding the trend of the severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2) is becoming crucial. Previous studies focused on predicting COVID-19 trends, but few papers have considered models for disease estimation and progression based on large real-world data.* **Methods:** *We used de-identified data from 60,938 employees of a major financial institution in Italy with daily COVID-19 status information between 31 March 2020 and 31 August 2021. We consider six statuses: (i) concluded case, (ii) confirmed case, (iii) close contact, (iv) possible-probable contact, (v) possible contact, and (vi) no-COVID-19 or infection. We conducted a logistic regression to assess the odds ratio (OR) of transition to confirmed COVID-19 case at each time point. We also fitted a general model for disease progression via the multi-state transition probability model at each time point, with lags of 7 and 15 days.* **Results:** *Employment in a branch versus in a central office was the strongest predictor of case or contact status, while no association was detected with gender or age. The geographic prevalence of possible-probable contacts and close contacts was predictive of the subsequent risk of confirmed cases. The status with the highest probability of becoming a confirmed case was concluded case (12%) in April 2020, possible-probable contact (16%) in November 2020, and close contact (4%) in August 2021. The model based on transition probabilities predicted well the rate of confirmed cases observed 7 or 15 days later.* **Conclusion:** *Data from industry-based surveillance systems may effectively predict the risk of subsequent infection.*

1. Introduction

Severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2), which causes Coronavirus disease 2019 (COVID-19), is responsible for substantial morbidity and mortality worldwide [1, CDC Data Tracker]. Since it emerged in late 2019, Coronavirus spread soon across continents and became a global pandemic. The largest number of cases were initially identified in China, but the virus' diffusion occurred quickly across other regions. In particular, Italy was hardly hit in the early phase of the pandemic [1], experiencing a mortality rate higher than most other countries, with 33,601 deaths [2] between February and July 2020.

Received 03.05.2023 – Accepted 11.01.2024

*Corresponding Author: Paolo Boffetta; E-mail: paolo.boffetta@stonybrookmedicine.edu

Health authorities have been deeply involved in trying to control infection transmission, adopting social distancing, smart working, and travel restrictions [3]. This is to contain the pandemic and prevent the overwhelming healthcare system [4].

Understanding the reasons why Italy was highly burdened by COVID-19 deaths could help identifying and planning optimal actions in future epidemics, both in Italy and in other countries [5]. Implementing public health interventions and targeting health system's efforts are based on population-level prevalence estimates and predictions of infection and disease. For this reason, assessing factors that can influence the prevalence of infection and disease can be crucial to preventing and predicting the trend of COVID-19 [1]. For example, seasonality has long been recognized as an important aspect of many viral infections [5-7]. Besides this, some contributing factors may be invariant, i.e., demographic characteristics, such as the age structure of a population. Conversely, some factors are potentially modifiable including the risk of transmission, which can be minimized by adopting hygienic precautions like hand sanitation, social distancing, and use of masks.

Considering the rapid spread of the disease in different waves, working units and providers must be prepared as much as possible before the infection has reached uncontrolled rates. This is critically important for public health system, which managed a markedly increased rate of hospitalization, as well as for the economic system, which experienced a severe slowdown [8]. Indeed, the COVID-19 pandemic constituted an unexpected challenge for international economy [8, 9]. It is, therefore, essential for a government to design a well-assessed and comprehensive policy to enable the country to recover from the crisis.

In this big framework, industry-based occupational surveillance systems can be crucial for monitoring the spread of infection. One study applied a Fama and French Three-Factor Model (pricing model developed in 1992) [10] to the US service to compare the performances of the service industries before and after COVID-19 [11]. This work highlighted the potential of industry-based surveillance, focusing on finance and returned stocks. Other studies [12] addressed the effect of travel limitations [13], psychological distress [14], modes of transmission (i.e., detailed travel and exposure history and identification of high-risk subjects) [15], and the occupational risks of COVID-19, for example underlying several occupational groups at increased risk of infection [16]. Also, some studies focused on predicting the risk of COVID-19 in the general population, for example, using data from hospital admission for non-COVID-19 diseases (non-tuberculosis pneumonia, influenza, acute bronchitis) [17], or using machine learning to assess the benefit of the mask [18]. Finally, some studies focused on prognostic models for patients diagnosed with COVID-19 aimed to predict progression to a more severe or critical status [19-22]. To our knowledge, no study focused on industry surveillance systems to estimate and predict the incidence of COVID-19 and to provide guidelines on infection control in the occupational setting.

This paper aims to provide the overall socio-economic picture during the COVID-19 pandemic and address the estimation and prediction of the incidence of COVID-19 in an occupational setting, adopting efficient regression and transition statistical approaches.

We used the database of a large Italian Financial Institution with daily information on the COVID-19 status for all employees to estimate at each point in time the risk of COVID-19, the probability of transitioning from one status to another, and the prediction of the confirmed cases in the following weeks.

2. Materials and Methods

2.1. Data Sources

The analysis was based on the anonymized file of contacts and infections reported daily between March 31, 2020, and August 31, 2021, to the Occupational Safety and Health Department (OSHD) of a major financial institution in Italy. The institute staff was distributed between central offices, mainly located in large Italian cities, and a large number of branches in all regions of the country, with a predominance in Northern regions.

Starting on March 31, 2020, each central office and branch was requested to report daily to the

OSHD the employees who fulfilled any of the following definitions:

- Concluded – Individuals that have been declared no-COVID-19 after having been declared confirmed cases;
- Confirmed – Positivity to a COVID-19 test (antigenic or molecular);
- Close contacts – A close contact with a positive or suspect positive individual (a colleague, a relative, etc);
- Possible contacts – A contact with a positive or suspect positive individual (a colleague, a relative, etc);
- Possible-probable contacts – Individual that satisfies the clinical criteria with an epidemiological link (i.e., waiting for the test result);
- No-Covid – Healthy person who has never been in any of the previous statuses.

  At the onset of the pandemic, the classification of individuals occurred according to a 4-tier system, including the so-called "possible case" that had to meet the following criteria:
- Symptoms (at least 1 primary or at least 2 secondary) not yet evaluated by the Public Health System (General Practitioner or Local Health Authority);
- Positive results in an antigen test, both symptomatic and asymptomatic until the execution of the molecular test;
- Symptomatic positives in a serological test until the execution of the molecular test;
- Presence of at least one symptom and being part of a working group where other colleagues with symptoms have been present in the last 14 days.

A "probable case" could be attributed to an individual who met clinical criteria: cough, fever, dyspnea, acute onset of anosmia/ageusia/dysgeusia, or radiological criteria consistent with COVID-19. Moreover, the "possible case" had not yet been taken over by the public health system, prudentially classified at the company level, while the "probable case" had already been taken over by the national health system.

Confirmed cases, close contacts, and possible-probable contacts were asked to quarantine according to the guidelines of the Ministry of Health and to inform the OSHD when their status was resolved. Quarantine is a preventive measure taken to separate and restrict the movement of individuals who may have been exposed to a contagious disease, such as COVID-19, to see if they develop symptoms. Subjects not included in any of the categories above were defined as negative (i.e., no-COVID-19). Precisely, not all the employees were registered in the COVID-19 bank database. To enter the COVID-19 bank database, a person should report close contact, possible contact, possible-probable contacts, or be a confirmed case of COVID-19.

Additional data sources were the anonymized file of all institution employees and the daily data of cases in the general population at the provincial level obtained from the National Authority for Civil Defense [23].

The following variables were analyzed at the individual level: sex, age (four categories, <45, 45-49, 50-54, 55+), place of work (address), type of employment (central office vs. branch), contact/case status. The large number of places of work required some grouping into units of adequate size for statistical analysis. After some preliminary analysis, we identified regional and 107 provincial units used for adjustment purposes. Their list is shown in Appendix Table 1.

The geographic unit used in the comparative analysis between the institute staff and the general population was the province (N=107). In all analyses, prevalence rates per 1,000 subjects were used, together with their Confidence Intervals (CI). Possible contacts and concluded cases were excluded from some analyses due to the small number of subjects.

2.2. Analysis of Prevalence, and Comparison with General Population

We analyzed the trend over time of the daily prevalence of each status between March 31, 2020, and August 31, 2021, and the prevalence ratio between workers in the branches and those in the central offices. We also compared the incidence of confirmed cases among subjects in the study population between 1 April 2020 and 31 August 2020 with that of the general Italian population, after standardization for the size of the population of each province. We calculated 7-day rolling means because of fluctuations in the daily incidence of the study population.

2.3. Individual-Level Analysis of Determinants of Prevalence of Infections and Contacts

Two sets of logistic regression models were fitted to the individual data, with the status of confirmed case, close contact, and possible-probable contact on each day between March 31, 2020, and August 31, 2021, as dependent variables. The first set of models included sex, age, and type of employment as potential determinants, the second set included also the geographical unit (region). In the analysis of each status, the cases of the other statuses were excluded to obtain a common reference category (i.e., no-COVID-19). The results are expressed as the odds ratio (OR) for each status, including their 95% Confidence Interval (CI).

2.4. Geographic Analysis of Determinants of Prevalence of Confirmed Cases

Multivariate linear regressions were performed with the prevalence of confirmed cases on August 31, 2021, as the dependent variable and the prevalence of possible contact, possible-probable contact, and close contact on each observation day between March 31, 2020, and August 31, 2021, as independent variables. They are reported in terms of z values, the normal deviation of the regression parameters of each contagion or contact indicator, for each day between 31 March 2020 and 31 August 2021. Confidence Intervals are provided, thus values of z higher than 1.96 or lower than -1.96 denote statistically significant associations (at α=0.05) between the indicator and the prevalence of confirmed cases as of 31 August 2021.

2.5. Analysis of Transition of Status

This analysis included prediction and estimations of the transition probability from one status to another, along the continuum from no-COVID-19 cases to confirmed cases. We carried out multi-state models [24, 25]. The status s(t) at which an individual moves at time t is conducted by a set of transition probabilities, qrs(t), where r and s are two different statuses. The intensities and estimation also depend on the time of the process t (i.e., which intervals of time, or lag, were chosen). For this reason, we choose two time-intervals, 7-day and 15-day lags. To fit a multi-state model to data, we need to estimate this transition intensity matrix, which identifies the immediate risk of moving from status *r* to status *s*:

$$q\_rs\ (t)=\text{lim┬}(\delta t\rightarrow 0)\ \text{〖}P(S(t+\delta t)=s \,\Big|\, S(t)=r)/\delta t\text{〗}$$

This analysis focused on Markov models [26, 27], assuming that future evolution only depends on the current status. We estimated the transition matrix via a likelihood approach. At each time point, we computed the corresponding time t transition probabilities [28-30].

Generally, the late status of the multi-state model is an "absorbing status", i.e., death. Thus, if a subject enters this status, the subject will remain with probability 1. In this approach, we assume that our last status is not absorbing, i.e., all the statuses have the same probability of transition. We also included sex, age, and type of employment in the model as potential confounders. CIs were computed with the maximum likelihood approach. All statistical analyses were carried out using the open-source statistical computing environment Python, with its libraries "MSMBuilder (3.8.0)" [31], "SciPy (1.5.4)" [32], "StatsModel (0.12.1)" [33].

3. Results

3.1. Study Population Characteristics

Table 1 shows socio-demographic characteristics of the entire study population. The largest age group was over 55 (32%), followed by participants aged between 50-54 (22%). The majority of the study subjects were located in a branch (59%) compared to the central office (41%). The Regions with the largest number of subjects were Lombardy (29%),

Table 1. Baseline socio-demographic characteristic of the study population.

Variable and category	Total	(%)
Sex		
Female	31,646	51.93%
Male	29,292	48.07%
Age(yr)		
< 45	18,184	29.84%
45-49	9,836	16.14%
50-54	13,618	22.35%
55+	19,300	31.67%
Site-location		
Central office	25,058	41.12%
Branch	35,880	58.88%
Region		
Abruzzo	752	1.23%
Basilicata	274	0.45%
Calabria	693	1.14%
Campania	3,831	6.29%
Emilia-Romagna	3,991	6.55%
Friuli Venezia Giulia	1,220	2.00%
Lazio	4,260	6.99%
Liguria	951	1.56%
Lombardia	17,568	28.83%
Marche	1,178	1.93%
Molise	106	0.17%
P.A. Bolzano	125	0.21%
P.A. Trento	231	0.38%
Piemonte	7,699	12.63%
Puglia	2,408	3.95%
Sardegna	963	1.58%
Sicilia	1,941	3.19%
Toscana	4,006	6.57%
Umbria	760	1.25%
Valle d'Aosta	106	0.17%
Veneto	7,875	12.92%
Covid Status (April 2020)		
Concluded	98	0.16%
Confirmed	179	0.29%
Close contacts	3,537	5.80%
Possible contacts	113	0.18%
Possible-probable contacts	543	0.89%
No-Covid	60	0.10%
Total	4,530	7.43%
Covid Status (November 2020)		
Concluded	9,536	15.65%
Confirmed	1,872	3.07%
Close contacts	7,240	11.88%
Possible contacts	238	0.39%
Possible-probable contacts	2,360	3.87%
No-Covid	60	0.10%
Total	21,306	34.96%
Covid Status (August 2021)		
Concluded	18,304	30.04%
Confirmed	4,764	7.82%
Close contacts	11,637	19.10%
Possible contacts	1,210	1.99%
Possible-probable contacts	3,809	6.25%
No-Covid	60	0.10%
Total	39,784	65.29%

Veneto (13%) and Piedmont (13%). Also, more participants were women (52%) than men (48%).

Moreover, Table 1 reports the COVID-19 statuses at different time points (i.e., April 2020, November 2020, and August 2021). Only the employees that reported confirmed, close contacts, possible contacts and possible probable contacts were collected in the data set, thus the total of COVID-19 statuses differs from one time to another. The total COVID status increases over time. Specifically, in April 2020, the total COVID statuses was 7.43% of the total population, with the majority of the cases reported to be close contacts (5.80%). In November 2020, the total COVID status was 34.96% of the total population, distributed almost equally between concluded (15.65%) and close contacts (11.88%). In August 2021, the total COVID status was 65.29%, mostly concentrated in the concluded status (30.04%) and close contacts (19.10%).

3.2. Comparison of Incidence with General Population

The comparison of the 7 days rolling mean of new confirmed cases rate between the Italian population and the study population showed that the infection rate of the study population followed the national trend, but it was lower, in particular at the beginning of the pandemic (Supplementary Figure 1, Figure 1).

3.3. Individual-Level Analysis of Determinants of Prevalence of Infections and Contacts

The analysis of individual data (Table 2) showed a strong association between all indicators analyzed

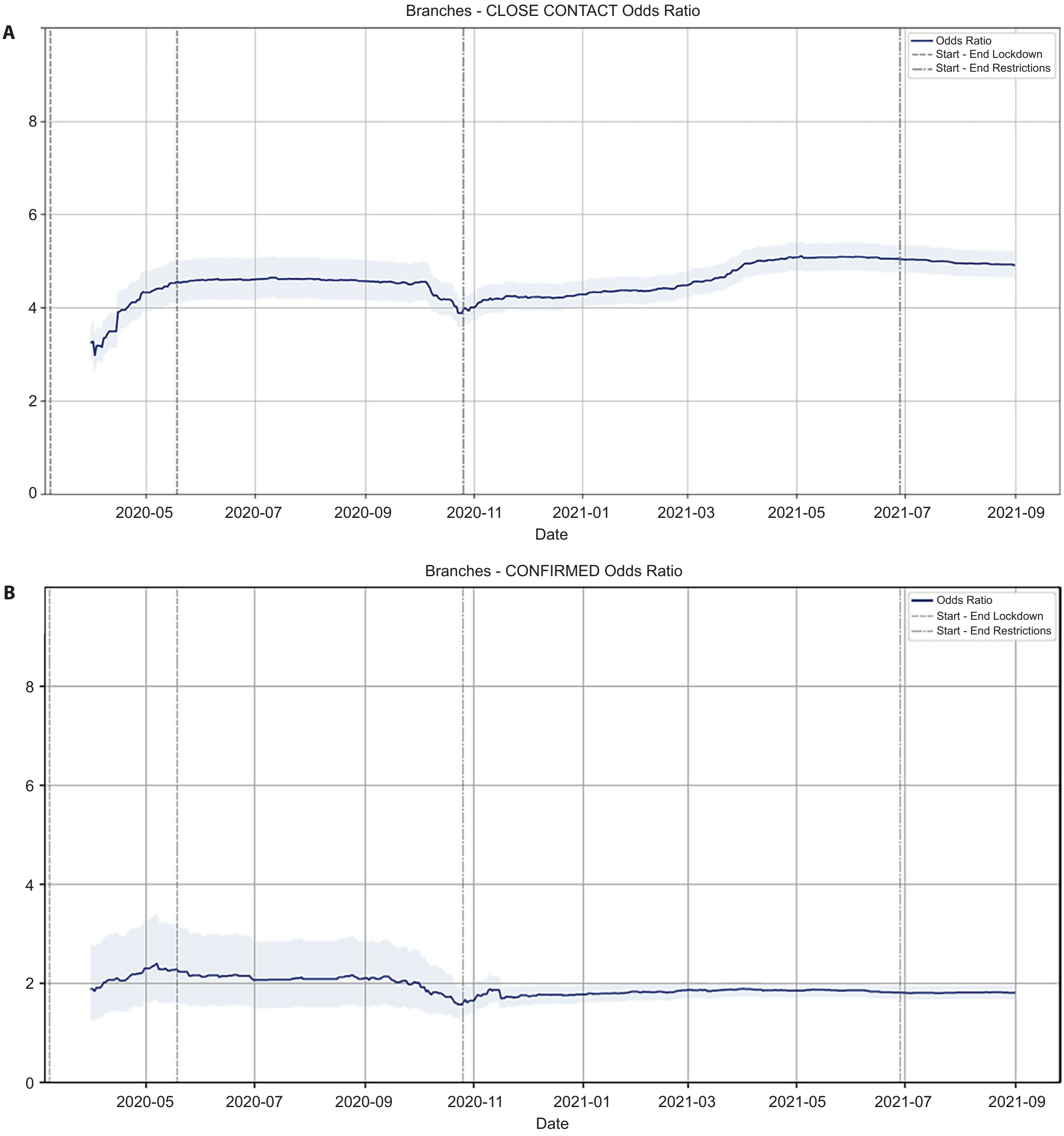

Figure 1. Odds ratio of confirmed case (A) close contact (B) and possible-probable contact (C) in branch workers vs central office workers, by date. Vertical lines indicate dates of lock-down strategies in Italy.

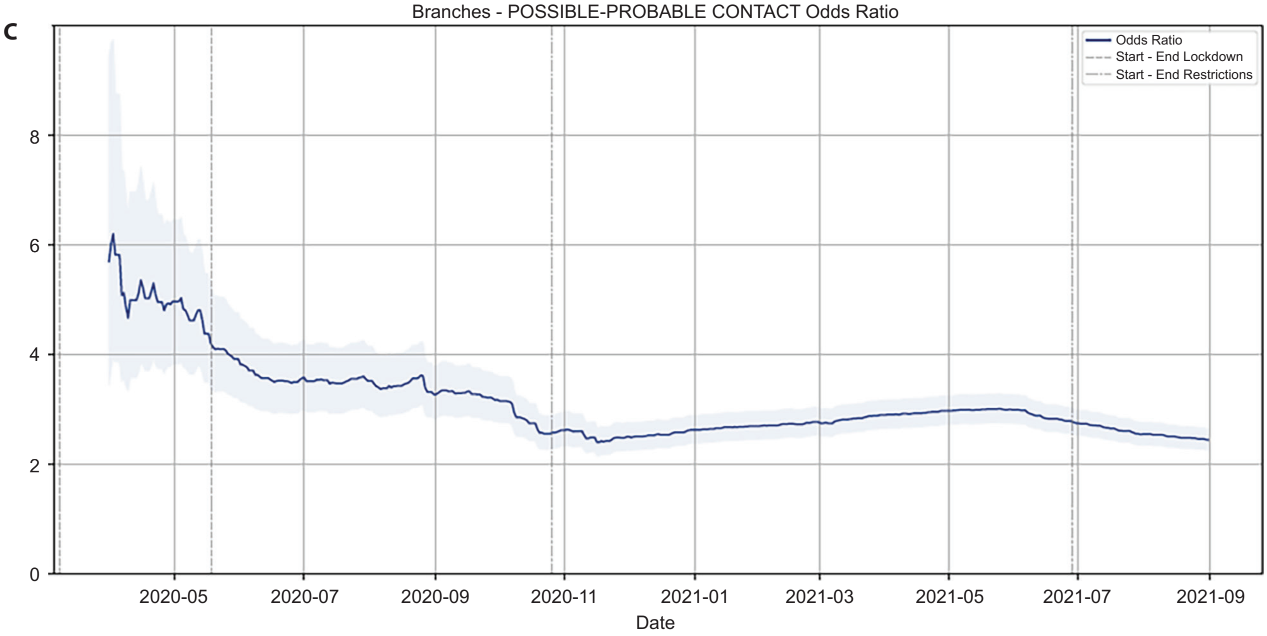

Figure 1. (C) (*Continued*)

and employment in a branch. The risk for the female was always higher than for male in all the statuses and for all the time points considered, the highest is the possible-probable contact on April 2020 (OR=1.77, CI: 1.27-2.47). The risk of close contact was higher in the age group 50-54 than in other age groups in all time points considered (April 2020: OR=1.39, CI: 1.23-1.58; November 2020: OR=1.12, CI: 1.04-1.20; August 2021: OR=1.17, CI: 1.10-1.24). The risk of possible-probable contact was higher in the age group below 45 years than in older age groups in August 2021, with the minimum risk in the age group 55+ (OR=0.72, CI: 0.66-0.79).

The comparison between employment in the central office and in-branch showed, for all indicators except possible-probable contact, a 3-4 times higher prevalence among those working in a branch compared to those working in central offices, with an increase of this ratio over time, with its peak for the close contact status on August 2021 (OR=4.92, 95% CI: 4.66-5.19). The risk of the confirmed Covid status is not significant for the gender and the age variables included in the model, and it is significant for the site variable, with higher risk for the bank branch than the bank central site in all time points (April 2020: OR=1.89, CI: 1.29-2.77; November 2020: OR=1.66, CI: 1.42-1.95; August 2021: OR=1.82, CI: 1.70-1.94). The analysis, including geographical units (Supplementary Table 1), showed an excess of possible-probable contacts and close contacts in the provinces of Lombardy and other areas with high a general population rate of infection during the corresponding period.

3.4. Geographic Analysis of Determinants of Prevalence of Confirmed Cases

The results of the analysis excluding provinces with no confirmed cases up to 31 August 2021 are presented in Supplementary Figure 2. The prevalence of close contacts showed the strongest association with that of confirmed cases; the prevalence of possible-probable contact was also strongly associated during the last months of the observation period. Corresponding analyses, including all provinces, are reported in Supplementary Figure 3. These results highlight how the prevalence of close contacts and of possible-probable contacts are predictive of the burden of confirmed cases.

Table 2. Odds ratio[a] (ORs) of each status and the corresponding 95% confidence intervals (minimum and maximum), for three time periods: April 2020; November 2020; and August 2021.

		April 2020			November 2020			August 2021		
Variable and category	**Status**	**OR**	**Min CI**	**Max CI**	**OR**	**Min CI**	**Max CI**	**OR**	**Min CI**	**Max CI**
Sex (M)	Close Contact	1[b]	-	-	1[b]	-	-	1[b]	-	-
Sex (F)	Close Contact	1.13	1.03	1.24	1.09	1.03	1.15	1.09	1.05	1.14
Age (< 45)	Close Contact	1[b]	-	-	1[b]	-	-	1[b]	-	-
Age (45-49)	Close Contact	1.21	1.05	1.40	1.06	0.97	1.15	1.10	1.03	1.17
Age (50-54)	Close Contact	1.39	1.23	1.58	1.12	1.04	1.20	1.17	1.10	1.24
Age (55+)	Close Contact	1.22	1.08	1.39	1.01	0.94	1.09	1.03	0.97	1.08
Site (Bank Central Site)	Close Contact	1[b]	-	-	1[b]	-	-	1[b]	-	-
Site (Bank Branch)	Close Contact	3.26	2.89	3.67	4.01	3.73	4.31	4.92	4.66	5.19
Sex (M)	Confirmed	1[b]	-	-	1[b]	-	-	1[b]	-	-
Sex (F)	Confirmed	0.88	0.62	1.24	0.92	0.79	1.06	1.03	0.97	1.09
Age (< 45)	Confirmed	1[b]	-	-	1[b]	-	-	1[b]	-	-
Age (45-49)	Confirmed	0.77	0.42	1.41	0.98	0.78	1.23	1.07	0.98	1.17
Age (50-54)	Confirmed	0.98	0.59	1.62	1.01	0.82	1.23	1.06	0.97	1.15
Age (55+)	Confirmed	1.50	0.98	2.29	1.10	0.91	1.32	1.02	0.94	1.10
Site (Bank Central Site)	Confirmed	1[b]	-	-	1[b]	-	-	1[b]	-	-
Site (Bank Branch)	Confirmed	1.89	1.29	2.77	1.66	1.42	1.95	1.82	1.70	1.94
Sex (M)	Possible-probable contact	1[b]	-	-	1[b]	-	-	1[b]	-	-
Sex (F)	Possible-probable contact	1.77	1.27	2.47	1.28	1.16	1.42	1.15	1.08	1.23
Age (< 45)	Possible-probable contact	1[b]	-	-	1[b]	-	-	1[b]	-	-
Age (45-49)	Possible-probable contact	1.12	0.70	1.78	1.01	0.88	1.16	0.97	0.88	1.07
Age (50-54)	Possible-probable contact	1.00	0.65	1.54	0.91	0.80	1.03	0.89	0.82	0.98
Age (55+)	Possible-probable contact	1.17	0.79	1.75	0.77	0.68	0.88	0.72	0.66	0.79
Site (Bank Central Site)	Possible-probable contact	1[b]	-	-	1[b]	-	-	1[b]	-	-
Site (Bank Branch)	Possible-probable contact	5.70	3.45	9.44	2.63	2.34	2.95	2.45	2.27	2.65

[a] *Obtained from a logistic regression models.*

[b] *Reference category.*

3.5. Analysis of Transition of Status

Table 3 presents the estimated transition probability matrix qrs for the seven-day lag observed in three different time frames: March 30 – April 30, 2020; November 1 – November 30, 2020; and August 1 – August 31, 2021. The diagonal of this matrix provides the estimated probability of remaining in that particular status. In April 2020, the status that shows the lowest probability of remaining the same (86%) is that of concluded case. In November 2020, the status that shows the lowest probability of remaining the same (52%) is that of close contact, followed by that of possible-probable contact (54%). Finally, in August 2021, the status that shows the lowest probability of remaining the same is possible contact (39%), followed by close contact (43%) and possible-probable contact (44%).

When focusing on the transition from one status to another, i.e., the non-diagonal element of the transition probability matrix, we found that the status with the greatest probability of becoming a

Table 3. Transition probability matrix q_{rs} with 7 lag days, at April 2020, November 2020, and August 2021.

Status	Period	Concluded	Confirmed	Close contact	Possible contact	Possible Probable	No-Covid
Concluded	April 2020	0.86	0.12	0.02	0.00	0.00	0.00
Confirmed	April 2020	0.00	0.97	0.00	0.00	0.01	0.01
Close contact	April 2020	0.00	0.00	0.99	0.00	0.00	0.00
Possible contact	April 2020	0.00	0.00	0.00	1.00	0.00	0.00
Possible Probable	April 2020	0.00	0.01	0.00	0.00	0.97	0.02
No-Covid	April 2020	0.00	0.00	0.01	0.00	0.00	0.99
Concluded	November 2020	0.95	0.04	0.00	0.00	0.00	0.00
Confirmed	November 2020	0.27	0.73	0.00	0.00	0.00	0.00
Close contact	November 2020	0.00	0.04	0.52	0.00	0.01	0.43
Possible contact	November 2020	0.00	0.01	0.02	0.70	0.00	0.27
Possible Probable	November 2020	0.01	0.16	0.00	0.00	0.54	0.30
No-Covid-19	November 2020	0.00	0.00	0.01	0.00	0.00	0.98
Concluded	August 2021	1.00	0.00	0.00	0.00	0.00	0.00
Confirmed	August 2021	0.29	0.71	0.00	0.00	0.00	0.00
Close contact	August 2021	0.00	0.04	0.43	0.00	0.00	0.53
Possible contact	August 2021	0.00	0.00	0.00	0.39	0.00	0.61
Possible Probable	August 2021	0.01	0.03	0.00	0.00	0.44	0.52
No-Covid	August 2021	0.00	0.00	0.00	0.00	0.00	1.00

confirmed case was concluded case (12%) in April 2020; possible-probable contact (16%) in November 2020, and close contact (4%) in August 2021. In April 2020, the remaining transition probabilities remained low (around 0.2-0.3%). In November 2020, the transition from close contact to no-Covid had the highest probability (43%), and the probabilities of transition from possible contact and possible-probable contact to no-Covid status were also high (27% and 30%, respectively), while that of the transition from possible-probable contact to the confirmed case was 16%. The probability of transition from close contact to confirmed case was 0.4%, and that from possible contact to confirmed case was 0.2%. Overall, the status with the greatest probability of becoming confirmed case was probable contact.

In August 2021, the transition from possible contact to no-Covid status had the highest probability (61%), followed by the transition from close contact and possible-probable contact to no-Covid status (53% and 52%, respectively). The remaining transition probabilities were lower than 0.5%.

Supplementary Table 5 shows the transition probability matrix with 15 days lag in the three-time frames, April 2020, November 2020, and August 2021. Supplementary Figure 4 shows the transition probability with 15 days lag, and the estimated probability of each status, with a lag of 15 days during the whole study period. These results confirm those based on a lag of 7 days (Table 3 and Supplementary Figure 6).

Supplementary Figures 4a and 4b show the estimated probability of transitioning from each status to that of confirmed case by gender, with a lag of 7 days. In general, there were no differences between women and men. Minor discrepancies were observed in July 2020, with a minimum for women compared with men, and in September 2020, with a peak for men compared to women. After November 2020, the two lines have a similar trend.

Supplementary Figure 6 shows the estimated probability of transitioning from each status to that of confirmed case, with a lag of 7 days. The status with the highest transition probability of becoming a confirmed case, excluding the confirmed case itself, is possible-probable contact. The grey dashed lines in the plot indicate the three different lock-down strategies adopted in Italy after the two first peaks of the disease. All three lock-down strategies appeared to have a beneficial effect. The probability of transitioning to the confirmed case decreased after each lock-down for all the statuses in the following 7 days.

Supplementary Figure 7 shows the probability of infection separately for central office and branch workers, with a lag of 7 days. Central offices showed a higher probability of transitioning to the confirmed case than the branch workers; conversely, the opposite trend was suggested for possible-probable contact.

Figure 2 compares the daily number of actual confirmed cases with the corresponding prediction derived from the multi-status transition model, with a lag of 7 days. The model predicted well the actual confirmed cases. Similar results were obtained with a lag of 15 days (Supplementary Figure 5).

4. Discussion

This is a large longitudinal study conducted within a major financial institution in Italy, based on the collection of individual-level daily information on COVID-19 status of all the employees. Our aim was to capture the overall socio-economic picture during the COVID-19 pandemic, providing new information on risk of infection in an occupational setting.

Two different approaches were considered overlaying six different COVID-19 statuses: (i) concluded case, (ii) confirmed case, (iii) close contact, (iv) possible-probable contact, (v) possible contact, and (vi) no contact or infection (no-Covid). The first approach consisted of logistic regression model conducted to assess the OR of transition to confirmed COVID-19 statuses at each time point; the second approach relied on the multi-state Markov model to estimate the probability of

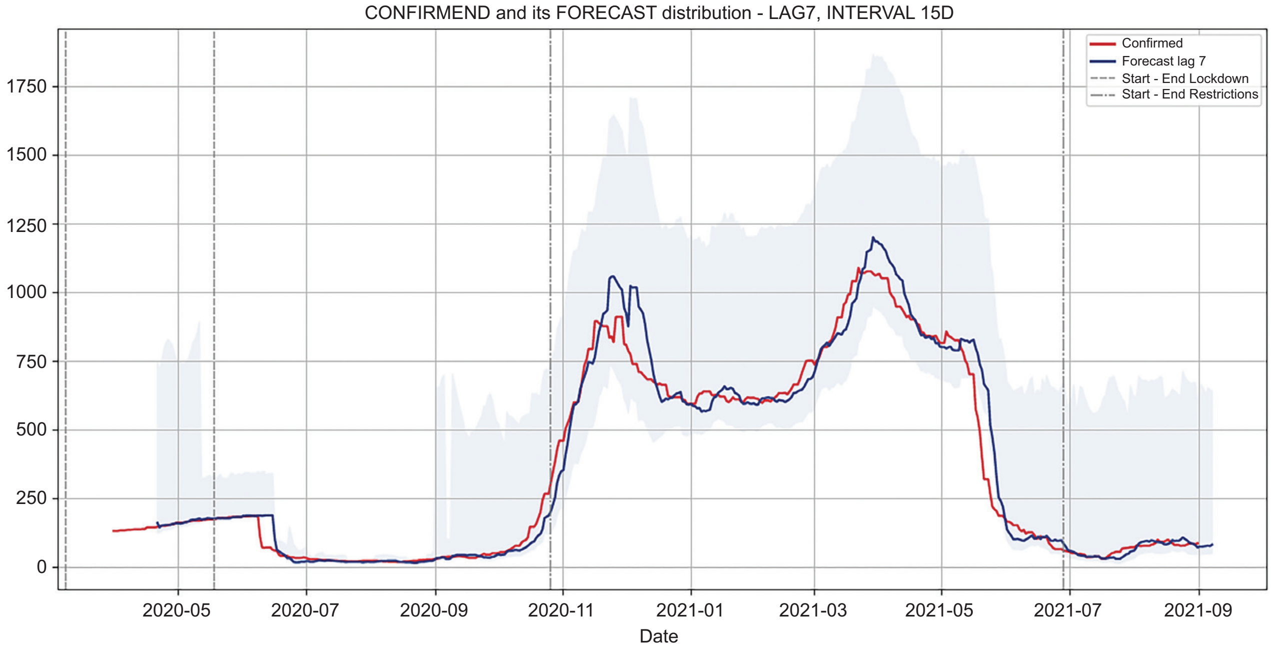

Figure 2. Number of confirmed cases (red line, 15 day intervals) and predicted number with 7-day lag.

Grey area indicates 95% Confidence Intervals.

Vertical lines indicate dates of lock-down strategies in Italy.

transition of each status at each time point. These methods optimally estimated and predicted the spread of COVID-19 at each time point, with a lag of 7 or 15 days.

This is one of the few studies to provide an accurate prediction of COVID-19 infection among workers, otherwise suitable for potential application in other settings, including hospitals (both among health care workers and hospitalized patients) and schools.

Many mathematical and statistical approaches with different complexity have been developed to predict the consequences and spread of this epidemic. COVID-19 outbreaks prediction has been the object of previous studies [34, 35], primarily based on artificial intelligence and other modeling. For example, the re-opening of schools in the UK was questioned in a paper which applied a statistical predictive approach considering hypothetical situations and predicting the possible infection rates [36]. When high-quality data are available, models such as SIR [35] are powerful in estimating the dynamics of a spread of any epidemics, including that of COVID-19 [36-38]. However, SIR-based models rely on accurate initial estimates of the spreading mechanism that are often unknown. Also, these models are often inflated, overestimating the epidemic severity [39] and underestimating mobility. Mobility is a crucial key in this pandemic, both at the large level – across the world – and at the small level – Italy. This issue had still a role during the lock-down phases, e.g., through patients transfer from one hospital to another. Our empirical methods prevent these two issues from avoiding the specification of the initial mechanism of disease spread and allowing mobility across regions. Also, our study identified some critical characteristics to be taken into account in predictive models for COVID-19 transmission, including the job category. This confirms the usefulness and reliability of risk assessment in predicting the occurrence of infection.

Our analyses revealed some crucial patterns. First, at the beginning of the pandemic (April 2020), the likelihood of an employee falling in the status "concluded" was the lowest. This pattern reflects the high number of contacts that people experienced in a period when remote working was not yet widespread, and the use of personal protection equipment (PPE) was limited based on limited recommendations and lack of availability. Second, in November 2020, during the COVID-19 second wave, the Italian situation appeared different: close contact was the status less likely to be maintained. This pattern highlights the multiple contact experiences in a short period. Third, in November 2020, our models estimated a high probability of subjects in a close contact status transitioning into no-Covid status. This pattern indeed is consistent with the fact that at the time people in Italy were more prepared to manage the infection risk, i.e., by wearing masks, sanitizing hands and maintaining the recommended interpersonal distance [37, 38]. The major attention and knowledge of the infection also led to better management of the incident cases, including contact tracing and self-isolation.

Across the entire study period, possible-probable contact was the status with the highest probability of transition into a confirmed case. Thus, our methods and the overall monitoring system appear to be coherent and reliable, reproducing by and large the risk assessment procedures developed in the healthcare setting. Close contact was not the strongest predictor of COVID-19 infection, as expected. A possible reason is a low specificity in the definition of close contact. Also, after the vaccination campaign, a larger number of subclinical infections may have been present in 2021, which prevented the disease but not so much the infection itself [39]. SARS-CoV-2 variants do not seem to play a role in these results, as the Delta variant was first reported in fall 2021 [40].

We observed a lower risk of transition to confirmed status in correspondence to the lock-down phases and the summer. While this latter result was expected, given the naturally downscale of respiratory-born infection spreading with increasing temperature, the first result reflects the effectiveness of the restriction strategies adopted by Italy and in particular at the workplace. This was seen in both sexes, all ages and independently from the occupational location, as well as in all the Italian regions. When focusing on the lock-down periods, notably, we observed a drastic reduction of the infection rates in this working setting, suggesting the infection control obtained through the severe restriction policy adopted in the country. In general, the containing measures introduced at different times could have impacted the transmission of infection among workers. This is consistent with other studies investigating and predicting the infection trends in Italy [41, 42].

Individuals working in the branches were more likely to transition from any status to confirmed. This is consistent with the fact that central office workers started working from remote early on in the pandemic. Employees in branches were therefore more exposed to the public and thus had a higher possibility of being in contact with infected people, including colleagues, than central office workers. However, workers who continued to work in the branches, rather than switching to remote work, had opportunities of exposure other than the workplace, such as transportation and more active social exchanges. Their higher risk of infection cannot therefore be attributed only to occupational circumstances. The comparison between the study population and the general Italian population showed a lower incidence of new cases in the former: this was probably due to the fact that the company under study adopted early in the pandemic severe actions in order to protect the staff and the customers such as remote working, the obligation to use personal protective equipment and a massive tracking of contacts.

It should be noted that, despite the different risks, there was no difference regarding the timing of vaccination. Indeed, both fell into the general population category addressed with vaccination schedules from March 2021. In addition, these differences seem not to depend on the socio-demographic characteristics of the participants, given the multiple adjustments included in the models. The difference by employment site was particularly evident for possible-probable contact (up to 5 folded risks in the earliest phase of the pandemic).

The apparent higher likelihood of transitioning status registered in August 2021 rather than November 2020 may be partially explained by the exit from lock-down. Noticeably, Italy was not only the first country to be hit by COVID-19 in Europe, but also the first and the most severe in introducing compulsory use of PPE and other preventive measures, including the mandatory vaccination for workers. For example, mask-wearing was mandatory from 16th August 2020 in areas where social distancing was not possible, and from 24th September 2020 some Italian regions introduced more strict measures, i.e., the obligatory use of masks everywhere, also outdoor [43]. The re-opening of economic activities was predicted to be followed by another increase in the infection rates [44].

This study has some limitations. Our results are based on data collected by the occupational medical service and, in turn, reported by the employees themselves. In particular, as discussed above, the definition of close contacts suffered from low specificity; this was a deliberated choice, aimed at the protection of the workers. Such misclassification likely was non-differential for confirmed cases given the need for a test to confirm it.

Also, data on vaccination status and date of vaccination were not available, limiting the interpretation of the results. Finally, we lack clinical data, as the symptoms and health status of the employees were not known.

It should be highlighted that the large majority of the study population were from a defined geographic area (i.e., Lombardy, Veneto, and Emilia Romagna), where the majority of the bank offices of this institution are located. This did not impact any statistical analyses and methods, which reported good significant results and thus revealing robustness of the methods here adapted.

Adopting this or a similar approach in the occupational setting may help the organization of a company, offering one week to adopt preventive measures, and balancing the internal sources and improving the working activity in a specific worksite. The perspective of a translation of this predictive model to other settings, such as hospitals and schools, is potentially critical given the importance of a system able to predict the infection rate without the specification of the initial mechanism of disease spread and allowing mobility across locations, i.e., here regions. Thus, the statistical approaches adopted in this manuscript can direct the public efforts and help taking decisions (e.g., expanding emergency rooms and medical personnel, adopting distant learning).

The results of our study underline the importance of further investigating the characteristics of the infection and its impact on the economic sector to reinforce the measures needed to guarantee the health and protection of workers. It would be necessary to consolidate remote working and reinforce organizational support with coaching and training tools. This would help in containing the risk of infection in case of emergency such that of Sars-Cov-2 spreading without compromising productivity. Epidemiological studies on seroprevalence of Sars-Cov-2, including elements related to occupational variables, will constitute further important contributions to the context analysis. In this respect, it is important to stress that this is among the first examples of detailed surveillance study of the COVID-19 epidemics conducted on the whole workforce of a major financial institution.

5. Conclusions

In conclusion, this study describes an accurate predictive statistical approach for COVID-19 infection developed based on the daily, individual-level data collected by the occupational surveillance of the employees of a large financial institution in Italy. The definition of infection status allowed to effectively distinguish pattern of Sars-Cov-2 transmission in the workforce with good prediction level at each time point and could be suitable for other settings. The proposed approach might contribute to early identification and control of the outbreaks of COVID-19 infection at the occupational level in the future. Models based on similar characteristics might be useful for the management of other possible epidemics.

Supplementary Materials: The following are available in the online version: Supplemental material (text) S1, Figure S1: Geographic analysis of determinants of prevalence of infections and contacts. The dashed grey lines represent the three lock-down strategies adopted in Italy, Figure S2: a) Probability of being confirmed for males and female for all the statuses, lag of 7 days, observation period 03/2020 – 09/2021, b) Probability of being confirmed for males and female for all the statuses excluding the confirmed status itself, lag of 7 days, observation period 03/2020 – 09/2021, Figure S3. Distribution of the confirmed cases (red line) and the estimated confirmed cases with prediction of 15 days (blue line), for a lag of 15 days. The grey intervals represent the confidence intervals for our estimates. The vertical dashed grey line represents the two lock-down strategies adopted in Italy. Table S1: Sociodemographic characteristic of the population for each statuses considered, at August 31, 2021 Table S2: Odds ratio (ORs) of each status for each Italian region, and the corresponding 95% confidence intervals (minimum and maximum), for three time periods, Table S3: Transition probability matrix qrs with 15 lag days, for three time points: March 30,2020; November 1,2020; and August 1/2021.

Author Contribution Statement: PB designed research plan for approach 1; RDV and PB designed research plan for approach 2; MM, PL: assisted with the research plan and writing of the manuscript; DR provided advice and a critical review of the manuscript; AC contributed to collect and organize the data, MM performed most of the statistical analyses, with contributions from RDV and PB; SC contributed to generate the graphical representations; PL and GC assisted with interpretation of results; RDV, PB wrote the paper; GC assisted in writing the Discussion section; MM, PL and DR provided critical review of the manuscript. RDV and PB had primary responsibility for final content. All authors reviewed and approved the final manuscript. The authors have declared no conflicts of interest.

Funding: The work was conducted with internal resources of the participating institutions.

Institutional Review Board Statement: The study does not require ethics approval; all the subjects were de-identified.

Informed Consent Statement: The study does not require ethics approval; all the subjects were de-identified.

Declaration of Interest: The authors declare no conflict of interest.

References

1. Boccia S, Ricciardi W, Ioannidis JPA. What Other Countries Can Learn from Italy during the COVID-19 Pandemic. *JAMA Intern Med.* 20201;180(7):927-928. Doi: 10.1001/jamainternmed.2020.1447.
2. Worldometer – COVID-19 CORONAVIRUS PANDEMIC. Available online: https://www.worldometers.info/coronavirus/.
3. Dowd JB, Andriano L, Brazel DM, et al. Demographic science aids in understanding the spread

and fatality rates of COVID-19. *Proc Natl Acad Sci USA*. 2020;117(18):9696-9698. Doi: 10.1073/pnas.2004911117/-/DCSupplemental

4. Aabed K, Lashin MMA. An analytical study of the factors that influence COVID-19 spread. *Saudi J Biol Sci*. 2021;28(2):1177-1195. Doi: 10.1016/j.sjbs.2020.11.067
5. Lowen AC, Steel J. Roles of Humidity and Temperature in Shaping Influenza Seasonality. *J Virol*. 2014;88(14):7692-5. Doi: 10.1128/jvi.03544-13
6. M. Moriyama, Hugentobler WJ, Iwasaki A. Seasonality of Respiratory Viral Infections. *Annu Rev Virol*. 2020;7(1):83-101. Doi: 10.1146/annurev-virology-012420
7. Thai PQ, Choisy M, Duong TN, et al. Seasonality of absolute humidity explains seasonality of influenza-like illness in Vietnam. *Epidemics*. 2015:13:65-73. Doi: 10.1016/j.epidem.2015.06.002
8. Martin A, Markhvida M, Hallegatte S, Walsh B. Socio-Economic Impacts of COVID-19 on Household Consumption and Poverty. *Econ Disaster Clim Chang*. 2020;4(3):453-479. Doi: 10.1007/s41885-020-00070-3
9. Fernandes N. Economic effects of coronavirus outbreak (COVID-19) on the world economy. 2020. Available online: https://ssrn.com/abstract=3557504
10. Fama EF, French KR. The Cross-Section of Expected Stock Returns. *J Finance*. 1992;47(2):427-465. Doi: 10.1111/j.1540-6261.1992.tb04398.x.
11. Liu S. Analysis of COVID-19 on service industry based on fama and French five-factor model. Proceedings – 2020 Management Science Informatization and Economic Innovation Development Conference, MSIEID 2020, Dec. pp. 154-157. Doi: 10.1109/MSIEID52046.2020.00035
12. Priyadarshini I. A Survey on some of the Global Effects of the COVID-19 Pandemic. 2020. Doi: 10.21203/rs.3.rs-20842/v1
13. Chinazzi M, Davis JT, Ajelli M, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. 2020;36B(6489):395-400. Doi: 10.1126/science.aba97
14. Qiu J, Shen B, Zhao M, Wang Z, Xie B, Xu Y. A nationwide survey of psychological distress among Chinese people in the COVID-19 epidemic: Implications and policy recommendations. *Gen Psychiatr*. 2020 ;33(2):e100213. Doi: 10.1136/gpsych-2020-100213
15. Wong JEL, Leo YS, Tan CC COVID-19 in Singapore-Current Experience: Critical Global Issues That Require Attention and Action. *JAMA*. 2020;323(13):1243-1244. Doi: 10.1001/jama.2020.2467
16. Koh D. Occupational risks for COVID-19 infection. *Occup Med*. 2020;70(1):3-5. Doi: 10.1093/occmed/kqaa036
17. DeCaprio D, Gartner J, Burgess T et al. Building a COVID-19 Vulnerability Index. Arxiv 2020. Available online: http://arxiv.org/abs/2003.07347
18. Jiang Z, Hu M, Fan L, et al. Combining Visible Light and Infrared Imaging for Efficient Detection of Respiratory Infections such as COVID-19 on Portable Device. Arxiv 2020. Available online: http://arxiv.org/abs/2004.06912
19. Knight SR, Ho A, Pius R, et al. Risk stratification of patients admitted to hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: Development and validation of the 4C Mortality Score. *BMJ*. 2020:370:m3339. Doi: 10.1136/bmj.m3339
20. Zhang H, Shi T, Wu X, et al. Risk prediction for poor outcome and death in hospital in-patients with COVID-19: derivation in Wuhan, China and external validation in London, UK. medRxiv. 2020; Doi: 10.1101/2020.04.28.20082222
21. Barda N, Riesel D, Akriv A, Levi J. Performing risk stratification for COVID-19 when individual level data is not available – the experience of a large healthcare organization. medRxiv. 2020. Doi: 10.1101/2020.04.23.20076976
22. Xie J, Hungerford D, Chen H, et al. Development and external validation of a prognostic multivariable model on admission for hospitalized patients with COVID-19. medRxiv. 2020. Doi: 10.1101/2020.03.28.20045997
23. Coronavirus. La situazione/desktop. Available online: https://mappe.protezionecivile.gov.it/it/mappe-emergenze/mappe-coronavirus/situazione-desktop
24. Gentleman RC, Lawless JF, Lindsey JC, Yan P. Multistate Markov models for analysing incomplete disease history data with illustrations for HIV disease. *Stat Med*. 1994;13(8):805-21. Doi: 10.1002/sim.4780130803
25. Andersen PK. Multistate models in survival analysis: a study of nephropathy and mortality in diabetes. *Stat Med*. 1988;7(6):611-70. Doi: 10.1002/sim.4780070605
26. Satten GA, Longini Jr IM. Markov chains with measurement error: Estimating the 'true' course of a marker of the progression of human immunodeficiency virus disease. *J R Stat Soc: Series C (Applied Statistics)*. 1996;45(3):275-295.
27. Jackson CH, Sharples LD, Thompson SG, Duffy SW, Couto E. Multistate Markov models for disease progression with classification error. *J R Stat Soc: Series D (The Statistician)*. 2003;52(2):193-209.
28. Cox DR and Mille HD. The theory of stochastic processes. Routledge, 2017.
29. Kalbfleisch JD, Lawless JF. The analysis of panel data under a Markov assumption. *JASA*. 1985;80(392):863-871.
30. Kay R. A Markov model for analysing cancer markers and disease states in survival studies. *Biometrics*. 1986;42:855-865.
31. MSMBuilder. Available online: http://msmbuilder.org/3.8.0/
32. SciPy. Available online: https://docs.scipy.org/doc/scipy-1.5.4/reference/
33. Statsmodels. Available onlne: https://www.statsmodels.org/devel/release/version0.12.1.html
34. Rabiolo A, Alladio E, Morales E, et al. Forecasting the COVID-19 epidemic by integrating symptom search behavior into predictive models: Infoveillance

study. *J Med Internet Res*. 2021;23(8):e28876. Doi: 10.2196/28876

35. Yu C-S, Chang S-S, Chang T-H, et al. A COVID-19 pandemic artificial intelligence–based system with deep learning forecasting and automatic statistical data acquisition: development and implementation study. *J Med Internet Res*. 2021;23(5):e27806. Doi: 10.2196/27806
36. Panovska-Griffiths J, Kerr CK, Stuart RM, et al. Determining the optimal strategy for reopening schools, the impact of test and trace interventions, and the risk of occurrence of a second COVID-19 epidemic wave in the UK: a modelling study. *Lancet Child Adolesc Health*. 2020;4(11):817-827. Doi: 10.1016/S2352-4642(20)30250-9
37. Coccia M. Preparedness of countries to face COVID-19 pandemic crisis: strategic positioning and factors supporting effective strategies of prevention of pandemic threats. *Environ Res*. 2022;203:111678.
38. De Nadai M, Roomp K, Lepri B, Oliver N. The impact of control and mitigation strategies during the second wave of coronavirus infections in Spain and Italy. Sci Rep. 2022;12(1):1073.
39. Bleier BS, Ramanathan M, Lane AP. COVID-19 Vaccines May Not Prevent Nasal SARS-CoV-2 Infection and Asymptomatic Transmission. *Otolaryngol Head Neck Surg*. 2021;164(2):305-307. Doi: 10.1177/0194599820982633
40. Planas D, Veyer D, Baidaliuk, et al. Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. 2021;596(7871):276-280. Doi: 10.1038/s41586-021-03777-9.
41. Palladino R, Bollon J, Ragazzoni L, Barone-Adesi F. Excess Deaths and Hospital Admissions for COVID-19 Due to a Late Implementation of the Lockdown in Italy. *Int J Environ Res Public Health*. 2020;17(16): 5644.
42. Caristia S, Ferranti M, Skrami E, et al. Effect of national and local lockdowns on the control of COVID-19 pandemic: a rapid review. *Epidemiol Prev*. 2020;44(5-6 Suppl 2):60-68.
43. Bontempi E. The europe second wave of COVID-19 infection and the Italy 'strange' situation. *Environ Res*. 2021;193vol. 193:110476. Doi: 10.1016/j.envres.2020.110476.
44. Qian, Ying, et al. "Investigating the effectiveness of re-opening policies before vaccination during a pandemic: SD modelling research based on COVID-19 in Wuhan." *BMC Public Health* 21.1 (2021): 1-18.