Australian & New Zealand Journal of Statistics

Aust. N. Z. J. Stat. **66**(1), 2024, 1–20 doi: 10.1111/anzs.12407

Latent heterogeneity in COVID-19 hospitalisations: a cluster-weighted approach to analyse mortality

Paolo Berta[1,*], Salvatore Ingrassia[2], Giorgio Vittadini[1] and Daniele Spinelli[1]

University of Milano-Bicocca and University of Catania

Summary

The COVID-19 pandemic caused an unprecedented excess mortality. Since 2020, many studies have focussed on the characteristics of COVID-19 patients who did not survive. From the statistical point of view, what seems to dominate is the large heterogeneity of the populations affected by COVID-19 and the extreme difficulty in identifying subpopulations who died affected by a plurality of contemporary characteristics. In this paper, we propose an extremely flexible approach based on a cluster-weighted model, which allows us to identify latent groups of patients sharing similar characteristics at the moment of hospitalisation as well as a similar mortality. We focus on one of the hardest hit areas in Italy and study the heterogeneity in the population of patients affected by COVID-19 using administrative data on hospitalisations in the first wave of the pandemic. Results highlighted that a model-based clustering approach is essential to understand the complexity of the COVID-19 patients treated by hospitals and who die during hospitalisation.

Key words: cluster-weighted models; comorbidities; COVID-19; hospitalisations; mortality.

1. Introduction

COVID-19 appeared first in China in December 2019 and rapidly spread in all countries. The COVID-19 pandemic had a strong impact on our lives and caused an unprecedented excess mortality. In the last years, the clinical debate on this disease has been intense given the great variability of the effects of COVID-19. In many cases, the disease was easily overcome with treatment without the need for hospitalisation, in other cases, hospitalisation was needed and the use of intensive care and extra-corporeal membrane oxygenation was often required. In many cases, COVID-19 also led to death.

Several studies have focussed on the characteristics underlying COVID-19 mortality. Contributing causes of death by COVID-19 have been investigated using, from the statistical point of view, different approaches to highlight the effect on mortality of age, gender, comorbidities, time of onset of illness and other characteristics. In this literature, it soon became evident that age is the main predictor of mortality (Levin *et al.* 2020), whereas findings on gender seem insufficient to conclude a higher prevalence of mortality risk in men compared to women (Dehingia & Raj 2021). In addition, a large body of literature

*Author to whom correspondence should be addressed.
[1]Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Via Bicocca degli Arcimboldi 8, Milan 20126, Italy. e-mail: paolo.berta@unimib.it
[2]Department of Economics and Business, University of Catania, Corso Italia 55, Catania 95129, Italy.

focussed on comorbidities (diabetes, obesity, renal failure and cancer) with evidence that some diseases expose patients to an increased risk of death (Dombrowski & Karounos 2013; Cheng *et al.* 2020; Huang, Lim & Pranata 2020; Palaiodimos *et al.* 2020; Ciardullo *et al.* 2021; Zhang *et al.* 2021). A remarkable issue on these studies is that often they concern a single clinical condition.

What seems to dominate from a statistical point of view is the large heterogeneity of the populations affected by COVID-19 and the extreme difficulty in identifying subpopulations who died based on a plurality of contemporary and concomitant characteristics, when classical statistical approaches, such as generalised linear models, are adopted. The most common empirical strategies look at the population affected by COVID-19 as a whole, without considering the opportunity to detect latent groups among that. To the contrary, we are aware that COVID-19 mortality should be studied adopting statistical models that account for the heterogeneity in the response distribution by splitting the population into a finite number of relatively homogeneous clusters.

To this end, we propose an approach based on a cluster-weighted model (CWM), also referred to in the literature as a mixture of regressions models with random covariates (Gershenfeld 1997; Ingrassia, Minotti & Vittadini 2012; Ingrassia *et al.* 2015). This class of models is extremely flexible and leads to identify unobservable (latent) groups of patients sharing similar characteristics at the moment of hospitalisation as well as similar risk of death. In this way, the CWM approach allows us to consider, simultaneously, comorbidities and demographic characteristics of patients, but, most important, to discover latent groups of patients without any superimposed structure. This approach, which is based on pre-existing observable risk factors, could also be adopted to provide a decisional support tool for healthcare managers to predict severe consequences for patients at the moment of hospital admission.

This paper focusses on one of the hardest hit areas in Italy at the beginning of the COVID-19 pandemic. We study the heterogeneity in the population of hospital patients affected by COVID-19 using administrative data on 2617 COVID-19 admissions occurred at the 'Spedali Civili', a public hospital with three facilities located in Brescia, (Lombardy, Italy) in the period from 21 January to 26 June 2020. Among these patients admitted during the so-called first wave, approximately 22% died before discharge.

Results allow us to disentangle different effects of COVID-19 by latent groups of patients, showing the main clinical conditions related to the worst outcome of this disease. Without any claim of a medical definition, our results shows that we can identify subgroups of patients who probably did not die because an immediate consequence of pre-existing health conditions, and others who died because of some concurrent diseases including COVID-19.

This paper proceeds as follows: in Section 2, we summarise some important results of clinical literature regarding COVID-19 mortality. In Section 3, we describe our data. Section 4 introduces the CWM and the empirical strategy adopted in this study. Section 5 summarises the empirical results, and a final discussion in Section 6 concludes.

2. Background literature

The Italian Statistical Institute (Istat 2020) provided a description of the effect of COVID-19 on mortality in Italy. The proportion of cases in which COVID-19 is the main

cause of death varies by age, ranging from 82% in people under 50 years to 92% in those aged 60–69 years. The same report also highlights heterogeneity in coexisting risk factors: the most frequent comorbidities associated with COVID-19 were hypertensive heart disease (18%), diabetes mellitus (16%), ischaemic heart diseases (13%) and neoplasms (12%). Chronic lower respiratory diseases, dementia and Alzheimer's disease, and obesity were also reported but with lower frequencies. However, 28.2% of the cases had no mention of other causes contributing to death besides COVID-19.

In another study, the Italian National Institute of Health (Grippo *et al.* 2021) described patterns of concomitant pathologies in COVID-19 patients who died. Comorbidities involved in mortality changed over time. Indeed, in February–April 2020, hypertensive heart disease was mentioned as a comorbidity in 18.5% of death certificates, followed by diabetes (15.9% of cases), ischaemic heart disease (13.1%) and neoplasms (12.1%), confirming what was noted by the Italian Statistical Institute. Moving to May–September, the most frequent comorbidities were neoplasms (17.3% of cases), hypertensive heart disease (14.9%), diabetes (14.8%), and dementia/Alzheimer's disease (11.9%). The age of patients dying from COVID-19 and their disease burden increased in the May-September 2020 period. A more serious disease burden was observed in this period, with a significantly higher frequency of chronic pathologies. In this phase of the pandemic, new protocols were defined, leading to a more accurate diagnosis and better outcomes. All these factors may have improved survival in COVID-19 patients, leading to a shift in mortality to older, more vulnerable and complex patients.

It is also worth citing an international study (Bastard *et al.* 2021) from 38 countries that collected plasma or serum samples from 3595 patients with proven critical COVID-19, 623 patients with severe COVID-19, 1639 asymptomatic or paucisymptomatic individuals with proven COVID-19 and 34,159 healthy controls. In this study, Bastard et al. (2021) detected the presence of neutralizing autoantibodies to type I interferon in plasma samples and observed that the incidence increased with age in the control cohort and sharply after the age of 70. These findings indicate that autoantibodies targeting type I interferons represent a type of immunodeficiency that contributes to about 20% of all COVID-19 fatalities.

A French study (deRoquetaillade *et al.* 2021) reported that among SARS-CoV-2 positive patients who died in the ICU, multiple organ dysfunction syndrome was the leading cause of death (37%), followed by secondary infection-related multiple organ dysfunction syndrome (26%), refractory hypoxaemia/pulmonary fibrosis (19%) and fatal ischaemic events (13%). Another study (Lundberg & Zeberg 2021) analysed how European countries that coped with previous epidemics would predict COVID-19 mortality even before the epidemic reached Europe and found that the inter-country variability in death rates during the winter influenza seasons of 2015–2019 correlated to excess mortality in 2020 during the COVID-19 outbreak. Since factors like age, population density, latitude, gross national product, governmental health expenditure, number of intensive care beds, degree of urbanisation or rates of influenza vaccination did not correlate, the authors hypothesised the existence of country-specific susceptibility. In China, a multicentre study (Gao *et al.* 2021) identified some risk factors connected with COVID-19 in-hospital mortality by means of Cox regression and survival curve analysis. These factors were age, comorbidity, Sequential Organ Failure Assessment (SOFA) > 3, Acute Physiology and Chronic Health Evaluation II (APACHE II) > 7, lymphopenia (< 800 ml), C-reactive protein > 52 mg/L, IL-6 > 120 pg/mL and PaO2/FiO2 < 200 mmHg.

Table 1. Most frequent comorbidities in Premier Healthcare Database Special COVID-19 Release (PHD-SR), March 2020–March 2021 (from Kompaniyets *et al.* 2021).

Age	Underlying medical condition (CCSR Category)	Number of cases (%)
18–39 ($n = 59,697$)	Obesity	22,055 (36.9)
	Essential hypertension	9964 (16.7)
	Anxiety and fear-related disorders	9031 (15.1)
	Asthma	8524 (14.3)
	Diabetes with complication	7737 (13.0)
	Tobacco-related disorders	7240 (12.1)
	Depressive disorders	5980 (10.0)
	Diabetes without complication	2911 (4.9)
40–64 ($n = 195,897$)	Essential hypertension	98,498 (50.3)
	Obesity	82,782 (42.3)
	Disorders of lipid metabolism	79,899 (40.8)
	Diabetes with complication	62,980 (32.1)
	Oesophageal disorders	42,121 (21.5)
	Anxiety and fear-related disorders	36,978 (18.9)
	Chronic kidney disease	31,911 (16.3)
	Sleep–wake disorders	31,499 (16.1)
≥ 65 ($n = 285,073$)	Disorders of lipid metabolism	182,267 (63.9)
	Essential hypertension	164,129 (57.6)
	Coronary atherosclerosis and other heart disease	103,987 (36.5)
	Diabetes with complication	101,010 (35.4)
	Chronic kidney disease	97,802 (34.3)
	Oesophageal disorders	86,699 (30.4)
	Obesity	73,316 (25.7)
	Neurocognitive disorders	71,741 (25.2)

The Centers for Disease Control and Prevention conducted one of the widest surveys about the pandemic with data extracted from 800 US hospitals (Kompaniyets *et al.* 2021). In this article, the authors provided descriptive statistics for different subpopulations (see Tables 1 and 2) among 540,667 COVID-19 hospitalised US patients from March 2020 through March 2021. We use this data as a benchmark with the purpose of highlighting the heterogeneity in COVID-19 admissions.

Table 1 presents the different comorbidities among adults hospitalised with COVID-19 stratified by age groups. Obesity is the main comorbidity in patients aged 18–39 and 40–64 years, while it is the third most common comorbidity for patients over 65 years. Anxiety and fear-related disorders are the second-highest comorbidities for those over 65 years of age and the third highest for the other two subpopulations. Complicated diabetes is the second-most common comorbidity for patients 40–64 years old and the fourth-most common for those over 65 years but does not appear for the age group 18–39 years. Chronic kidney disease is ranked fifth for the 40–64 and over 65 age groups but does not appear for the 18–39 age group. Chronic obstructive pulmonary disease and bronchiectasis is listed sixth for the over-65-year subpopulation and fifth for the 40–64 age group but does not appear for the 18–39 age group. Aplastic anaemia is ranked fifth for the 40–64 years subpopulation but does not appear to constitute a major risk of comorbidity for the other two subpopulations. Coronary atherosclerosis and other heart diseases are listed seventh in the 40–64 years and over 65 years subpopulations but do not appear in the 18–39 years subpopulation. It follows that the population of COVID-19 hospitalised is not homogeneous

Table 2. Characteristics of COVID-19 Admissions (Premier Healthcare Database Special COVID-19 Release (PHD-SR), March 2020-March 2021 (from Kompaniyets *et al.* 2021).

Variable	Value	Full sample	Died	Proportion
Age	18–39	59,697	1299	2.2%
	40–49	51,591	2710	5.3%
	50–64	144,306	14,867	10.3%
	65–74	121,832	21,421	17.6%
	75–84	103,012	23,308	22.6%
	≥85	60,229	16,569	27.5%
Sex	Female	261,078	32,939	12.6%
	Male	279,317	47,211	16.9%
	Unknown	272	24	8.8%
No. of conditions	0	27,375	740	2.7%
	1	39,776	2087	5.2%
	2–5	212,429	25,893	12.2%
	6–10	167,706	31,310	18.7%
	>10	93,381	20,144	21.6%

but can be divided into heterogeneous subpopulations. Table 2 shows that the mortality dramatically changed among subpopulations of hospitalised patients belonging to different age groups, as expected, determining a first great heterogeneity between subpopulations of COVID-19 patients. In the same table, the comparison by gender shows that mortality is higher among males than females. Finally, in Table 2 only 2.7% of patients without any other existing clinical conditions died, whereas the mortality rate dramatically increased as the number of comorbidities increased.

In summary, to study COVID-19, it is worth exploring and analysing the heterogeneity underlying COVID-19. The scientific literature suggests that COVID-19 affects heterogeneous subpopulations, differing for age, comorbidities, the ability of hospital care to cope with changes over time, different healthcare systems and geographical situations. However, there may be unobservable sources of heterogeneity that result in latent groups. We address this issue in COVID-19 patients exploiting a suitable statistical approach for detecting latent COVID-19 hospitalised and deceased subpopulations.

3. Data: An Italian case study

The first Italian case of COVID-19 was diagnosed on 21 February 2020, in Codogno, a small municipality in the province of Lodi, when a 38-year-old healthy man was admitted to the public hospital of Codogno with mild pneumonia resistant to therapy (Cereda *et al.* 2020). At the beginning, the situation seemed to be limited to Codogno and some neighboring municipalities, but it quickly became clear that the spread of the virus concerned the whole Lombardy. In fact, on 8 March 2020, all of Lombardy was locked down into red zones, and the whole country was locked down in a national red zone a few days later, starting from 11 March 2020 (Angelici *et al.* 2023). Within Lombardy, the municipality of Brescia and the surrounding province were the earliest centres of the COVID-19 outbreak along with Bergamo and Lodi.

The impact on the healthcare systems is one of the most severe consequences of the COVID-19 pandemic. In Lombardy, where the municipality of Brescia is located, the

healthcare system reacted to this pandemic mainly in three ways: first, planned admissions were stopped and a large amount of beds capacity was dedicated to COVID-19 patients. Second, the number of ICU beds grew in three weeks from 800 to 1500. Last, in some cases, the emergency departments arranged two pathways: one dedicated to the admissions for suspected COVID-19 patients and a second one for non-COVID-19 patients. In particular, to cope with the huge number of COVID-9 patients asking for hospitalisations, the Spedali Civili—the main hospital located in Brescia—was transformed into a COVID-19 hospital hub (Casiraghi *et al.* 2020). This transformation was expressed in a systematic modification at the structural and organisational levels. First, the emergency department was structurally modified, adopting a dual-track system of admissions. For the suspected COVID-19 admissions, a specific triage was implemented in a new external emergency tent. Furthermore, at the organisational level, the existing staff were primarily involved in handling the emergency. Physicians, nurses and sanitary workers were involved and received specific training on COVID-19 management (Rossi *et al.* 2021).

In this framework, the Spedali Civili collected a large amount of information and provided the administrative hospital discharge charts analysed in this paper. The study is designed considering the 2617 COVID-19 hospitalisations occurred in the period from 21 January to 26 June 2020 in three hospitals located in Brescia. The data include several patient characteristics: (a) demographic information (age, gender); (b) information on hospitalisations, such as length of stay, special-care unit admission and in-hospital mortality; (c) up to six co-diagnosis codes and procedures defined according to the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). An important feature that can be obtained combining the six co-diagnosis relates to the patients' health status, and it approximates the level of risk that affects each patient due to their clinical conditions. This variable (*Elix*) is based on the count of comorbidities identified adopting the Elixhauser algorithm (Elixhauser *et al.* 1998). The set of comorbidities composing the Elixhauser's index is the following: congestive heart failure, cardiac arrhythmias, valvular disease, pulmonary circulation disorders, peripheral vascular disorders, hypertension, paralysis, other neurological disorders, chronic pulmonary disease, diabetes, hypothyroidism, renal failure, liver disease, peptic ulcer disease, excluding bleeding, AIDS/HIV, lymphoma, metastatic cancer, solid tumour without metastasis, rheumatoid arthritis/collagen vascular diseases, coagulopathy, obesity, weight loss, fluid and electrolyte disorders, blood loss anaemia, deficiency anaemia, alcohol abuse, drug abuse, psychoses and depression. *Elix* variable is coded as a categorical with three levels: no comorbidities, one comorbidity and more than one comorbidity.

We also consider the week of admission in the hospital (*Week*), which is measured as the number of weeks that occurred from the beginning of the year. We include this covariate in our analysis to capture the degree of stress affecting the health system in coping with the COVID-19 outbreak, which can be related to a higher risk of in-hospital mortality.

3.1. Descriptive statistics

Tables 3–7 provide summary statistics for our sample. Table 3 highlights the role of age on COVID-19, confirming that mortality increases with age. In our context of analysis, mortality is higher in male than in female (70.3% vs. 29.7%, Table 4). To this

Table 3. Brescia COVID-19 data: number of cases and percentage by age.

Age	Number of cases (%)	Number of deaths (%)
<18	35 (1.3)	0 (0.0)
18–39	125 (4.8)	0 (0.0)
40–49	188 (7.2)	7 (1.3)
50–64	702 (26.8)	46 (8.4)
65–74	678 (25.9)	146 (26.7)
75–84	668 (25.5)	251 (46.0)
>84	221 (8.4)	96 (17.6)
Total	2617 (100.0)	546 (100.0)

Table 4. Brescia COVID-19 data: number of cases and percentage by sex.

Sex	Number of cases (%)	Number of deaths (%)
Male	1687 (64.5)	384 (70.3)
Female	930 (35.5)	162 (29.7)
Total	2617 (100.0)	546 (100.0)

Table 5. Brescia COVID-19 data: number of cases and percentage by admission week.

Admission week	Number of cases (%)	Number of deaths (%)
≤ 9	191 (7.3)	57 (10.5)
10	392 (15.0)	88 (16.1)
11	513 (19.6)	126 (23.1)
12	517 (19.8)	126 (23.1)
13	367 (14.0)	60 (11.0)
14	221 (8.5)	35 (6.4)
15	135 (5.2)	19 (3.5)
16	106 (4.1)	13 (2.4)
≥17	175 (6.7)	22 (4.0)
Total	2617 (100.0)	546 (100.0)

Table 6. Brescia COVID-19 data: number of cases and percentage by number of comorbidities.

	Number of cases (%)			
# of Elixhauser's comorbidities	Number of cases		Number of deaths	
0	1897	(72.5)	376	(68.9)
1	552	(21.1)	129	(23.6)
2	138	(5.3)	33	(6.0)
3	27	(1.0)	7	(1.3)
4	3	(0.1)	1	(0.2)

Table 7. Brescia: comorbidities by age group.

Age	Underlying medical condition	Number of cases (%)
18–39	Deficiency anaemia	3 (1.9)
	Pulmonary circulation disorders	3 (1.9)
	Chronic pulmonary disease	3 (1.9)
	Other neurological disorders	2 (1.3)
	Obesity	2 (1.3)
	Renal failure	2 (1.3)
	Liver disease	2 (1.3)
	Diabetes, uncomplicated	2 (1.3)
40–64	Pulmonary circulation disorders	48 (5.4)
	Diabetes, uncomplicated	37 (4.2)
	Renal failure	36 (4.0)
	Hypertension, uncomplicated	23 (2.6)
	Cardiac arrhythmias	20 (2.2)
	Liver disease	11 (1.2)
	Other neurological disorders	10 (1.1)
	Solid tumour without metastasis	10 (1.1)
≥65	Diabetes, uncomplicated	80 (5.1)
	Cardiac arrhythmias	75 (4.8)
	Hypertension, uncomplicated	67 (4.3)
	Renal failure	64 (4.1)
	Congestive heart failure	63 (4.0)
	Pulmonary circulation disorders	60 (3.8)
	Chronic pulmonary disease	26 (1.7)
	Other neurological disorders	25 (1.6)

matter, Elgendy et al. (Elgendy & Pepine 2020) suggested that, along with behaviours more frequently associated with the male sex (i.e. smoking habit), women could be protected by oestrogen receptors or by a stronger immune response to viral infections (Klein & Flanagan 2016).

Table 5 outlines the evolution of the pandemic, showing that the peak and median admission period of our patients was in the 12th week of 2020 (from 16 March to 22 March 2020): this is in line with the first wave of the COVID-19 pandemic in Italy. The first admission was observed on 21 January 2020. In Table 6, we observe that one or more comorbidities were present at admission in 720 (27.5%) patients, accounting for the 31% of deceased patients. To detail the role of comorbidities, we included Table 7. Although age groups exhibit some common patterns (pulmonary circulation disorders, renal failure, diabetes and hypertension), comorbidities vary between age ranges. For instance, while in the youngest group the most frequent comorbidities are equally distributed, diabetes has the highest percentage in the oldest group, whereas pulmonary circulation disorders are common in the 40–64 range.

A comparison between Tables 1 and 2 and Tables 3–6 allows us to appreciate the presence of latent heterogeneity related to COVID-19 patients. For instance, mortality in patients hospitalised in Brescia was higher than those observed in the United States (20.9% vs. 14.9%). For the group of patients aged 75–84 years, the mortality in our sample is double than what was observed in Kompaniyets *et al.* (2021). In addition, the proportion of males among hospitalised and deceased patients was much higher in Brescia than in the US sample (64.5% vs. 51.7% and 70.3% vs. 58.9%).

We acknowledge that direct comparison between a single Italian hospital and United States' data should be done cautiously, as there are many sources of unobserved heterogeneity that may affect the results. Such sources may be related to characteristics of the population, of the healthcare system and of the virus variants. However, such a comparison highlights that, even though age, gender, comorbidities and period of admission explain some heterogeneity in COVID-19 mortality, there is a large unobserved heterogeneity in the populations of COVID-19 patients. Thus, we explore latent heterogeneity in COVID-19 hospitalisations adopting a multivariate approach that allows us to characterise latent groups using the patient characteristics outlined in this section.

4. Cluster-weighted models

Data heterogeneity has been modelled here according to the CWM (Gershenfeld 1997; Ingrassia, Minotti & Vittadini 2012). Here we first introduce a quite general model, called the generalised CWM (Ingrassia *et al.* 2015); afterwards, we specialise the model for the analysis of the COVID-19 data.

The CWM belongs to the class of the mixture of regression models. In literature, this model has been also referred to as *Mixture Model with Random Covariates*. Let $(\mathbf{X}^\top, Y)^\top$ be a pair of a random vector $\mathbf{X}$ and a random variable Y defined on Ω with joint probability $p(\mathbf{x}, y)$, where $\mathbf{X}$ is the d-dimensional input vector with values in some space $\mathcal{X} \subseteq \mathbb{R}^d$ and Y is a response variable having values in $\mathcal{Y} \subseteq \mathbb{R}$. While the mixture of regressions models the conditional probability density $p(y|\mathbf{x})$, the CWM models the joint probability density $p(\mathbf{x}, y)$.

Assume that Ω can be partitioned into G disjoint groups, say $\Omega_1, \ldots, \Omega_G$, that is $\Omega = \Omega_1 \cup \cdots \cup \Omega_G$. The general formulation of a CWM is

$$p(\mathbf{x}, y) = \sum_{g=1}^{G} \pi_g p(y|\mathbf{x}; \Omega_g) p(\mathbf{x}; \Omega_g), \tag{1}$$

where $\pi_g = p(\Omega_g)$ is the mixing weight of group Ω_g, $p(\mathbf{x}|\Omega_g)$ is the probability density of $\mathbf{x}$ given Ω_g and $p(y|\mathbf{x}; \Omega_g)$ is the conditional density of the response variable Y given the predictor vector $\mathbf{x}$ and the group Ω_g, $g = 1, \ldots, G$.

The posterior probability $p(\Omega_g|\mathbf{x}; y)$ that the pair $(\mathbf{x}, y)$ belongs to the gth group $(g = 1, \ldots, G)$ is given by:

$$p(\Omega_g|\mathbf{x}; y) = \frac{p(\mathbf{x}, y; \Omega_g)}{p(\mathbf{x}, y)} = \frac{p(y|\mathbf{x}; \Omega_g) p(\mathbf{x}; \Omega_g) \pi_g}{\sum_{j=1}^{G} p(y|\mathbf{x}; \Omega_j) p(\mathbf{x}; \Omega_j) \pi_j}. \tag{2}$$

CWMs have been first proposed in a context of media technology under Gaussian assumptions; subsequently, they have been investigated from a statistical point of view in a sequence of papers: Ingrassia, Minotti & Vittadini (2012) reformulated the CWM in a statistical setting and showed that it is a general and flexible family of mixture models; Ingrassia, Minotti & Punzo (2014) presented a family of twelve CWMs, nested in the linear t-CWM, for model-based clustering; Subedi *et al.* (2013) addressed the problem of applicability of the CWM in high-dimensional $\mathbf{X}$-spaces by assuming latent factors for the covariates in each mixture component. Moreover, in healthcare, the multilevel CWM has

been proposed for the hospital evaluation (Berta *et al.* 2016), and extended in (Berta & Vinciotti 2019) for binary response variables.

In this paper, we consider the *generalised linear mixed CWM* (Ingrassia *et al.* 2015; Mazza, Punzo & Ingrassia 2018), where the component conditional distributions are assumed to belong to the exponential family and the covariates are allowed to be of mixed-type.

In this framework, the vector of covariates can be written as $\mathbf{X} = \left(\mathbf{U}^\top, \mathbf{V}^\top\right)^\top$, where $\mathbf{U}$ is a p-variate vector of continuous covariates and $\mathbf{V}$ is a q-variate vector of categorical covariates, with number of levels $c_1, \ldots, c_q$, respectively, being $p + q = d$. In this case, $\mathcal{X} = \mathbb{R}^p \times \{1, \ldots, c_1\} \times \cdots \times \{1, \ldots, c_q\}$. Moreover, we assume that $\mathbf{U}$ and $\mathbf{V}$ are 'locally' independent; that is, they are independent within each mixture component. The conditional distribution of the variable $\mathbf{U}$ given Ω_g is modelled here according to a p-variate Gaussian density with mean $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$, i.e. $p\left(\mathbf{u}; \boldsymbol{\psi}_g^*\right) = \phi\left(\mathbf{u}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right)$. As for the variable $\mathbf{V}$, we assume that each categorical covariate can be represented by a binary vector $\mathbf{V}^r = \left(V^{r1}, \ldots, V^{rc_r}\right)^\top$, where $V^{rs} = 1$ if V_r is equal to the category s, with $s \in \{1, \ldots, c_r\}$, and $V^{rs} = 0$ otherwise.

Furthermore, we assume that the q categorical covariates are independent given the mixture component. Then, we have

$$p(\mathbf{v}; \boldsymbol{\alpha}_g) = \prod_{r=1}^{q} \prod_{s=1}^{c_r} \left(\alpha_{grs}\right)^{v^{rs}}, \quad g = 1, \ldots, G, \tag{3}$$

where $\boldsymbol{\alpha}_g = \left(\boldsymbol{\alpha}_{g1}^\top, \ldots, \boldsymbol{\alpha}_{gq}^\top\right)^\top$, with $\boldsymbol{\alpha}_{gr} = \left(\alpha_{gr1}, \ldots, \alpha_{grc_q}\right)^\top$, $\alpha_{grs} > 0$ and $\sum_{s=1}^{c_r} \alpha_{grs} = 1$, $r = 1, \ldots, q$. In particular, the density $p\left(\mathbf{v}; \boldsymbol{\alpha}_g\right)$ in (3) is given by the product of q conditionally independent multinomial distributions of parameters $\boldsymbol{\alpha}_{gr}$, $r = 1, \ldots, q$.

Based on the above assumptions, model (1) assumes the form

$$p(\mathbf{x}, y; \boldsymbol{\vartheta}) = \sum_{g=1}^{G} q(y|\mathbf{x}; \boldsymbol{\beta}_g)\ \phi(\mathbf{u}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\ p(\mathbf{v}; \boldsymbol{\alpha}_g)\ \pi_g, \tag{4}$$

where $q(y|\mathbf{x}; \boldsymbol{\beta}_g)$ denotes the conditional density of $Y|\mathbf{x}; \Omega_g$ with parameter $\boldsymbol{\beta}_g$. Model (4) is referred to as the generalised linear mixed CWM, where the prefix 'generalised linear' refers to the local relation of Y given $\mathbf{x}$, and the term 'mixed' underlines the mixed-type nature of the random covariates.

4.1. Cluster-weighted approach for modeling COVID-19 data

Our approach for modelling COVID-19 data aims at identifying latent groups in COVID-19 hospital mortality. Hence, our response variable Y is a binary indicator that is equal to one if the patient dies during the hospitalisation and zero otherwise (i.e. we assume that Y takes values in $\mathcal{Y} = \{0, 1\}$ and that $Y|\mathbf{x}, \Omega_g$ follows a Bernoulli distribution with parameters $\mu_g\left(\mathbf{x}; \boldsymbol{\beta}_g\right)$). In this case, in (4) we have

$$q(y|\mathbf{x}; \boldsymbol{\beta}_g) = \left[\mu_g(\mathbf{x}; \boldsymbol{\beta}_g)\right]^y \left[1 - \mu_g(\mathbf{x}; \boldsymbol{\beta}_g)\right]^{1-y}, \tag{5}$$

where

$$\mu_g\left(\mathbf{x};\boldsymbol{\beta}_g\right)=\frac{\exp(\beta_{0g}+\boldsymbol{\beta}_{1g}^{\top}\mathbf{x})}{1+\exp(\beta_{0g}+\boldsymbol{\beta}_{1g}^{\top}\mathbf{x})}. \tag{6}$$

Model (4), with conditional distributions (5), will be called the Bernoulli CWM (a special case of the Binomial CWM, which belongs to the family of the generalised CWM (Ingrassia *et al.* 2015).

Computational details

Model (4) has been fitted on our COVID-19 data according to the maximum likelihood approach, using an expectation maximisation (EM) algorithm (Ingrassia *et al.* 2015). In fact, CWM can be viewed as a situation of incomplete data (McLachlan & Peel 2004) and the adopted EM algorithm identifies the posterior probability that each observation belongs to each predefined latent clusters. Afterwards, patients were clustered according to the maximum posterior probability. Our EM algorithm follows an iterative process that starts using the available data (E-step) and then maximizing the expected log-likelihood (M-step). The iterative process continues until a predefined convergence criterion is met. The convergence is guaranteed when the Aitken acceleration index (Aitken 1927) is lower than a defined threshold, which is typically set to 10^{-4}. From a computational point of view, EM algorithms can be sensitive to the starting point, and several initialisation strategies can be implemented (Biernacki, Celeux & Govaert 2003; Karlis & Xekalaki 2003). In this case, we have considered five repeated runs of the k-means algorithm, which is faster and more stable than random draws (Berta & Vinciotti 2019). The k-means algorithm is adopted as a starting point for the initial allocation of the observations to one of the unobserved clusters exploiting the relationship between the dependent variable and only the continuous covariates (*Age* and *Week*).

At the end of the estimation process, the Akaike information criterion (AIC) and Bayesian information criterion (BIC) have been used to deselect models.

CWM specification

In this paper, the Bernoulli CWM is defined by three components. To begin with, the vector of covariates $\mathbf{X}$ includes both numerical and categorical variables: patient's age (*Age*), sex (*Female*), and severity proxied by comorbidities (*Elix*). These patients' characteristics are typically used in healthcare as risk-adjustment covariates; in particular, they are included to control for clinical conditions pre-existing to the hospitalisation, and they can be considered as a risk factor for in-hospital mortality. The vector $\mathbf{X}$ also includes the variable week of admission (*Week*), which is the number of weeks that occurred from the beginning of the year. We include this covariate with the aim of capturing the evolution of pandemic, as a proxy for severity of contagious in the population and the stress experienced by the healthcare system.

The numerical variables *Age* and *Week* are modelled according to a multivariate Gaussian distribution with vector of means $\boldsymbol{\mu}_g$ and variance–covariance matrix $\boldsymbol{\Sigma}_g$, considering Gaussian parsimonious models (Celeux & Govaert 1995; Punzo & Ingrassia 2016). As for the categorical variables, *Elix* is assumed to be Poisson distributed with mixture component-specific mean λ_g and *Female* is assumed to be Bernoulli distributed with probability ψ_g.

Thus in (4), we set $\mathbf{v} = (v_1, v_2)'$ where $v_1 = Female$ and $v_2 = Elix$. In summary, in (4) we have

$$\begin{aligned} \phi(\mathbf{u}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) &= \phi(Age, Week; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \\ p(v_1; \psi_g) &= \psi_g^{\text{Female}}(1-\psi_g)^{(1-\text{Female})}, \\ p(v_2; \lambda_g) &= \frac{\lambda_g^{\text{Elix}} \exp(-\lambda_g)}{\text{Elix!}}. \end{aligned} \tag{7}$$

As for the conditional distribution of $Y|\mathbf{x}; \Omega_g$, from (5) and (6), we get

$$q(y|\mathbf{x}; \boldsymbol{\beta}_g) = \left(\frac{\exp(\alpha_g + \boldsymbol{\beta}_g^\top \mathbf{x})}{1+\exp(\alpha_g + \boldsymbol{\beta}_g^\top \mathbf{x})}\right)^y \left(1 - \frac{\exp(\alpha_g + \boldsymbol{\beta}_g^\top \mathbf{x})}{1+\exp(\alpha_g + \boldsymbol{\beta}_g^\top \mathbf{x})}\right)^{(1-y)}. \tag{8}$$

We remark that the some analysis suggested including both variables *Week* and $Week^2$ to improve the model fit to the data.

Finally, here we assumed that the conditional distribution of the continuous covariates in each group is multivariate Gaussian. Anyway, once the groups provided by the model are obtained, according to the maximum posterior probability, we get truncated distributions and then the normality assumption cannot be checked. As a matter of fact, an analysis of the distributions of posterior probabilities of the units points out that in many cases, the value of the largest probability is quite close to the value of the second largest probability and this means that there is some overlap between the probability densities of the groups.

5. Empirical results

In this section, we describe the results of the application of the CWM to detect unobservable subgroups in the population of COVID-19 patients hospitalised in Brescia. Stata was used for the whole analysis, the command cwmglm (Spinelli, Ingrassia & Vittadini 2022) was used to fit the CWMs. We adopted a CWM approach defined by the number of clusters G and the parametrisation of the variance–covariance matrix $\boldsymbol{\Sigma}_g$ of the Gaussian covariates (*Age*,*Week*). We compare CWMs and finite mixture of regressions (FMR with $G = 2$ because over 2 groups the estimation process did not converge) based on $2 \leq G \leq 5$ latent clusters and a logistic regression (i.e. a CMW with $G = 1$). We remark that comparison between FMR and CWM from both theoretical and applied perspective is analysed in Ingrassia, Minotti & Vittadini (2012); in particular it is shown that FMR can be considered as a particular case of CWM under suitable constrains. Moreover, such theoretical results allow to compare model selection criteria like BIC and AIC between FMR and CWM. FMRs have also been estimated using Stata command `cwmglm`.

Each CWM is combined with the parsimonious models of the variance/covariance matrix defined by Celeux & Govaert (1995). According to Dang *et al.* (2017), such parametrisations define the volume (equal or variable), the shape (equal or variable) and the orientation (axis-aligned, equal or variable) of $\boldsymbol{\Sigma}_g$. The combinations of volume, shape and orientations lead to models labelled as EEI, VII, EEI, VEI, EVI, VVI, EEE, VVV, EVV, VEV, EVE, VEE, VVE and VVV. Therefore, up to fourteen models are originated for each value of G.

Table 8. Comparison among alternative CWM and FMR.

Model	G	AIC	BIC
FMR	2	48,537	48,666
CWM	1	45,434	45,516
CWM	2	42,536–46,445	42,670–46,573
CWM	3	43,418–45,285	43,629–45,484
CWM	4	45,196–45,229	45,506–45,544
CWM	5	44,331–45,229	44,451–44,698

Starting from the information displayed in Table 8, we selected CWM with $G = 3$ and EEV variance–covariance matrix, which has an AIC equal to 43,418 and a BIC equal to 43,629. According to both AIC and BIC, models with either $G = 2$ or $G = 3$ could be taken into account; anyway, Kadane & Lazar (2004) stated that: 'there is no particular reason to choose a single best model according to some criterion. Rather, it makes more sense to 'deselect' poor models, maintaining a subset for further considerations. Sometimes this subset might consist of a single model, but sometimes perhaps not'. Therefore, we selected the model yielding the clearest interpretation among the models that attained the best values. We also tested CWM with a number of predefined latent groups greater than 5, but the estimation process fails to converge.

The results obtained at the end of the estimation process are presented in Table 9, and they can be summarised as follows:

1 The identified mixture components are labelled as g_1, g_2 and g_3. For g_1, the prior membership probability is 6.7%. The same parameters equal to 78.9% for g_2 and to 14.4% for g_3.
2 Patients in group g_1 on average are 74 years old and have been admitted on the 18th week of 2020 (27 April–03 May). These are the latest admissions compared to the other mixture components. Moreover, according to the variance–covariance matrix of the normal covariates, age and admission week are positively correlated. Finally, we remark that these patients present the highest number of comorbidities (Elixhauser's index is equal to 0.7) and male and female patients are equally represented (50%).
3 The second group (g_2) comprises the majority of patients. Here, on average patients are 70 years old and have been admitted to the hospital in the 11th week (9 March–15 March which corresponds to the first week of lockdown in Italy). In addition, this group is characterised for a strong prevalence of male (67%) and a number of comorbidities equal to 0.345.
4 The final group (g_3) includes 14.4% of patients. This is the youngest group (on average patients are 46 years old) and the healthiest group (on average, 0.155 comorbidities). They have been admitted approximately on the 12th week of 2020 (16 March–22 March).
5 Groups g_1 and g_2 show similar mortality rates (i.e. 0.211 and 0.243, respectively) which are considerably higher than mortality rate in group $g3$ (0.039), as expected observing the age and the low number of comorbidities characterizing this group.
6 The off-diagonal elements of the covariance matrices of the Gaussian covariates in (9) have negative elements in $\mathbf{\Sigma}_2$ and $\mathbf{\Sigma}_3$. This means that the relationship between the

Table 9. Parameter Estimates according to the CWM (marginal density).

Distribution	Variable	g_1 ($\pi_1 = 6.7\%$)	g_2 ($\pi_2 = 78.9\%$)	g_3 ($\pi_3 = 14.4\%$)
Prior	Mortality	0.211	0.243	0.039
Normal (EEV)	Admission week (Mean, μ_{Week})	18.417	11.877	12.119
Normal (EEV)	Admission week (Variance $\mathbf{\Sigma}_{\text{Week,Week}}$)	3.931	4.714	4.288
Normal (EEV)	Age (Mean, μ_{Age})	73.998	70.238	46.207
Normal (EEV)	Age (Variance $\mathbf{\Sigma}_{\text{Age,Age}}$)	134.849	134.067	134.492
Normal (EEV)	Covariance Age-Admiss. Week ($\mathbf{\Sigma}_{\text{Age,Week}}$)	1.398	−10.188	−6.972
Binomial	Female (Ψ)	0.505	0.33	0.415
Poisson	# of Elixhauser's comorb. (λ)	0.703	0.345	0.155

Table 10. Regression coefficient estimates in the conditional distribution of $Y|\mathbf{x}; \Omega_g$, see $\boldsymbol{\beta}_g$ in (8).

$\boldsymbol{\beta}$	g_1	g_2	g_3
Week†	−0.693	−0.283	−0.701
Week^2 †	0.04	−0.051	0.031
# of Elixhauser's comorbidities=1	−0.734	0.128	7.484
# of Elixhauser's comorbidities ≥ 2	−0.496	−0.006	3.423
Age	0.045	0.096	0.242
Female	−0.392	−0.51	−0.589
Intercept	−1.842	−7.98	−21.035
log-likelihood at convergence		−21,673.438	
number of iterations		777	

Notes: †*Week* is mean centered.

Age and the *Week* is negative in groups g_2 and g_3. Therefore, the average age of hospitalised patients decreased as the pandemic progressed by week. Conversely, the off-diagonal element of $\mathbf{\Sigma}_1$ means that, in g_1, patients' age increased in time. The covariance matrices are as follows:

$$\mathbf{\Sigma}_1 = \begin{pmatrix} 3.931 & 1.398 \\ 1.398 & 134.849 \end{pmatrix}, \mathbf{\Sigma}_2 = \begin{pmatrix} 4.714 & -10.188 \\ -10.188 & 134.067 \end{pmatrix}, \mathbf{\Sigma}_3 = \begin{pmatrix} 4.288 & -6.972 \\ -6.972 & 134.92 \end{pmatrix}. \tag{9}$$

Thus, according to Table 9, it can be concluded in the first instance that we have three groups of COVID-19 hospitalisations: two composed of elderly patients and one of younger patients. The central group g_2 does not differ so much in mortality and age from the other more limited elderly group g_1, except for the date of admission, Elixhauser's index and gender. In contrast, young people in group g_3 are less exposed to the risk of death, compared to the other two groups: in this group, COVID-19 mortality increases with increased comorbidity, male sex and earlier admission (Table 10). As we can see, the adoption of the CWM model to describe the characteristics of in-hospital COVID-19 mortality allows us to detect latent groups, showing their different characteristics simultaneously and clearly.

A visual comparison of the three groups allows to better understand the relevance of a clustering approach in this kind of analysis to detect latent groups of patients. Figure 1 shows the distributions of the Gaussian covariates of the selected CWM (*Age* and *Week*); g_1 represented with a solid line pattern in the top and mid panels and a hollow square in the bottom panel, g_2 with dotted lines and circles and g_3 with dashed lines and triangles. The starting time period is Week 4, as the first COVID-19 case observed in our dataset was dated 21 January 2020. Observations are allocated to groups and identified by colours according to their maximum posterior probabilities. The top and central panels are two density plots obtained starting from the histogram of the distributions (by groups) of the same variables. The bottom panel jointly represents *Age* and *Week*. Figure 1 confirms that groups g_1 (solid line) and g_2 (dotted) have substantial overlapping with respect to *Age*, while mixture component g_3 (dashed) is younger and well separated. Thus, we verified that the third group relates to young patients, while the first and second groups refer to the elderly. This determines a strong difference in terms of risk of mortality.

The second interesting element which defines the differences between groups is given by the week of admission. In Figure 1 we observe that patients in group g_2 are, on average, those who were hospitalised first (11th week), similarly to patients belonging to group g_3, which are admitted 1 week later (12th week). Differently, patients belonging to group g_1 are admitted in the final period of the first wave (18th week). In addition, Figure 2 allows us to understand the role of week of admission on mortality. The vertical axis represents the predicted mortality in each cluster as a function of the week of admission. The other variables (age, gender and comorbidities) are set to the cluster average. In all three groups, mortality decreases over time as a function of the week of admission. This can be explained by the ability to care which has grown during the first wave, and this had a higher effect on groups g_1 and g_2, which was affected by a higher risk of death due to their age. An effect is also observable on young people despite their lower mortality.

This confirms that mortality is primarily related to patients' age, as it is supported by observing the results of logistic regressions in Table 10. In this table, the coefficients for age in the columns for groups g_1 and g_2 have a small magnitude compared to the same coefficient in group g_3. Overall, the positive coefficients confirm that an increase in age lead to an increase in mortality. In addition we observe that this effect is stronger in group g_3, meaning that in this group of younger patients, despite the lower mortality rate, there is a higher heterogeneity of risk of mortality related to the age.

Proceeding with Table 10, differences between groups are observed also by gender and comorbidities. In each group, females have a lower risk of mortality regardless any other patients characteristics. In addition, the correlation between comorbidities and mortality is non-monotonic: the risk associated with a single comorbid condition is always higher than those associated with more than one comorbidity.

The comorbidities in group g_1 have negative effect on mortality, whereas in group g_2 they seem to be irrelevant on mortality. Moreover, it is worth to notice that the few comorbidities among young people (group g_3) have lethal effects. In other words, in the first group of elderly, age is the most relevant covariate explaining the risk of mortality, while within the youngest group, the comorbidities seem to be the main driver in predicting mortality.

The examination of the type of comorbidity allows us to further develop the results achieved so far. Among the different groups some common patterns emerge: for example, pulmonary circulation disorders, diabetes and cardiac arrhythmias are the most frequent

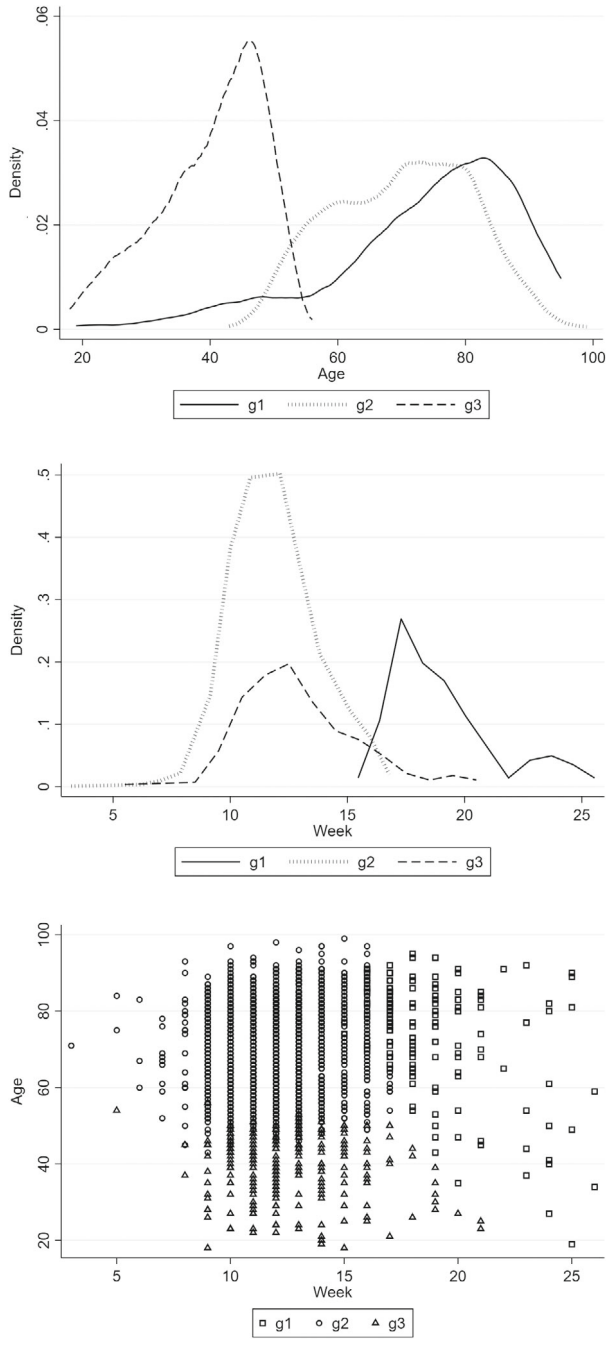

Figure 1. Distributions of the variables *Age* and *Admission week* in the groups: univariate density functions and scatter plot. In the top and mid plot groups are distinguished by line pattern: Group 1 (g_1) as a solid line, Group 2 (g_2) with a dotted pattern and Group 3 (g_3) with dashed pattern. In the bottom plot g_1 is characterized by squared markers, g_2 and g_3 by circular and dashed markers, respectively.

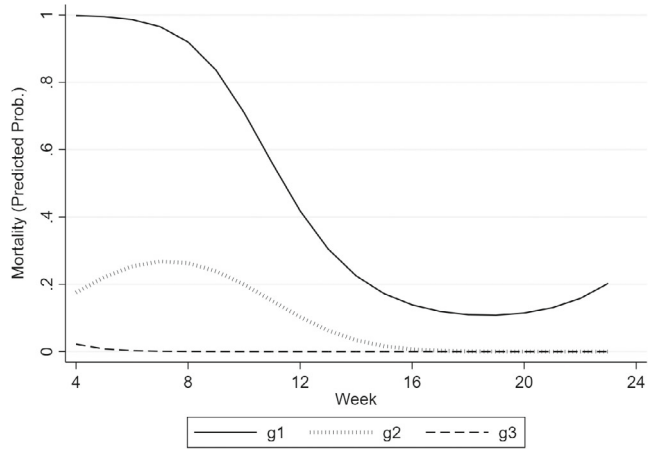

Figure 2. Predicted mortality versus time. Group 1 (g_1) as a solid line, Group 2 (g_2) with a dotted pattern and Group 3 (g_3) with dashed pattern.

diagnoses related to hospital admission. However some differences emerge when we analyse the results in each group. We comment on the second group, g_2, the one composed by the majority of patients. We note that the most relevant comorbidities are diabetes uncomplicated (0.050), pulmonary circulation disorders (0.045), renal failure (0.042), hypertension uncomplicated (0.039), cardiac arrhythmias (0.038) and congestive heart failure (0.027). Group g_1 contains the smallest proportion of patients and the most frequent comorbidities are: cardiac arrhythmias (0.094), congestive heart failure (0.075), diabetes, uncomplicated (0.063), pulmonary circulation disorders (0.059), other neurological disorders (0.049), hypertension uncomplicated (0.045), chronic pulmonary disease (0.042) and renal failure (0.037). As we can see, the typology does not differ so much from those observed in group g_1, where the proportion of people affected by comorbidity is much larger. Finally, the young patients of group g_3, where the comorbidities play an important role on mortality, are mainly affected by pulmonary circulation disorders (0.027), renal failure (0.026) and diabetes uncomplicated (0.013); these comorbidities are also the most observed comorbidities in group g_2 (Table 11).

6. Discussion

In this article, we focussed on COVID-19 patients and their in-hospital mortality. We exploit a model-based clustering approach in order to assess heterogeneity in the data and to detect latent clusters of patients. In fact, from our descriptive analysis and the comparison with the Centers for Disease Control and Prevention data (US), we observe that a standard statistical approach is not suitable to distinguish the plausible existence of latent groups of patients.

We assume that heterogeneity within COVID-19 subpopulations varies according to different situations. That is, the type of heterogeneity in different contexts and conditions is not predictable. Such heterogeneity includes also different earlier clinical conditions in

Table 11. Overview on clinical conditions in each mixture component.

# of Elixhauser's comorbidities	g_1	g_2	g_3
0	0.456	0.720	0.859
1	0.385	0.214	0.127
≥ 2	0.159	0.066	0.014

Comp.	Comorbidity	Proportion
g_1	Cardiac arrhythmias	0.094
	Congestive heart failure	0.075
	Diabetes, uncomplicated	0.063
	Pulmonary circulation disorders	0.059
	Other neurological disorders	0.049
	Hypertension, uncomplicated	0.045
	Chronic pulmonary disease	0.042
	Renal failure	0.037
g_2	Diabetes, uncomplicated	0.050
	Pulmonary circulation disorders	0.045
	Renal failure	0.042
	Hypertension, uncomplicated	0.039
	Cardiac arrhythmias	0.038
	Congestive heart failure	0.027
	Solid Tumour without metastasis	0.014
	Other neurological disorders	0.012
g_3	Pulmonary circulation disorders	0.027
	Renal failure	0.026
	Diabetes, uncomplicated	0.013
	Cardiac arrhythmias	0.009
	Deficiency anaemia	0.009
	Liver disease	0.008
	Obesity	0.008
	Other neurological disorders	0.006

patients affected by COVID-19 and often it is not always clear if COVID-19 was the main cause of death or a more or less serious concurrent cause.

The empirical analysis focusses on the hospital discharge record of the Spedali Civili in Brescia, one of the earliest and most hit location in Western countries. Our data include information about patients' age and gender, which are recognised for being two predictors of COVID-19 mortality. In addition, we have also the opportunity to analyse the week of hospital admission, a variable that provides another source of heterogeneity related with the evolution of the pandemic and the level of stress on the healthcare system. An additional source of heterogeneity comes from the type of comorbidities associated with mortality in Brescia, classified according to the hospital discharge records.

Based on our cluster-weighted analysis, three latent groups of patients are detected. The main drivers characterizing these groups are: patients' age, their comorbidities and week of admission. The three groups detected by our CWM are exposed to different risk of death and this support our empirical approach, instead of typical statistical approaches which do not consider latent heterogeneity. The results largely simplify group description and appear more realistic and intelligible.

To the best of our knowledge, this is the first study attempting to detect unobservable groups of patients while considering group characteristics respect to the in-hospital mortality.

Due to data limitation, we cannot access any information following the patients' discharge. In this case, for example, we would have the opportunity to consider not only the in-hospital mortality but also the post-discharge mortality which is a widely adopted outcome in the healthcare literature. In addition, the data do not provide information on behaviours that could affect the likelihood of dying, such as smoking habits.

Data availability statement

The data that support the findings of this study are available from Spedali Civili of Brescia and can be obtained only under a research agreement with the healthcare administration of the hospital and cannot be shared publicly. The authors are willing to fully co-operate providing any assistance and information on how the administrative data we used can be obtained with the purpose to replicate our analysis.

REFERENCES

AITKEN, A.C. (1927). XXV.—On Bernoulli's numerical solution of algebraic equations. *Proceedings of the Royal Society of Edinburgh*, **46**, 289–305.

ANGELICI, M., BERTA, P., COSTA-FONT, J. & TURATI, G. (2023). Divided we survive? Multilevel governance during the COVID-19 pandemic in Italy and Spain. *Publius: The Journal of Federalism*, **53**, 227–250.

BASTARD, P., GERVAIS, A., LE VOYER, T., *et al.* (2021). Autoantibodies neutralizing type I IFNS are present in 4% of uninfected individuals over 70 years old and account for 20% of COVID-19 deaths. *Science Immunology*, **6**, eabl4340.

BERTA, P., INGRASSIA, S., PUNZO, A. & VITTADINI, G. (2016). Multilevel cluster-weighted models for the evaluation of hospitals. *Metron*, **74**, 275–292.

BERTA, P. & VINCIOTTI, V. (2019). Multilevel logistic cluster-weighted model for outcome evaluation in health care. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, **12**, 434–443.

BIERNACKI, C., CELEUX, G. & GOVAERT, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, **41**, 561–575.

CASIRAGHI, A., DOMENICUCCI, M., CATTANEO, S., *et al.* (2020). Operational strategies of a trauma hub in early coronavirus disease 2019 pandemic. *International Orthopaedics*, **44**, 1511–1518.

CELEUX, G. & GOVAERT, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, **28**, 781–793.

CEREDA, D., TIRANI, M., ROVIDA, F., *et al.* (2020). The early phase of the COVID-19 outbreak in Lombardy, Italy arXiv preprint, arXiv:2003.09320.

CHENG, Y., LUO, R., WANG, K., *et al.* (2020). Kidney disease is associated with in-hospital death of patients with COVID-19. *Kidney International*, **97**, 829–838.

CIARDULLO, S., ZERBINI, F., PERRA, S., *et al.* (2021). Impact of diabetes on COVID-19-related in-hospital mortality: a retrospective study from northern Italy. *Journal of Endocrinological Investigation*, **44**, 843–850.

DANG, U.J., PUNZO, A., MCNICHOLAS, P.D., INGRASSIA, S. & BROWNE, R.P. (2017). Multivariate response and parsimony for Gaussian cluster-weighted models. *Journal of Classification*, **34**, 4–34.

DEHINGIA, N. & RAJ, A. (2021). Sex differences in COVID-19 case fatality: do we know enough? *The Lancet Global Health*, **9**, e14–e15.

DE ROQUETAILLADE, C., BREDIN, S., LASCARROU, J.B., *et al.* (2021). Timing and causes of death in severe COVID-19 patients. *Critical Care*, **25**, 224. https://doi.org/10.1186/s13054-021-03639-w

DOMBROWSKI, N.C. & KAROUNOS, D.G. (2013). Pathophysiology and management strategies for hyperglycemia for patients with acute illness during and following a hospital stay. *Metabolism*, **62**, 326–336.

ELGENDY, I.Y. & PEPINE, C.J. (2020). Why are women better protected from COVID-19: clues for men? sex and covid-19. *International Journal of Cardiology*, **315**, 105–106.

ELIXHAUSER, A., STEINER, C., HARRIS, D.R. & COFFEY, R.M. (1998). Comorbidity measures for use with administrative data. *Medical Care*, **36** (1), 8–27.

GAO, J., ZHONG, L., WU, M., *et al.* (2021). Risk factors for mortality in critically ill patients with COVID-19: a multicenter retrospective case-control study. *BMC Infectious Diseases*, **21**, 602. https://doi.org/10.1186/s12879-021-06300-7

GERSHENFELD, N. (1997). Nonlinear inference and cluster-weighted modeling. *Annals of the New York Academy of Sciences*, **808**, 18–24.

GRIPPO, F., GRANDE, E., MARASCHINI, A., *et al.* (2021). Evolution of pathology patterns in persons who died from COVID-19 in Italy: a national study based on death certificates. *Frontiers in Medicine*, 8. https://www.frontiersin.org/articles/10.3389/fmed.2021.645543/full

HUANG, I., LIM, M.A. & PRANATA, R. (2020). Diabetes mellitus is associated with increased mortality and severity of disease in COVID-19 pneumonia–a systematic review, meta-analysis, and meta-regression. *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, **14**, 395–403.

INGRASSIA, S., MINOTTI, S.C. & PUNZO, A. (2014). Model-based clustering via linear cluster-weighted models. *Computational Statistics and Data Analysis*, **71**, 159–182.

INGRASSIA, S., MINOTTI, S.C. & VITTADINI, G. (2012). Local statistical modeling via the cluster-weighted approach with elliptical distributions. *Journal of Classification*, **29**, 363–401.

INGRASSIA, S., PUNZO, A., VITTADINI, G. & MINOTTI, S.C. (2015). The generalized linear mixed cluster-weighted model. *Journal of Classification*, **32**, 85–113.

ISTAT (2020). *Impact of COVID-19 Epidemic on Mortality: Causes of Death in COVID-19 Laboratory Confirmed Cases*. Technical report. Rome: Istat. https://www.istat.it/it/files/2020/07/Report_ISS_Istat_Inglese.pdf

KADANE, J.B. & LAZAR, N.A. (2004). Methods and criteria for model selection. *Journal of the American Statistical Association*, **99**, 279–290.

KARLIS, D. & XEKALAKI, E. (2003). Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics & Data Analysis*, **41**, 577–590.

KLEIN, S.L. & FLANAGAN, K.L. (2016). Sex differences in immune responses. *Nature Reviews Immunology*, **16**, 626.

KOMPANIYETS, L., PENNINGTON, A.F., GOODMAN, A.R., BELAY, H., KO, B.J.Y., *et al.* (2021). Underlying medical conditions and severe illness among 540,667 adults hospitalized with COVID-19, March 2020–March 2021. *Preventing Chronic Disease*, **1** (18), E66. https://doi.org/10.5888/pcd18.210123

LEVIN, A.T., HANAGE, W.P., OWUSU-BOAITEY, N., COCHRAN, K.B., WALSH, S.P. & MEYEROWITZ-KATZ, G. (2020). Assessing the age specificity of infection fatality rates for COVID-19: systematic review, meta-analysis, and public policy implications. *European Journal of Epidemiology*, **35**, 1123–1138. https://doi.org/10.1007/s10654-020-00698-1

LUNDBERG, J.O. & ZEBERG, H. (2021). Longitudinal variability in mortality predicts COVID-19 deaths. *European Journal of Epidemioliogy*, **36**, 599–603. https://doi.org/10.1007/s10654-021-00777-x

MAZZA, A., PUNZO, A. & INGRASSIA, S. (2018). flexCWM: a flexible framework for cluster-weighted models. *Journal of Statistical Software*, **86**, 1–30.

MCLACHLAN, G. & PEEL, D. (2004). *Finite Mixture Models*. Wiley Series in Probability and Statistics. New York: Wiley.

PALAIODIMOS, L., KOKKINIDIS, D.G., LI, W., *et al.* (2020). Severe obesity, increasing age and male sex are independently associated with worse in-hospital outcomes, and higher in-hospital mortality, in a cohort of patients with COVID-19 in the Bronx, New York. *Metabolism*, **108**, 154262.

PUNZO, A. & INGRASSIA, S. (2016). Clustering bivariate mixed-type data via the cluster-weighted model. *Computational Statistics*, **31**, 989–1013.

ROSSI, C., BERTA, P., CURELLO, S., *et al.* (2021). The impact of COVID-19 pandemic on AMI and stroke mortality in Lombardy: evidence from the epicenter of the pandemic *medRxiv*.

SPINELLI, D., INGRASSIA, S. & VITTADINI, G. (2022). Cwmglm: stata module to estimate cluster weighted models (CWM). *Statistical Software Components*. https://ideas.repec.org/c/boc/bocode/s459090.html

SUBEDI, S., PUNZO, A., INGRASSIA, S. & MCNICHOLAS, P. (2013). Clustering and classification via cluster-weighted factor analyzers. *Advances in Data Analysis and Classification*, **7**, 5–40.

ZHANG, H., HAN, H., HE, T., *et al.* (2021). Clinical characteristics and outcomes of COVID-19–infected cancer patients: a systematic review and meta-analysis. *JNCI: Journal of the National Cancer Institute*, **113**, 371–380.