THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Multimorbidity profiles and stochastic block modeling improve ICU patient clustering

Citation for published version:
Restocchi, V, Villegas, JG & Fleuriot, JD 2022, Multimorbidity profiles and stochastic block modeling improve ICU patient clustering. in M Fazio, DK Panda, R Prodan, V Cardellini, B Kantarci, O Rana & M Villari (eds), *Proceedings of the 22nd IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing, CCGrid 2022.* Institute of Electrical and Electronics Engineers (IEEE), pp. 925-932, 22nd IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing, CCGrid 2022, Taormina, Italy, 16/05/22. https://doi.org/10.1109/CCGrid54584.2022.00112

Digital Object Identifier (DOI):
10.1109/CCGrid54584.2022.00112

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings of the 22nd IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing, CCGrid 2022

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 19. Apr. 2024

Multimorbidity profiles and stochastic block modeling improve ICU patient clustering

Valerio Restocchi
School of Informatics
The University of Edinburgh
Edinburgh, United Kingdom
v.restocchi@ed.ac.uk

Jorge Gaete Villegas
School of Informatics
The University of Edinburgh
Edinburgh, United Kingdom
j.a.gaete@ed.ac.uk

Jacques D. Fleuriot
School of Informatics
The University of Edinburgh
Edinburgh, United Kingdom
jdf@ed.ac.uk

***Abstract*—Identifying groups of patients with similar morbidity profiles can help us understand the relationships between their pre-existing conditions and the risks of adverse events in the ICU. To find such groups, common approaches apply clustering algorithms such as k-means and latent class analysis. However, these techniques present drawbacks such as the lack of principled methods for choosing the number of clusters, the need for assumptions about the relationships between variables, and outputs which are hard to explain. To overcome these limitations, we map the problem of patient clustering to that of community detection in complex networks. We construct a bipartite network in which nodes represent patients and their features, including morbidities and demographics. Then, we find homogeneous groups of patients using stochastic block modeling (SBM), an unsupervised probabilistic approach to find structure in networks. We show that this approach has several advantages over traditional clustering methods, and enables us to retrieve more fine-grained clusters that are commonly missed by existing approaches. We also show that these clusters have a stronger relationship with mortality and sepsis rates of patients in the ICU.**

***Index Terms*—Multimorbidity, Patient clustering, Critical care, AI for healthcare, Community detection, Unsupervised learning, Stochastic block modeling, Networks for health**

I. Introduction

With the global population getting older and older, the prevalence of long-term illnesses has become the main challenge in healthcare [1]. This becomes even more apparent in critical care, where the medical outcome of a patient depends on the interaction of a number of factors. The coexistence of multiple long-term illnesses, or multimorbidity, is one of the most important of such factors and it heavily contributes to the heterogeneity of adverse outcomes [2, 3, 4, 5].

For this reason, clustering techniques, such as k-means, have been used to identify multimorbidity profiles and contextualize the potential impact of specifically adapted treatments, including those for sepsis [6, 7, 8]. However, even when they are successful, these computational techniques commonly rely on heuristics to optimize an objective function, and have been criticized due to their dependency on the quality of the data and simplicity of their models [9]. In response to such concerns, latent class analysis (LCA) has been proposed as an alternative [9, 10]. Recently, it has been used on laboratory and demographic data to identify four multimorbidity profiles with different prevalence of sepsis and mortality [11]. Similarly, using data on pre-existing conditions, Zador et. al. identified six multimorbidity profiles [12].

Despite being a more principled approach and addressing some of the issues of other clustering algorithms, LCA still shares some of the same drawbacks: First, it does not use a principled method to define the optimal number of clusters, which could lead to poor results [5, 13]. Second, similar to traditional clustering techniques, LCA makes assumptions about the causal relationship between variables, assumptions that might however be unrealistic, consequently affecting the quality of the results [5]. Finally, LCA uses unobserved variables to find clusters, making them hard to understand as they cannot be interpreted directly from the observed data [14].

Network science is the study of complex systems based on the connections between their constituting elements [15], which makes it particularly suitable to analyze the complex interactions between long-term disorders in patients. In complex networks, the analogous of clustering is called community detection, a problem whereby the entities constituting a network (nodes) are grouped based on their connectivity patterns [16]. However, most of the methods proposed to do this are heuristic and, therefore, show similar problems to those of traditional clustering techniques [17, 18]. To overcome this problem and provide a more robust solution to the task of patient clustering, we propose the use of stochastic block modeling (SBM) [19, 20]. This encompasses a family of generative models commonly used for community detection that, thanks to their probabilistic approach, are not prone to the same issues that affect heuristic methods [18, 21]. Specifically, to accurately address the shortcomings of the clustering techniques discussed above, we use the hierarchical stochastic block model (hSBM), a non-parametric version of SBM that provides hierarchical clusters and therefore a better resolution, and has been already successfully applied to clustering outside standard community detection [21, 22, 23].

One of the main advantages of this approach is that the

VR and JF are partially funded by the National Institute for Health Research (NIHR) [Artificial Intelligence for Multiple Long-Term Conditions (NIHR202639)]. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. JG is funded by the Chilean National Agency for Research and Development (ANID) [DOCTORADO BECAS CHILE/2017-72180000]

hSBM is an unsupervised, non-parametric method, and therefore it does not require any input or assumption about the data. Additionally, this guarantees that the model cannot overfit and find structure where there is none [21, 22]. Finally, the model presents a hierarchical cluster structure that can facilitate the visualization of the solutions, and enhance their interpretability [24].

We use the hSBM to detect clusters of patients based on their multimorbidity and demographic information. Our results show that it finds clusters with homogeneous demographic and multimorbidity profiles that explain the data in more detail than in recent work [12], also uncovering important groups of patients missed by existing approaches. Additionally, these groups show distinct, statistically significant sepsis and mortality rates which are more informative than those suggested by existing methods used in critical care.

II. Methods and Data

In this section we briefly describe the dataset used for our analysis, and present the hSBM and its use in the context of our work.

A. Dataset

To perform our analysis, we use information on 38,417 patients from the Medical Information Mart in Critical Care (MIMIC-III), an anonymized dataset comprising information on the admissions to the critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012, which has been widely used since its release [25, 26].

Following Zador et al. [12], we consider age, sex, admission type (elective, non-elective) and secondary diagnoses as features for our study. Patients under 16 years of age are not considered, and the age of the rest is discretized into the following ranges: 16-24, 25-44, 45-64, 65-84, and over 85. We only consider the first ICU admission for patients with multiple ones. The resulting dataset is consistent with that reported by Zador et. al in terms of both demographic (gender and age) and morbidity distributions.

Morbidities are computed from the rich collection of secondary diagnoses by using the Elixhauser comorbidity index [27], a well established method to detect long-term disorders in patients based on ICD-9 codes. Finally, sepsis is computed following the definition given by Angus et al., whose implementation is available from the official MIMIC-III code repository [26, 28]: A patient is considered to have sepsis if it is explicitly recorded in the their history, if there exists an infection (bacterial/fungal) and organ dysfunction, or if there exists an infection (bacterial/fungal) and the patient is under mechanical ventilation.

B. Hierarchical stochastic block model for patient clustering

The stochastic block model (SBM) is a generative model commonly used for community detection in complex networks, and is based on the assumption that nodes have a given probability to connect to each other, and that this probability solely depends on the community (or block) to which they belong [29, 30]. A major limitation of the SBM is that the model requires prior information on the number of blocks the network possesses [17]. To address this issue, recent studies have proposed a series of Bayesian, non-parametric versions of the SBM [21, 22]. One of such versions, the hierarchical – or nested – stochastic block model (hSBM), provides hierarchical clustering and was introduced to improve the resolution limits of the non-parametric SBM, allowing the model to discover finer-grained clusters [22].

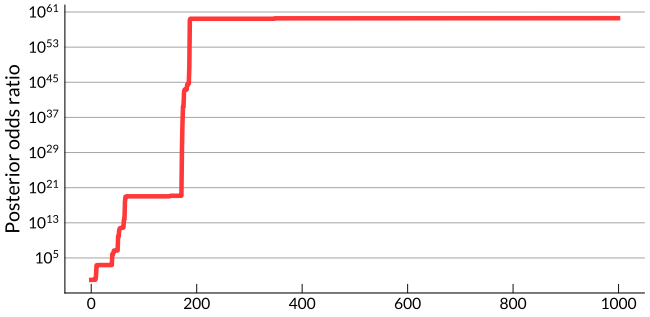

Fig. 1. This figure displays the posterior odds ratio obtained at every batch run of the merge-split MCMC. We run the algorithm for 1,000 batches of 10 runs each. The posterior probability of a partition is computed for each of these batches. Then, we compute the posterior odds ratio by dividing the posterior probability of the partition obtained at a given batch run by the posterior probability of the partition we obtained after running the agglomerative multilevel MCMC 100 times. It is possible to note that after around 200 batch runs, the posterior odds ratio stabilises, after having reached a state whose partition is $\approx 10^{60}$ more likely than the one found initially by the agglomerative multilevel MCMC.

In this paper, we use the hSBM to identify informative clusters of homogeneous patients. To this end, we create a bipartite unweighted network, in which one set of nodes represents patients, whereas the other represents their features, including demographics, admission type, and morbidities. Each patient node is connected to its features, as shown in Fig. 2. Then, we use the hSBM to find clusters of patients. There are several ways by which the hSBM can find the best partition, so we decide to follow a procedure that gives the highest probability of not getting stuck in a local minimum. Specifically, we first use an agglomerative multilevel Markov Chain Monte Carlo (MCMC) algorithm that starts by partitioning each node in a different cluster and then, at each step, proposes moving nodes to different clusters [31]. These moves are accepted with a given probability based on their resulting minimum description length gain. Given its stochastic nature, this algorithm is not guaranteed to always find the best partition. To limit this possibility, we run the algorithm 100 times and keep the run that yields the highest posterior probability. However, this still does not ensure that the partition obtained is optimal, as there is still a small chance that the algorithm found a solution corresponding to a local minimum. For this reason, we further refine the resulting partition by running another optimization algorithm on it, the merge-split MCMC, proposed to address this issue [32]. We run this 10,000 times, in batches of 10, to ensure that no further significant improvement is

possible. Indeed, we find that, after roughly 200 run batches, the improvement in the posterior likelihood reaches a plateau, as can be seen in Fig. 1. This result suggests that the clusters obtained by the hSBM are either optimal or near-optimal.

III. Cluster analysis

We compare our clusters with those obtained by Zador et al. using LCA [12] through a discussion of their composition and relationship to the prevalence of sepsis and mortality among the patients that belong to them.

A. The benchmark - LCA for patient clustering

Zador et al. find groups of patients with similar morbidity profiles and the relationship of such groups with sepsis and mortality rates [12]. We use this work as our benchmark for three reasons: First, this is one of the few studies that cluster patients based on their multimorbidity profiles. Second, to achieve this they use MIMIC-III. Third, they focus their analysis of clusters on adverse outcomes such as mortality and sepsis. To the best of our knowledge, this is currently the state-of-the-art for studies that possess all three of the aforementioned characteristics.

Zador et al. find six groups of patients with statistically significant different multimorbidity profiles. They compute the prevalence of sepsis and organ dysfunction for all the clusters, and the associated mortality rates. Their results suggest that two commonly used scores by clinicians to assess the risk of sepsis and mortality at the time of admission, namely the Oxford Acute Severity of Illness Score (OASIS)[33] and the Sequential Organ Failure Assessment (SOFA) [34], provide predictions which are in contrast to the observed prevalence of adverse events in their clusters. This implies that, by using information on multimorbidity, it is possible to improve the assessment of adverse outcomes. Despite this promising finding though, not all the clusters they obtain show discernible mortality and sepsis rates [12]. Also, these clusters miss some categories of patient with far higher or lower probabilities to develop sepsis or die, such as those patients with elective admissions or those with no long-term illnesses, respectively. In the next section, we show that our approach can uncover these groups of patients, which are missed by both LCA, and OASIS and SOFA scores.

B. Clinically homogeneous multimorbidity clusters

Fig. 2 presents the clusters and their hierarchy as found using the hSBM. At the highest hierarchical level, it is possible to see that the hSBM captures the bipartite structure of the network, separating features (on the left-hand side) from patients (on the right-hand side).

At the intermediate hierarchical level, the hSBM identified six patient clusters. This is the same number of clusters identified by Zador et al. with LCA. However, despite some similarities, the clusters provided by the two methods are very different. In both cases, it does not seem that gender plays a significant role in assigning patients to a group. Moreover, both approaches identified clusters of similar age profiles.

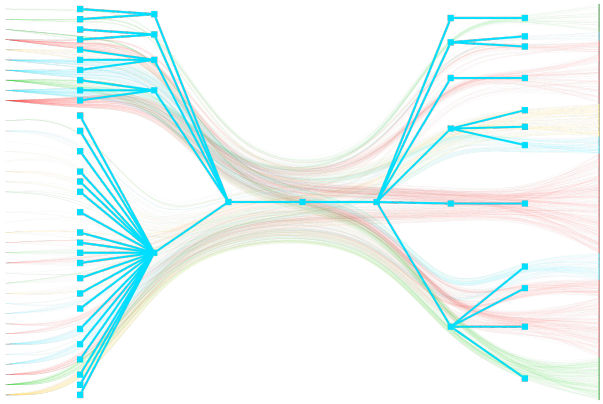

Fig. 2. The bipartite network of patients and their features. Patients are displayed on the right-hand side, and features are displayed on the left-hand side of the network. Only 1000 edges out of the 235,137 between patients and their features are randomly sampled to be displayed here, to reduce visual clutter. The tree illustrates the hierarchical structure of the clusters we find, and is discussed in Sec. III.

For instance, the hSBM includes most patients in the age ranges 16-24 and 25-44 in the same clusters – clusters A and B, Fig. 3 –, as does LCA [12]. Similarly, both groups of patients over 85 and patients in the age range 65-84 are clustered separately with both methods. Morbidities aside, the main difference is that the admission type (i.e. elective vs non-elective admissions) plays a role in several of the clusters found by the hSBM. We can clearly see this in clusters B and D, in which patients have a low prevalence of elective admissions (roughly 60% lower than average). This difference becomes even more pronounced when considering clusters at the bottom hierarchical level, with clusters B1, D1, and D2 showing little to no elective admissions. This is a level of detail that LCA could not capture.

Another major difference in cluster composition emerges when examining multimorbidity profiles: LCA separates patients based on several groups of diseases, suggesting that multimorbidity is the main separation criterion for this algorithm [12]. Furthermore, they show that in all clusters patients have a multimorbidity count, which suggests that only the morbidity type plays an effective role in grouping patients.

Our results are in stark contrast to these. In fact, the hSBM discerns patients not only based on multimorbidity profiles, but also their combinations with demographics, admission type and, importantly, the number of morbidities patients have (see Fig. 5). Specifically, we see that only two clusters, D and F, include patients with a heterogeneous multimorbidity count, whereas the others include patients with a definite number of morbidities. For instance, a remarkable finding by our approach is that none of the 2716 patients in cluster A have morbidities. This cluster, in fact, represents younger patients with no long-term conditions who have been admitted to the hospital following some traumatic event, such as a car accident, a stab or gun wound. Importantly, this category of patients is expected and of particular interest, but is unde-

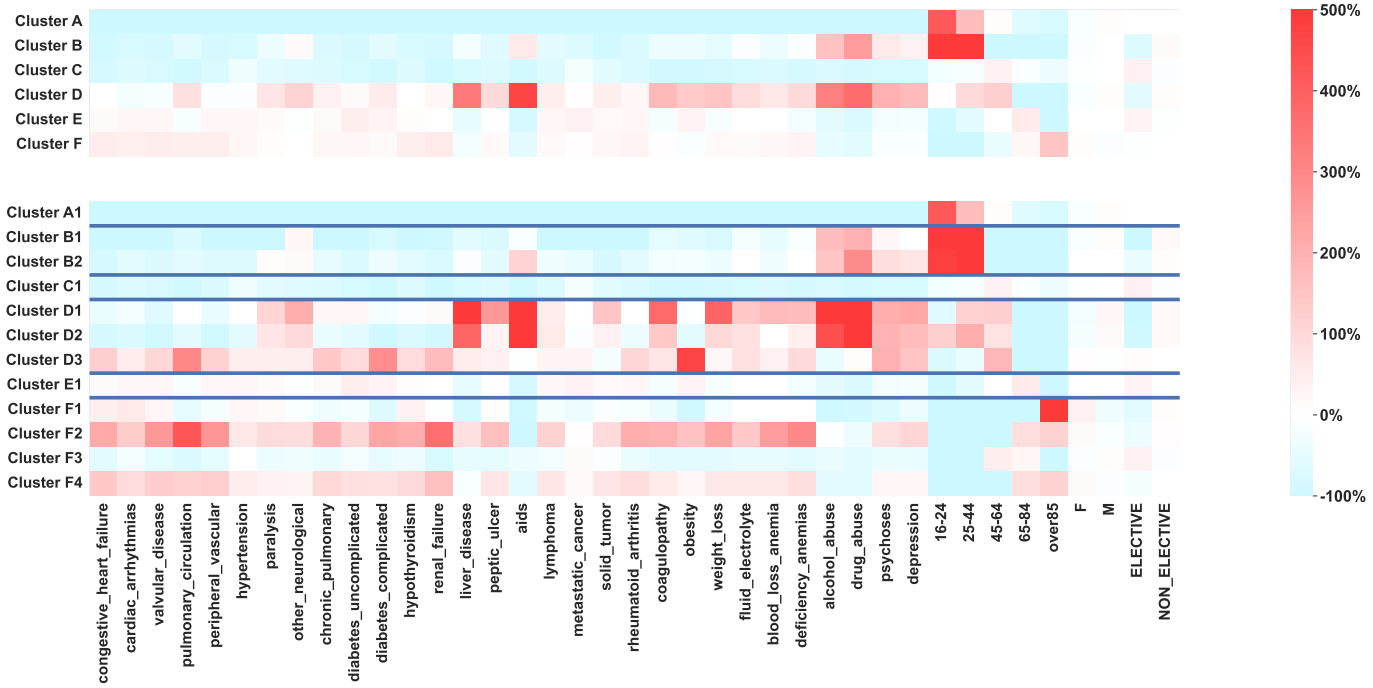

Fig. 3. This figure illustrates the multimorbidity composition of patient clusters inferred by the hSBM. For example cluster B splits into clusters B1 and B2. The lines between clusters at the bottom level group clusters that originate from the same parent cluster at the intermediate level. The heatmap shows the relative difference in the prevalence of morbidity between each cluster and the whole dataset. Although relative changes can be larger, we cap the colormap at 500% to improve readability.

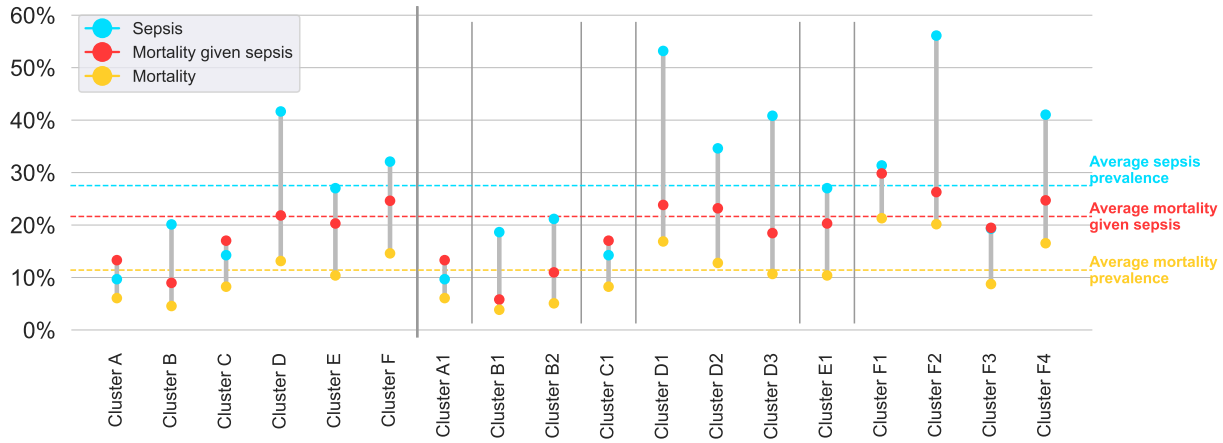

Fig. 4. This figure displays the heterogeneity in the prevalence of sepsis, mortality, and mortality given sepsis across the clusters we find. These are compared to the respective averages in the whole subset of 38,417 patients we use in our analysis. Especially considering the fine-grained clusters at the bottom hierarchical level, starting from A1, it is possible to see that some clusters present far higher than average prevalence of sepsis and mortality, such as D1, F1, and F3, which is not captured by the OASIS and SOFA scores computed at admission (Fig. 7). Similarly, some of the clusters we uncover display a high prevalence of patients with low risk of developing sepsis and die, such as clusters from A1 to C1, which is once again in contrast with the information provided by SOFA and, in particular, OASIS scores.

tectable by LCA. This is also reflected in the much lower than average mortality and sepsis rates for these patients (see Fig. 4), which is not captured by either LCA or the SOFA and OASIS scores (see Fig. 7).

C. Analysis of fine-grained clusters

In the previous section, we showed how the six clusters obtained at the intermediate hierarchical level display information which could not be retrieved with traditional clustering and LCA. At the lowest hierarchical level, these clusters split into 12, more fine-grained ones, that discriminate even further between complex patient and multimorbidity profiles and provide greater insights on sepsis and mortality in the ICU.

Given the different number of clusters, an accurate compari-

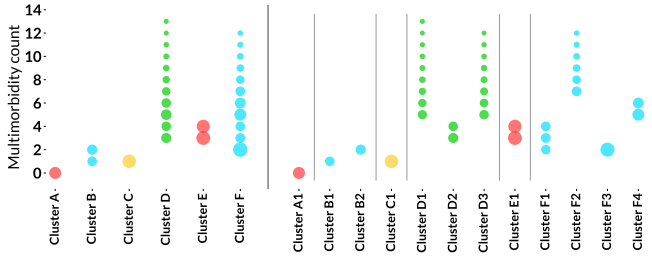

Fig. 5. This figure displays the multimorbidity count – i.e. the number of co-existing long-term disorders – by each cluster at both the intermediate and bottom hierarchical levels. The circle size is proportional to the frequency of patients who have exactly that number of morbidities. It is interesting to see that, contrary to existing literature, we find several clusters in which the patients have only few, if any, long-term illnesses. Importantly, the fact that clusters such as A1, B1, C1, etc. are composed of patients who all have the same multimorbidity count, suggests that the number of morbidities has a major role in determining the clusters. The exact number seems to matter less though when the multimorbidity count becomes higher, such as in the case of clusters D1, D3, and F2.

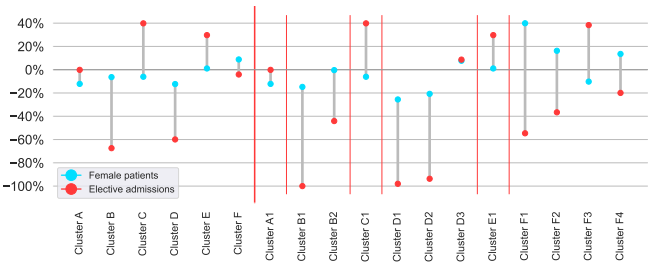

Fig. 6. This figure shows the relative change of gender and admission type prevalence between each cluster and the average among all patients. It is immediate to see that gender plays a major role in clustering, with only two clusters – D1 and D2 – having a significantly lower than average prevalence of female patients, and only F1 having a significantly higher prevalence. Conversely, we can see that the admission type is more influential. For instance, D1 and D2 have little to no patients with elective admissions, whereas C1 is composed of many patients scheduled critical surgery.

son between the hSBM and LCA at this point would hardly be significant. Instead, we provide a detailed analysis of each of these 12 clusters next, focusing on their composition, and on the mortality and sepsis rates for the patients they represent. We will further discuss the similarities and differences with LCA clusters in the discussions in Sec. IV.

Cluster A1 – Young patients without morbidities. This cluster has a much higher than average prevalence of younger patients, who have no morbidities and expectedly show a far lower prevalence of sepsis (9.68% vs 27.52%) and mortality given sepsis (13.3% vs 21.6%) than average (Fig. 4).

Clusters B1 and B2 – Younger patients with substance abuse issues and non-elective admissions. These two clusters are also defined by patients under 45 years of age, but compared to A1 they present a higher proportion of patients in the age range 25-44. Besides age, from Fig. 3 it is possible to see that patients in these two clusters are distinguished by a much lower than average prevalence of elective admissions, and by a prevalence of drug abuse which is 3 and 3.89 times higher than average for B1 and B2, respectively. Similarly, alcohol abuse is present in 22.74% and 21.17% of the patients, whereas the average in the dataset is 8.42%. It is worth noting that the main factor that distinguishes these two – otherwise similar – clusters is that in cluster B1 all patients only have one morbidity each, whereas in B2 they all have exactly two (see Fig. 5). This is reflected in the higher prevalence of AIDS, psychoses and depression that we can see in B2, but also in the fact that patients grouped in B1 have a lower mortality rate after they develop sepsis (Fig. 4). Overall, these two clusters have a low mortality rate, both with and without sepsis, and also lower than average prevalence of sepsis, which is coherent with the fact that patients in these clusters are younger and have only one or two long-term conditions.

Cluster C1 – Elective admissions of middle-aged patients with exactly one morbidity. This cluster is the one with the highest rate of elective admissions, which is 38.38% higher than average (Fig. 6). From Fig. 5, we can see that patients in this cluster only have one morbidity, but on average, the prevalence of all morbidities is lower than in the whole dataset (Fig. 3). To gain a better idea of who these patients are, we inspect the most common causes of admission. We find that these are mostly patients who are receiving coronary artery bypass graft surgery, which is compatible with the fact that most admissions in this cluster are elective. This might also explain the second lowest prevalence of sepsis among the clusters.

Clusters D1, D2, D3 – Patients with mental and neurological disorders. At the intermediate hierarchical level, cluster D showed a significantly high prevalence of substance abuse, followed by a number of other conditions with higher than average prevalence such as AIDS, coagulopathy, liver disease and fluid electrolyte disorder, all compatible with substance abuse [35, 36]. However, the split of this cluster into D1, D2, and D3 tells a much more complex story. Clusters D1 and D2 still present the highest prevalence of substance abuse, AIDS, and liver disease among all clusters (see Fig. 3). Similar to B1 and B2, the main difference among D1 and D2 is the multimorbidity count, which is either 3 or 4 for all patients in D2 but is far higher and more heterogeneous for patients in D1. If analyzed together with the age profiles – in D1 patients are older than in D2 – this suggests that D1 groups those patients who are in the later stage of substance abuse: these patients have developed a number of long-term illnesses that are not as present in their younger counterpart, such as higher prevalence of liver disease, coagulopathy, peptic ulcer, and weight loss. This is unsurprisingly reflected in their far higher sepsis and mortality rates (Fig. 4).

Interestingly, despite sharing the same parent cluster as D1 and D2, cluster D3 does not show an higher than average substance abuse prevalence, but instead displays the highest number of obese patients – 28.22% vs an average of 4.92% – and a number of associated cardiovascular and metabolic conditions, including diabetes with and without complications and pulmonary circulation disorders. Although at first sight D3 may seem very different from D1 and D2, it is possible to see

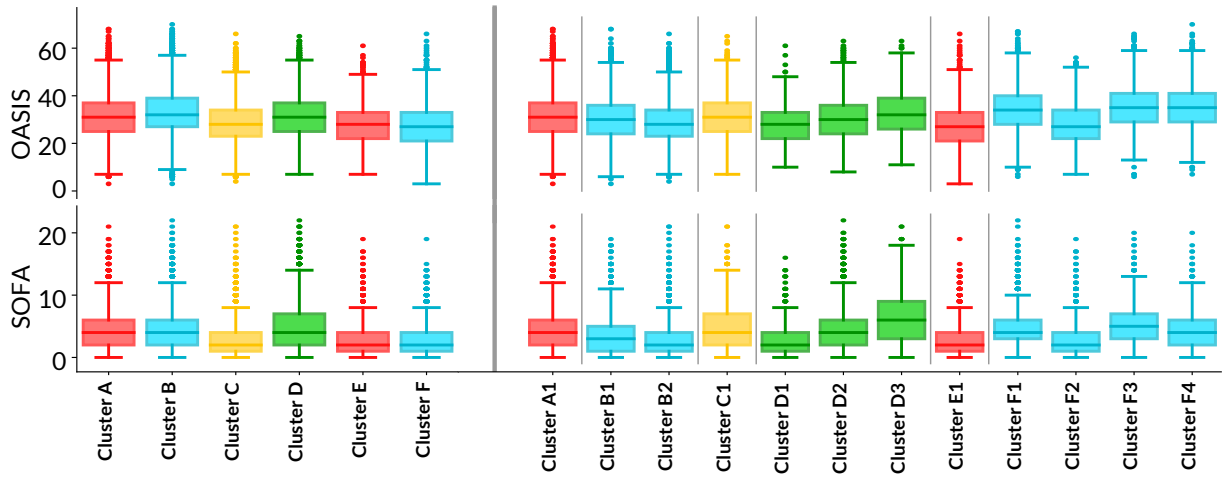

Fig. 7. This figure shows the distribution of OASIS and SOFA scores among the patients in our clusters. As discussed in Sec. III-D, we find that the actual rates of sepsis and mortality in our clusters follow different patterns that what suggested by these scores.

that all these three clusters share a very similar prevalence of neurological disorders, psychoses, depression, and even paralysis. Further, D3 and D1 are linked by a very similar – almost identical – multimorbidity count distribution.

Cluster E1 – Elderly patients with multimorbidity and elective admissions. The defining characteristics of this cluster are age – with a higher prevalence of elderly patients– elective admissions – 29.71% higher than average – and the number of morbidities, either 3 or 4. Apart from this, patients in this cluster show a much lower than average prevalence of substance abuse, AIDS and liver disease, but slightly higher prevalence of everything else. This cluster is representative of the dataset, so it does not come as a surprise that its sepsis and mortality rates are close to the average.

Clusters F1, F2, F3, and F4 – Patients over 45 with heterogeneous multimorbidity profiles. At the intermediate level, cluster F shows that its patients are predominantly older – with a peak of patients over 85 years old – and have a number of co-occurring morbidities. By inspecting its ramifications at the bottom level from Fig. 3, we can see that we can further divide this category of patients into two groups. Clusters F1 and F3 are mostly composed of patients who do not have a high number of morbidities, between 2 and 4 and exactly 2, respectively. The main difference between these two clusters is the age of the patients, which is strictly over 85 for F1 and between 45 and 84 for F3. Remarkably, this difference allows us to uncover significantly different mortality rates between the two groups, which is far higher at 21.29% for the older patients of F1, compared to 8.75% of patients that belong to F3, as it can be seen in Fig. 4. The other two clusters of this group, F2 and F4, are also fairly similar in terms of multimorbidity profiles and demographics, but differ in the number of morbidities their patients have. In fact, although patients in both clusters have a high multimorbidity count, those who belong to F4 have either 5 or 6, whereas patients in F2 strictly have at least 7 morbidities (Fig. 5). This is clearly reflected in their sepsis prevalence, which is the highest among all clusters at 56.11%, more than double the average of all patients. A similar observation can be made for their mortality rate, which is 21.29%, or 86.5% higher than the average. These observations are to be expected, given both the age of the patients and the fact that they have complex multimorbidity profiles which include a number of cardiovascular diseases [37].

D. Comparison of sepsis and mortality rates with OASIS and SOFA scores

OASIS and SOFA are commonly used scores for patient risk assessment at time of admission. OASIS focuses on assessing the risk of mortality, whereas SOFA is used to evaluate the risk of organ dysfunction (and consequently sepsis) [33, 34]. These scores have the merit of being easy to compute, not requiring any laboratory data, and easy to understand. However, these scores do not take into account multimorbidity, despite it being a key factor in determining the outcome of a patient, as we have seen from our results in Sec. III-C. Our results suggest that these scores alone are not accurate enough to be effectively used in the risk assessment of a patient. Specifically, we want to investigate whether the relative risk of mortality and sepsis across our clusters as predicted by OASIS and SOFA is coherent with the prevalence of these two adverse outcomes we find. By inspecting our results from Fig. 7 and Fig. 4 it is possible to see that OASIS and SOFA provide scores which are in disagreement with the actual prevalence of adverse outcomes we find in our clusters. There are two stark examples of this that we will use to illustrate our point. First, it is cluster D, whose *children* clusters D1, D2, and D3 have progressively higher OASIS and SOFA scores. However, if we analyze the actual prevalence of mortality and sepsis, we see that the highest are found in cluster D1. In fact, we find that mortality rates progressively decrease, rather than increase, in these clusters. The second, and perhaps most powerful example is represented by F2, which is the cluster with the

lowest average SOFA score but the highest prevalence of sepsis and the second highest mortality rate across all clusters. These results show that our approach provides significantly different insights into sepsis and mortality rates than common scores and can potentially be used to assess the risk of a patient more accurately.

IV. Discussion

By representing patients' data as a bipartite network, we can map the problem of patient clustering to community detection, and find structure in our data using the hSBM, a non-parametric Bayesian generative model. Thanks to this approach, we are able to unveil complex relationships between multimorbidity and adverse events in the ICU, and find more significant profiles than existing methods. Our results show that our approach has three distinctive advantages over LCA.

First, we are able to retrieve clusters that would otherwise be undetected, such as those with low multimorbidity prevalence. For instance, cluster A1 only includes patients with no pre-existing conditions, who have been admitted to the hospital primarily due to traumatic events. Not surprisingly, this cluster has the lowest sepsis rates of all. Moreover, the two clusters in which patients only have one or two long-term disorders, namely B1 and B2, have the lowest mortality rates. Second, from our comparison with LCA it is immediate to see that the hSBM captures more fine-grained relationships. In fact, although our approach still identifies clusters similar to those reported by Zador et al., revolving around substance abuse, cardiovascular diseases, diabetes, etc., it is also capable to differentiate them into more detailed depictions. This is clear from analyzing, for instance, clusters B and D. They both represent substance abuse clusters but, within cluster B, patients are younger and have a low prevalence of multimorbidity, whereas in cluster D patients are older and have a higher prevalence of disorders commonly associated with drug abuse (Fig. 3). These differences create a significant divide in mortality and sepsis rates, which is not observed in Zador et al. [12]. Third, the hSBM is non-parametric, and it does not even require input on the number of clusters. Thanks to the network representation of the data, the fact that the model is non-parametric ensures that no overfitting occurs, and, consequently, that the resulting clusters are completely unbiased, even in presence of highly unbalanced data. A remarkable consequence of these features is that we find several clusters in which patients display an exceedingly high prevalence of characteristics which have, instead, a particularly low prevalence in the whole dataset, such as being between 16 and 24 years old (2.9%), obesity (4.9%), peptic ulcer (0.82%), and AIDS (0.57%). A second, equally important, consequence is that, unlike recent work, all our clusters include patients with a largely homogeneous number of long-term disorders [11, 12].

Besides comparison with other clustering methods, we also show that the sepsis and mortality rates found in our clusters largely differ from the predictions made by OASIS and SOFA scores. For this reason, we argue that our results constitute robust evidence that multimorbidity should be included in critical care risk assessment.

V. Conclusion

There is currently a limited understanding of the co-occurrence of long-term health conditions and their associated health outcomes due to the complex interactions between morbidities and also interactions with other factors such as demographics. Our work shows that hierarchical stochastic block modeling and, more generally, a network representation of patient data offer several intrinsic advantages, such as the elucidation of fine-grained associations, over traditional clustering methods. It is an original contribution to a growing number of research efforts aimed at mapping and identifying disease clusters and understanding adverse health outcomes – in our case sepsis and death in the ICU – for people with complex sets of pre-existing conditions.

References

[1] K. Barnett, S. Mercer, M. Norbury, G. Watt, S. Wyke, and B. Guthrie, "Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study," *The Lancet*, vol. 380, pp. 37–44, 2012.

[2] P. Kalgotra, R. Sharda, and J. M. Croff, "Examining health disparities by gender: A multimorbidity network analysis of electronic medical record," *International journal of medical informatics*, vol. 108, pp. 22–28, 2017.

[3] J. C. Forte, A. Perner, and I. C. van der Horst, "The use of clustering algorithms in critical care research to unravel patient heterogeneity," *Intensive care medicine*, pp. 1–4, 2019.

[4] K. Reddy, P. Sinha, C. M. O'Kane, A. C. Gordon, C. S. Calfee, and D. F. McAuley, "Subphenotypes in critical care: translation into clinical practice," *The Lancet Respiratory Medicine*, vol. 8, no. 6, pp. 631–643, 2020.

[5] L. Busija, K. Lim, C. Szoeke, K. M. Sanders, and M. P. McCabe, "Do replicable profiles of multimorbidity exist? systematic review and synthesis," *European journal of epidemiology*, vol. 34, no. 11, pp. 1025–1053, 2019.

[6] C. W. Seymour, J. N. Kennedy, S. Wang, C.-C. H. Chang, C. F. Elliott, Z. Xu, S. Berry, G. Clermont, G. Cooper, H. Gomez *et al.*, "Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis," *JAMA*, vol. 321, no. 20, pp. 2003–2017, 2019.

[7] G. Geri, P. Vignon, A. Aubry, A.-L. Fedou, C. Charron, S. Silva, X. Repessé, and A. Vieillard-Baron, "Cardiovascular clusters in septic shock combining clinical and echocardiographic parameters: a post hoc analysis," *Intensive care medicine*, vol. 45, no. 5, pp. 657–667, 2019.

[8] G. Papin, S. Bailly, C. Dupuis, S. Ruckly, M. Gainnier, L. Argaud, E. Azoulay, C. Adrie, B. Souweine, D. Goldgran-Toledano *et al.*, "Clinical and biological clusters of sepsis patients using hierarchical clustering," *PloS one*, vol. 16, no. 8, p. e0252793, 2021.

[9] D. Rindskopf and W. Rindskopf, "The value of latent class analysis in medical diagnosis," *Statistics in medicine*, vol. 5, no. 1, pp. 21–27, 1986.

[10] J. A. Hagenaars and A. L. McCutcheon, *Applied latent class analysis*. Cambridge University Press, 2002.

[11] Z. Zhang, G. Zhang, H. Goyal, L. Mo, and Y. Hong, "Identification of subclasses of sepsis that showed different clinical outcomes and responses to amount of fluid resuscitation: a latent profile analysis," *Critical Care*, vol. 22, no. 1, pp. 1–11, 2018.

[12] Z. Zador, A. Landry, M. D. Cusimano, and N. Geifman, "Multimorbidity states associated with higher mortality rates in organ dysfunction and sepsis: a data-driven analysis in critical care," *Critical Care*, vol. 23, no. 1, pp. 1–11, 2019.

[13] K. Nasserinejad, J. van Rosmalen, W. de Kort, and E. Lesaffre, "Comparison of criteria for choosing the number of classes in bayesian finite mixture models," *PloS one*, vol. 12, no. 1, p. e0168838, 2017.

[14] M. Moshkovitz, S. Dasgupta, C. Rashtchian, and N. Frost, "Explainable k-means and k-medians clustering," in *International Conference on Machine Learning*. PMLR, 2020, pp. 7055–7065.

[15] M. E. J. Newman, *Networks*. Oxford university press, 2018.

[16] M. A. Javed, M. S. Younis, S. Latif, J. Qadir, and A. Baig, "Community detection in networks: A multidisciplinary review," *Journal of Network and Computer Applications*, vol. 108, pp. 87–111, 2018.

[17] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Physics reports*, vol. 659, pp. 1–44, 2016.

[18] T. P. Peixoto, "Descriptive vs. inferential community detection: pitfalls, myths and half-truths," *Pre-print: arXiv:2112.00183v4*, 2022.

[19] P. Holland, K. Blackmond Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social Networks*, vol. 5, pp. 109–137, 1983.

[20] B. Karrer and M. E. J. Newman, "Stochastic blockmodels and community structure in networks," *Physical Review E*, vol. 83, no. 016107, 2011.

[21] T. P. Peixoto, "Nonparametric bayesian inference of the microcanonical stochastic block model," *Physical Review E*, vol. 95, no. 1, p. 012317, 2017.

[22] ——, "Hierarchical block structures and high-resolution model selection in large networks," *Physical Review X*, vol. 4, no. 1, p. 011047, 2014.

[23] M. Gerlach, T. P. Peixoto, and E. G. Altmann, "A network approach to topic models," *Science advances*, vol. 4, no. 7, p. eaaq1360, 2018.

[24] A. Vellido, "The importance of interpretability and visualization in machine learning for applications in medicine and health care," *Neural computing and applications*, vol. 32, no. 24, pp. 18 069–18 083, 2020.

[25] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[26] A. E. W. Johnson, D. J. Stone, L. A. Celi, and T. J. Pollard, "The MIMIC code repository: enabling reproducibility in critical care research," *Journal of the American Medical Informatics Association*, vol. 25, no. 1, pp. 32–39, 2018.

[27] A. Elixhauser, C. Steiner, D. R. Harris, and R. M. Coffey, "Comorbidity measures for use with administrative data," *Medical care*, pp. 8–27, 1998.

[28] D. C. Angus, W. T. Linde-Zwirble, J. Lidicker, G. Clermont, J. Carcillo, and M. R. Pinsky, "Epidemiology of severe sepsis in the united states: analysis of incidence, outcome, and associated costs of care," *Critical Care Medicine*, vol. 29, no. 7, pp. 1303–1310, 2001.

[29] E. M. Airoldi, D. M. Blei, E. A. Erosheva, and S. E. Fienberg, *Handbook of mixed membership models and their applications*. CRC press, 2015.

[30] P. Doreian, V. Batagelj, and A. Ferligoj, *Advances in Network Clustering and Blockmodeling*. John Wiley & Sons, 2019.

[31] T. P. Peixoto, "Efficient monte carlo and greedy heuristic for the inference of stochastic block models," *Physical Review E*, vol. 89, p. 012804, 2014.

[32] ——, "Merge-split markov chain monte carlo for community detection," *Physical Review E*, vol. 102, p. 012305, 2020.

[33] A. E. Johnson, A. A. Kramer, and G. D. Clifford, "A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy," *Critical care medicine*, vol. 41, no. 7, pp. 1711–1718, 2013.

[34] J.-L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. Reinhart, P. Suter, and L. G. Thijs, "The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure," 1996.

[35] H. S. Ballard, "The hematological complications of alcoholism," *Alcohol health and research world*, vol. 21, no. 1, p. 42, 1997.

[36] D. Salmon-Ceron, C. Lewden, P. Morlat, S. Bévilacqua, E. Jougla, F. Bonnet, L. Héripret, D. Costagliola, T. May, and G. Chêne, "Liver disease as a major cause of death among HIV infected patients: role of hepatitis C and B viruses and alcohol," *Journal of Hepatology*, vol. 42, no. 6, pp. 799–805, 2005.

[37] A. M. Arnold, B. M. Psaty, L. H. Kuller, G. L. Burke, T. A. Manolio, L. P. Fried, J. A. Robbins, and R. A. Kronmal, "Incidence of cardiovascular disease in older americans: The cardiovascular health study," *Journal of the American Geriatrics Society*, vol. 53, no. 2, pp. 211–218, 2005.