A new method for patient stratification based on multi-layer network modeling and molecular data integration

Manuela Petti*, Davide Mascolo†, Caterina Alfano‡, Lorenzo Farina*

*Dept. of Computer, Control and Management Engineering, Sapienza University of Rome, Italy.
†MSc Program in Data Science, Sapienza University of Rome, Italy.
‡PhD Program in Network Oncology and Precision Medicine, Sapienza University of Rome, Italy.

***Abstract*—With the development of increasingly complex and heterogeneous diseases, there has been a need to use computational methods to support traditional medicine and gain a deeper understanding of the optimal treatment for patients with complex diseases such as cancer. This work focused on developing a new approach based on multi-layer network modeling and integrating molecular data for patient stratification. We applied the proposed method considering two different datasets containing molecular data collected among 215 patients affected by Glioblastoma Multiforme (GBM) and 393 patients suffering from Non-Small Cell Lung Cancer (NSCLC). The proposed method outperformed the state-of-the-art algorithm Similarity Network Fusion (SNF) obtaining a better stratification of GBM patients for overall survival. Considering the NSCLC patients, there is a significant stratification of patients ($p - value \leq 0.05$) with clinical characterization by sex, and age that opens up future debates and studies on the topic of sexual dimorphism. The proposed method successfully addresses both challenges of patients' stratification and biomedical data integration with potential clinical impact in therapy identification.**

***Index Terms*—Network-based, omics integration, multi-layer network, patient stratification, non-small cell lung cancer, glioblastoma multiforme.**

I. Introduction

Precision medicine is considered an emerging approach for disease treatment that takes into account individual characteristics of patients in terms of molecular profile, lifestyle, and clinical history [1] [2]. The advantage of this approach is used through computational methods that can integrate several data sources to create a comprehensive view of the disease. The emerging paradigm driving these methods is network science, which models patient relationships across multiple data layers using network approaches to achieve meaningful patient stratification [3]. This is an important goal for biomedical research, given that within a unique cohort of patients affected by the same disease, it is possible to identify subgroups based on their clinical features and/or molecular profiles, especially for complex and heterogeneous diseases such as cancer. Patient Similarity Network (PSN) [8] is the most common paradigm used for patient stratification where each node in a network is seen as a patient and the edges between nodes represent the similarity between patients. As reported in Fig. 1, this approach starts with several data layers for which the idea is to build a PSN. The next step is integrating the different PSNs into a single multi-layer network model that will be the patient stratification step's input.

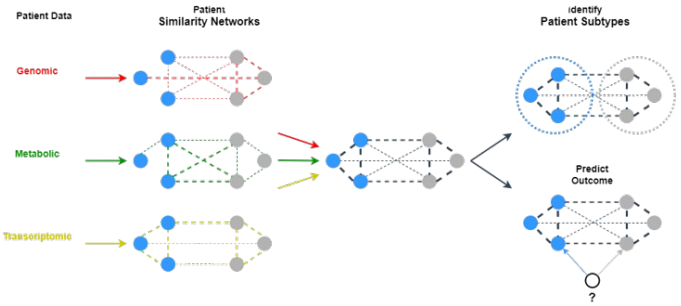

Fig. 1. PSN Framework

Until this study, there were different successful PSN-based methods for patient stratifications and the state-of-the-art is represented by the SNF algorithm [4]. Wang and colleagues developed it to integrate a set of unimodal PSNs and it overcomes the limitations of Network-Based Stratification [5] related to the use of a not well-defined network to map the mutations. The SNF has been applied with success in many different applications [6] [7] but even in the case of this method, several studies proposed some changes to improve and expand SNF. Building on this knowledge gap, the authors propose a new approach for patient stratification that leverages multi-layer network modeling and molecular data integration to surpass state-of-the-art methods.

Sapienza University of Rome supported this work, "Progetto di Ateneo 2021", grant number: RM12117A8B3C3033 and "Progetto di Ateneo 2023", grant number: RMRM123188F7C3D2B3.

Email: manuela.petti@uniroma1.it, davidemascoloofficial@gmail.com, caterina.alfano@uniroma1.it, lorenzo.farina@uniroma1.it

II. Materials and Methods

A. Data

To test the proposed method, the authors have used two independent cohorts of patients suffering from different types of cancer. We downloaded GBM data for 215 patients from the TCGA website selecting the three types: gene expression, miRNA expression, and DNA methylation. The set of patients is composed mostly of men, precisely 62% of men and women represented by 38% as reported in Fig. 2.

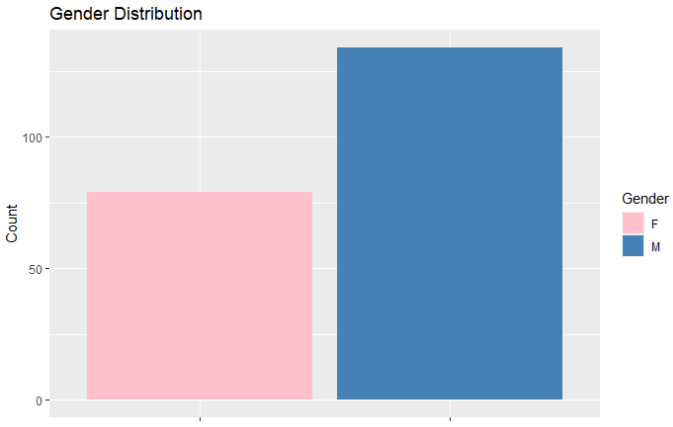

Fig. 2. Gender Distribution in GBM patients

For NSCLC data, the set of 393 patients is downloaded considering the work done by Ravi and colleagues [10], with clinical response information, and patients ranging in age from 29 to 90 years, composed of 182 men and 207 women. All the patients were treated with an anti-PD-(L)1 agent for immunotherapy considering the genomic and transcriptomic profile.

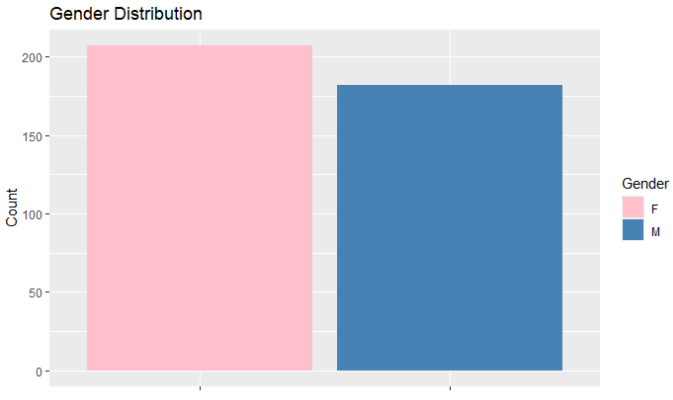

Fig. 3. Gender Distribution in NSCLC patients

B. Framework

For each data type in the two datasets, we generated the PSN by calculating patient similarity using Euclidean distance. Then we stack the unimodal PSNs obtaining a multi-layer network model and we applied for 1000 iterations genLouvain [11] to identify a common partition. The output of [11] is a *patient x iteration* matrix on which the Lloyd k-means algorithm is applied to extract the most stable partition from the 1000 iterations and to evaluate the survival curves of the identified clusters. The Generalized Louvain algorithm was applied using a MATLAB package developed by Lucas G. and colleagues [12]. Using this approach, it is need to build the similarity matrices *patient - patient* for each data layer. After that, these matrices are used to generate the multilayer matrix. To do that it is necessary to define two parameters: intralayer parameter γ and interlayer parameter ω. For this study, γ was set to 0.82 and ω equal to 1. The output of the Generalized Louvain algorithm is the partition matrix represented in Fig. 4, where each row represents a different patient, and the columns represent the number of iterations.

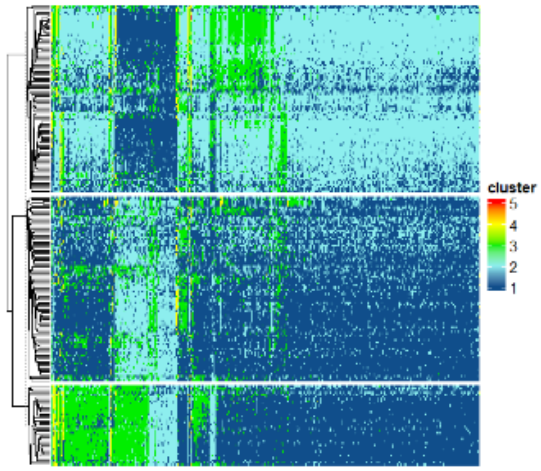

Fig. 4. Heatmap patient x iteration

Considering the Fig. 4, it is clear that many patients are classified often into the same cluster, index of homogeneity of the result. Moreover, in most cases, the algorithm identified two or three clusters, reaching a maximum of five clusters.

III. Results

A. GBM Patients

Fig. 5 shows the overall survival of the 3 subgroups of GBM patients identified with the proposed approach. The three curves appear well separated from the beginning. No alternative combination of the three layers (use of single datasets or pairs) has returned a stratification associated with such a clear separation of the survival curves. Also, SNF returned a significant stratification (Log-rank test, $p - value = 2x10^{-4}$), but 2 subtypes are not well separated [4].

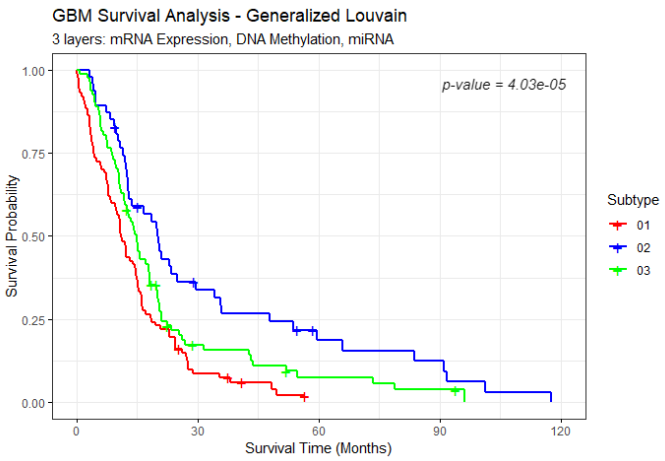

Fig. 5. Survival curves of the Patients' Subtypes in GBM

The goal was to identify the most impactful clinical variable for classifying the identified clusters. Looking at Fig. 6 it is possible to note the statistical significance of the *Age* variable. Subtype 2 is the youngest (*median age = 41 years*) but with the greatest intra-group variability. Subtype 1, which corresponds to the lowest survival probability curve, is older (*median age = 60 years*). In the middle, there is Subtype 2 (*median age = 55.5 years*).

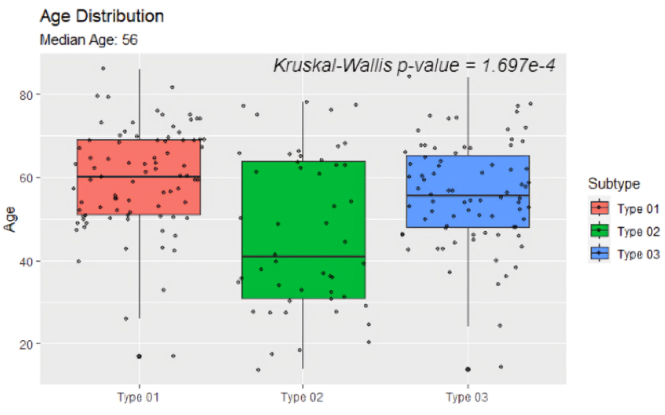

Fig. 6. Age Distribution of the Patients' Subtypes in GBM

B. NSCLC Patients

For NSCLC patients, the same type of analysis was carried out. In particular, the proposed approach identified two well-separated subgroups as shown in Fig. 7, and this result is obtained by integrating genomic and transcriptomic layers with clinical signatures (curated, myeloid, and immune).

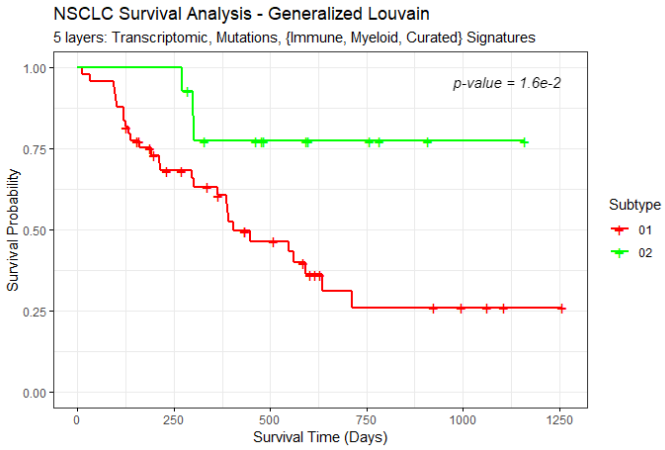

Fig. 7. Survival curves of the Patients' Subtypes in NSCLC

Compared with the survival curves obtained with the SNF method, there is a clear separation between the two groups identified from the beginning, and the two curves never overlap, an effect that was the case with the SNF method. The most impactful clinical variables for the classification are *Age*, *Gender*, and *Line of Therapy*, which were tested using the clinical information available for the cohort and the Kruskal-Wallis test. Starting to study the *age* variable, is shown in Fig. 8 the distribution conditioning for the identified subtypes. It follows that Subtype 1 is older (*median age = 65 years*), while Subtype 2 is younger (*median age = 56 years*).

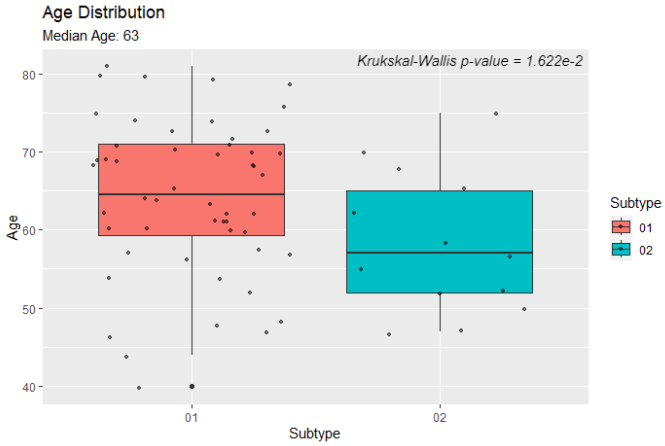

Fig. 8. Age Distribution of the Patients' Subtypes in NSCLC

The *Gender* variable emerges as a very interesting key to interpretation regarding the still-open problem of Sexual Dimorphism [13]. In particular, as shown in Fig. 9, the gender distribution is also significant with Subtype 2 characterized by more than 85% women patients. At the same time, patients within Subtype 1 are mainly women of an advanced age who have a lower probability of survival than patients in Subtype 2. The low survival in older women (Subtype 1) could be dictated by a menopausal state [13].

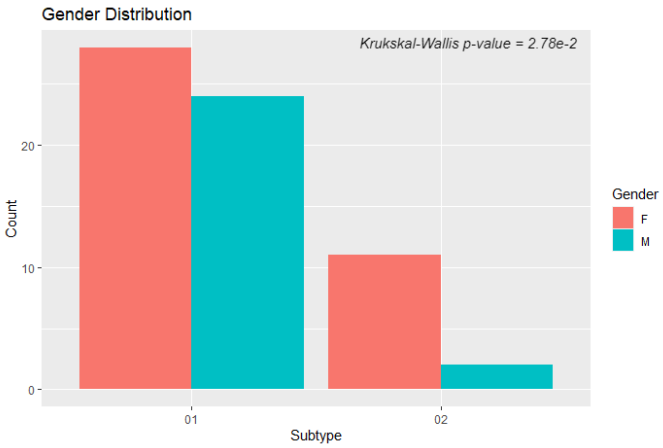

Fig. 9. Gender Distribution of the Patients' Subtypes in NSCLC

Sex	Type 01 (%)	Type 02 (%)
Female	52.9	85.7
Male	47.1	14.3

TABLE I
PERCENTAGE DISTRIBUTION AGE VARIABLE IN NSCLC

The last statistically significant clinical variable, was the line of therapy, consisting of five different levels depending on the patient's therapy status. As shown in Fig. 10, Subtype 1 includes patients who received all lines of therapy, that is, from the first line of therapy to the fifth. Subtype 2, on the other hand, which corresponds to the group with a higher probability of survival, consists only of patients who have been treated with the first and second lines of therapy. Thus, reviewing the history of clinical variables previously analyzed, it is clear that the subtype with the least likelihood of survival for patients with NSCLC is composed of individuals of women sex, in an advanced age state, and who due to sexual dimorphism resist immunotherapy treatment, which is why with Subtype 1, it is up to the fifth line of therapy.

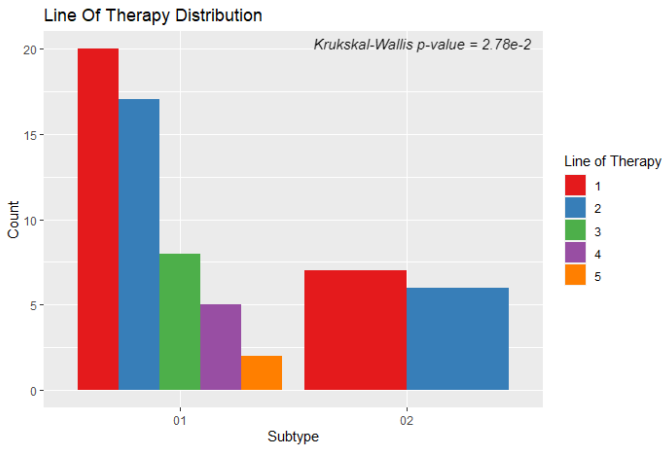

Fig. 10. Line of Therapy Distribution of the Patients' Subtypes in NSCLC

Therapy	Type 01 (%)	Type 02 (%)
1	37.3	57.1
2	33.3	42.9
3	15.7	-
4	9.8	-
5	3.9	-

TABLE II
PERCENTAGE DISTRIBUTION LINE OF THERAPY VARIABLE IN NSCLC

CONCLUSIONS

From the proposed work, it is clear the importance of using computational methods based on patient similarity networks, and the integration of omics data to achieve statistically significant stratification of patients within the same cohort. The proposed method, not only outperforms the SNF algorithm but also brilliantly overcomes both the challenges of patient stratification and biomedical data integration with tremendous clinical potential in therapy identification and optimization. From a technical point of view, the proposed method outperformed the SNF algorithm obtaining a better stratification of GBM patients in terms of overall survival, identifying three statistically significant subtypes, with the first group of patients oldest and also with the lowest probability of survival compared with the other two groups. Considering NSCLC patients, there is a significant stratification of patients with clinical characterization mainly by sex, and age. Looking at the survival curves it is clear that patients who are within Subtype 1 are the ones who have the lowest probability of survival. Similarly, among the statistically significant clinical variables, patients in Subtype 1 have the highest median age of 65 years. In addition, the variable Age suggests that the majority of patients in Subtype 1 are women. From this analysis, the low probability of survival for older women emerges. Considering the immunotherapy treatment received by the patients on whom data were recorded, and the median age of patients in Subtype 1, it is likely that this result is related to the phenomenon of sexual dimorphism [13], which refers to biological and physiological differences between the

sex of a species, including variations in physical characteristics, hormone levels, and immune responses. In particular, in women undergoing menopause, the significant decrease in estrogen can negatively affect the immune response, reducing the effectiveness of immunotherapy, which uses the immune system to fight cancer. The decrease of estrogen can alter the tumor microenvironment and affect tolerance to side effects of therapy, leading to different responses than premenopausal women and making it more difficult to benefit from treatment [14]. The proposed method captures the division by sex of the two subtypes and highlights the biological importance of the sex variable.

REFERENCE

1) Petti M, Farina L. Network medicine for patients' stratification: From single-layer to multi-omics. WIREs Mech Dis. 2023 Nov-Dec;15(6):e1623. doi: 10.1002/wsbm.1623. Epub 2023 Jun 15.
2) Erikainen, S., & Chan, S. (2019). Contested futures: envisioning "Personalized," "Stratified," and "Precision" medicine. New Genetics and Society, 38(3), 308–330. https://doi.org/10.1080/14636778.2019.1637720.
3) Farina L. Network as a language for precision medicine. Ann Ist Super Sanita. 2021 Oct-Dec;57(4):330-342. doi: 10.4415/21.04.08. PMID:35076423.
4) Wang, B., Mezlini, A., Demir, F. et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods 11, 333–337 (2014). https://doi.org/10.1038/nmeth.2810
5) Hofree, M., Shen, J., Carter, H. et al. Network-based stratification of tumor mutations. Nat Methods 10, 1108–1115 (2013). https://doi.org/10.1038/nmeth.2651
6) Bhalla S, et al. Patient similarity network of newly diagnosed multiple myeloma identifies patient subgroups with distinct genetic features and clinical implications. Sci Adv. 2021 Nov 19;7(47):eabg9551. doi: 10.1126/sciadv.abg9551. Epub 2021 Nov 17. PMID: 34788103; PMCID: PMC8598000.
7) Cavalli FMG, et al. Intertumoral Heterogeneity within Medulloblastoma Subgroups. Cancer Cell. 2017 Jun 12;31(6):737-754.e6. doi: 10.1016/j.ccell.2017.05.005. PMID: 28609654; PMCID: PMC6163053.
8) Pai S, Bader GD. Patient Similarity Networks for Precision Medicine. J Mol Biol. 2018 Sep 14;430(18 Pt A):2924-2938. doi: 10.1016/j.jmb.2018.05.037. Epub 2018 Jun 1. PMID: 29860027; PMCID: PMC6097926
9) D. Mascolo, L. Farina and M. Petti, "Multi-layer network modelling of genomic and transcriptomic data to investigate the response to checkpoint inhibitors in NSCLC," 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Istanbul, Turkiye, 2023, pp. 2825-2830, doi: 10.1109/BIBM58861.2023.10385375.
10) Ravi, A., et al. Genomic and transcriptomic analysis of checkpoint blockade response in advanced non-small cell lung cancer. Nature Genetics, 55(5), 807–819.
11)] Pj Mucha et al. Community Structure in Time – Dependent Multiscale, and Multiplex Networks, Science, 2010 May.
12) Mucha, P. J., Richardson, T., Macon, K., Porter, M. A. Onnela, J.-P. Community structure in time-dependent, multiscale, and multiplex networks. Science 328, 876-878 (2010).
13) Klei, S., et al. Sex differences in immune responses. Nat Rev Immunol 16, 626-638 (2016).
14) Klein SL, Flanagan KL. Sex differences in immune responses. Nat Rev Immunol. 2016 Oct;16(10):626-38. doi: 10.1038/nri.2016.90. Epub 2016 Aug 22. PMID: 27546235.