Erkenntnis (2023) 88:2911–2932
https://doi.org/10.1007/s10670-021-00483-w

ORIGINAL RESEARCH

Patient Similarity in the Era of Precision Medicine: A Philosophical Analysis

Giovanni Boniolo[1] · **Raffaella Campaner**[2] · **Massimiliano Carrara**[3]

Received: 22 September 2020 / Accepted: 17 October 2021 / Published online: 19 November 2021

Abstract

According to N. Goodman, the Carnapian notion of similarity is useless in science and without interest for philosophy. In our paper we suggest that, given the current role that the notion of similarity has in managing biomedical big data, this drastic position should be revised, and similarity should be provided a scientifically useful philosophical interpretation. With the advent of the new sequencing technologies, imaging technologies and with the improvements of health records, the number of genomics, post-genomics and clinical data has exponentially increased. The deluge of data has urged, among others, to devise a new way of stratifying patients. A solution has been found and it is based exactly on the notion of similarity. By discussing two examples focusing on similarity among breast cancer patients, in the paper we illustrate such a use, and analyze it from a philosophical standpoint by resorting to A. Tversky's features matching approach. We believe that the latter can foster some better understanding of the meaning and current use of similarity in the context of biomedical big data, and that, therefore, be the focus of further reflections in the philosophy of science, in particular in the philosophy of biomedicine.

✉ Massimiliano Carrara
massimiliano.carrara@unipd.it

Giovanni Boniolo
giovanni.boniolo@unife.it

Raffaella Campaner
raffaella.campaner@unibo.it

[1] Department of Neuroscience and Rehabilitation, University of Ferrara, Via Fossato di Mortara, 64a, 44121 Ferrara, Italy

[2] Departments of Philosophy and Communication Studies, University of Bologna, Via Zamboni, 38, 40126 Bologna, Italy

[3] FISPPA Department, University of Padua, P. zza Capitaniato 3, 35139 Padova, Italy

1 Introduction

According to Goodman (1972), the Carnapian notion of similarity is useless in science and without interest for philosophy. We believe that, after fifty years, this position needs to be reconsidered in the light of personalized medicine, where big data have a central role.

Over the centuries, biomedical research has devised conceptual ways of grouping people with the same set of pathological features, to treat them in the same way. The advent of precision medicine has significantly affected the scenario, introducing an unprecedented amount of data in medical research and clinical practice, and thus forcing experts to rethink also classificatory practices. This biomedical novelty has been accompanied by new computational tools able to manage statistical models of vast sets of data. In particular, patients' molecular and clinical uniqueness, and the overwhelming abundance of information on their lifestyles and on the environments in which they live, can now be dealt with computational and statistical tools, such as the cluster theory. The notion of *similarity* has proved to be central for this approach allowing for rethinking classificatory practices: instead of grouping individuals on the basis of either having or not having certain features, it is possible to group them on the basis of molecular and clinical characteristics that render them mutually similar to some extent. This procedure has hence at its core the choices, respects and degrees of similarity/dissimilarity on which patients' grouping is grounded.

We believe that similarity among breast cancer patients can be taken as an illustrative case study for how to approach patient similarity in medicine and in biomedicine more generally. The relevant aspects of similarity will then differ for each area of biomedicine. But what do we exactly mean by *similarity* in this specific context? Actually, this question is twofold, and will be tackled by examining the possible use of some philosophical tools to deal with new approaches to patienthood. First of all (Sect. 2), we will present some crucial moments in the philosophical discussion on similarity which we believe can provide some relevant insights on the topic, starting from some brief recall of Goodman's and Carnap's positions (Sect. 2.1). We will then introduce the *Feature Matching Approach* (hereafter, FMA) proposed by A. Tversky in the late Seventies (1977), since it allows a proper philosophical framework of the idea of similarity at stake (Sect. 2.2). In Sect. 3, we will illustrate the Integrative Cluster approach and the Patient Similarity approach, stressing what has motivated their introduction in the medical context. It will then be argued that the use of similarity in current biomedicine can have, as just mentioned, Tversky's view as its philosophical counterpart (Sect. 4). After analyzing also some examples, we conclude that FMA offers adequate conceptual tools to better understand similarity and its use in biomedicine.

Here, we do not want to enter the wider philosophical debate on family resemblance terms, cluster concepts, and/or natural kind, which has touched upon a number of disciplines and cases. With respect to natural kinds, for the purpose of the paper let us just mention here that the sort of classificatory practices

investigated in the paper are not employed in biomedicine with the aim of grasping some hidden structure, providing some metaphysical carving of nature at its joint, or fixing some taxonomies reflecting ontological divisions in nature, but with the aim of finding clinically interesting statistical correlations. The specific scope of this paper is, rather, to show the actual relevance in contemporary biomedical practice of the notion of similarity, and to highlight, through Tversky's approach, that it is worthy of deeper investigation by means of philosophical tools.

2 Similarity and Distance from a Philosophical Perspective

Let us start recalling the main philosophical views on the notion of similarity, to introduce some theoretical tools and sketch the overall framework in which the paper's proposal is set.

2.1 The Geometrical Model

According to the original "geometrical model" (see, e.g., Carnap, 1928/1967) the notion of similarity is obtained via that of *similarity space*.[1] What is needed to define a similarity space is a set of points—the space—and a metric on this set of points; the metric is simply a set-theoretic function that for every pair of points in the space taken as arguments gives a real number as value. A metric space is an ordered pair $<X, d>$ where X is a space and d a metric such that it has the three following properties (where a and b are two points in the space):

- *Minimality*: $d(a, b) \geq 0$ and $d(a, a) = 0$.
- *Symmetry*: $d(a, b) = d(b, a)$.
- *Triangle Inequality*: $d(a, b) + d(b, c) \geq d(a, c)$.

The metric d represents the similarity relation (but it can be read also a dissimilarity relation); d(a, b) could then be taken as the real number representing the similarity between a and b (or the dissimilarity between a and b). Intuitively, taking d as expressing dissimilarity instead of similarity, *Minimality* just claims that an object is not dissimilar to itself (the real number associated by d to the pair formed by a and itself is 0) and that everything in the defined space is comparable. In other terms, that means that for an arbitrary pair of distinct points in the space, the degree of similarity or dissimilarity between them is always defined.

[1] It is worth recalling that the similarity notion based on the metric came out in geometry around 1906, thanks to the work of the French mathematicians René Fréchet (even if the name is due to Felix Hausdorff) when he discussed the notion of distance between two points of a topological space. Carnap's work was very close to the dawn of the mathematical birth of that notion.

Symmetry corresponds to the widely popular idea that similarity relations are symmetric, and the meaning of the axiom is that the degree of similarity between a and b is the same as that between b and a.

Finally, *Triangle inequality* corresponds to the idea that if b is similar/dissimilar to a certain degree to both a and c, then the degree of similarity/dissimilarity between a and c should be smaller than or even equal to the sum of the degree of dissimilarity between a and b and b and c. The intuitive idea is that the similarity of a to b and that of b to c constraints the similarity of a to c, namely that if a is quite similar to b and b is quite similar to c, then a and c cannot be very dissimilar from each other. Triangle inequality therefore corresponds to the idea that similarity relations are somewhat transitive relations or at least transitive with respect to a certain lower bound.

It is usually observed (see Decock & Douven, 2011 for a survey) that the geometrical model allows for a simple way of representing the similarity and/or dissimilarity between objects as a metric distance between the respective points in some uniform space, and therefore is able to offer a method to constructing spatial representations of similarity and dissimilarity relations, a similarity space. Another recognized advantage of the geometrical model is that it gives a straightforward methodology to compare similarity relations. Suppose you aim to model the claim that objects a and b are more similar to each other than objects c and d. To obtain it, it is sufficient to prove that $d(a, b) \geq d(c, d)$.

If the geometrical model works, we have a powerful tool to describe similarity space and similarity relations. It works in physics, where all the discussions concerning distance (both in classical physics, in relativity and quantum mechanics) adopt that geometric model. Nevertheless, physics is not the only field where distance and similarity can be used. Concerning this point, we should recall the strong criticism advanced by Goodman (1972) of the adoption of the geometrical interpretation in these different fields. He observed that one of the main difficulties of adopting a similarity relation is that it is highly contextual: "Comparative judgments of similarity often require not merely selection of relevant properties but a weighting of their relative importance, and variation in both relevance and importance can be rapid and enormous. Consider baggage at an airport checking station. The spectator may notice shape, size, color, material, and even make of luggage; the pilot is more concerned with weight, and the passenger with destination and ownership. Which pieces are more alike than others depends not only upon what properties they share, but upon what makes the comparison, and when [...] circumstances alter similarities" (Goodman, 1972, 445). That is a big issue for the geometrical model of similarity: it cannot represent the contextual dependence of similarity relations out of physics. The reason is that one of its fundamental assumptions, given in terms of the Minimality requirement, is that similarity measures are done within a unique, acontextual, space of comparison (for a hint on the debate, see Decock & Douven, 2011; Carrara & Morato, 2011). But, as Goodman rightly observed, what is similar in a certain context might be completely dissimilar given another context. And this is extremely important for grouping patients since their being patients of a certain kind strongly depends on the molecular and clinical features designing their pathological context.

There is a second problem for the geometrical model. Contrary to popular opinion, similarity relations are not in general taken as *symmetric* ones. This fact has been shown also by a series of psychological data. Tversky (1977) has shown that similarity judgements are often asymmetric: for example, people tend systematically to judge Tel Aviv as being more similar to New York than New York similar to Tel Aviv. Or, again, take three individuals you, your brother and another individual, call him "Sam". Sam, from a morphological point of view, is a sort of blend of you and your brother. Assume further that Sam is the person most similar to you (within a certain class of comparison). But suppose also that the degree of similarity between your brother and Sam is greater than the degree of similarity between you and Sam. Therefore, Sam is the person most similar to you, but you are not the most similar person (within the same comparison class) to Sam. And the same goes for patients. Let it be that the patient A is the most similar to patient B, and that the similarity between the patient B and a patient C is greater than the similarity between the patient A and the patient B. Therefore, the patient A is the individual most similar to the patient B, but the patient B is not the most similar to A, being most similar to C.

Finally, consider *triangle inequality*. Again, one can easily find a counterexample to the above-mentioned property of the geometrical model. Consider the following example. Cuba is similar to Jamaica for a certain degree (they are both Caribbean islands) and Cuba is similar to China (for their political affinity), but Jamaica is definitely not similar to China: the degree of dissimilarity between China and Jamaica is surely greater than the sum of the degrees of dissimilarity between Jamaica and Cuba and that between Cuba and China. So, also triangle inequality fails, at least in some cases, in particular when patients are at stake. If the patient A is similar to the patient B with respect to a certain set of molecular and clinical features, and if the patient B is similar to the patient C with respect to a different set of molecular and clinical features, the patient A is not similar to the patient C neither with respect to the first set nor with respect to the second set of features.

2.2 The Feature Matching Approach

There are two main different ways of bypassing such problems. The first one is the Tversky's FMA (Feature Matching Approach) (1977), the second one is the *Conceptual Space approach* proposed by Gärdenfors (2004). Let us focus on Tversky's account, since—as it will be shown—is more useful to our aims, and leave aside Gärdenfors' *conceptual spaces theory*, usually conceived as a refinement of the old Carnapian geometrical model. Gärdenfors' proposal is much more semantically and cognitively oriented than Tversky's and therefore less proper to our analysis, where the idea of grouping collection of features is central for our purpose. Indeed, similarity relations hold in the FMA for objects characterized as collections of features, whereas in the geometrical approach the class of objects over which the similarity relation has to be defined are points in a geometrical space. This is precisely the point we start from when grouping patients: a set of individuals characterized by a set of clinical features.

Given two objects, a and b, belonging to a certain domain D and characterized, respectively, by the set of features A and B, d(a,b) is a measure of the similarity of a to b. This means that anytime we have $d(a,b) > d(a,c)$ we have that a is more similar to b than to c. In the FMA (Tversky, 1977), similarity has to satisfy three conditions:

- The *Matching condition*, according to which the degree of similarity between two objects a and b is a function F of three sets: i) the set of their common features ($A \cap B$); ii) the set of the distinctive features possessed by a and not by b (A–B); iii) the set of the distinctive features possessed by b and not by a (B–A). That is, $d(a, b) = F(A \cap B, A–B, B–A)$
- The *Monotonicity condition*, which constraints similarity comparisons among objects, given a certain domain. Informally, the idea behind is that an object a is more similar to an object b than it is to an object c iff the common features of a and c are a subset of the common features of a and b and the distinctive features of a and c are subsets of the distinctive features of those of a and b. It follows that similarity increases with the addition of common features or deletion of distinctive features. That is, $d(a, b) \geq d(a, c)$ whenever $(A \cap B)$ is subset of $(A \cap C)$, (A–C) is subset of (A–B), (C–A) is subset of (B–A).
- The *Independence condition,* according to which the degree of similarity due to the joint effect of two features is independent of the degree of similarity that depends on the third feature. In other terms the *Independence* condition states the following: assume that the pairs of objects (a, b) and (c, d) as well as the pairs (a′, b′) and (c′, d′) agree on the same two features, while the pairs (a, b) and (a′, b′) as well the pairs (c, d) and (c′, d′) agree on a third feature. If this is the case, *Independence* predicts that a is more similar to b than a′ to b′ if and only if c is more similar to d than c′ is to d′; formally: $d(a, b) \geq d(a', b') \leftrightarrow d(c, d) \geq d(c', d')$

For Tversky any similarity relation that satisfies the Monotonicity, Matching and Independence conditions is a *matching function*. Matching functions F are used to measure degree of similarity, that is, they are analogues to distances in the geometrical model. There is an important characteristic of the functions that satisfy Monotonicity, Matching and Independence; for any similarity relation that satisfies the three conditions above, there are two interval measurement scales S and f such that:

1. $S(a,b) \geq S(c,d) \leftrightarrow d(a,b) \geq d(b,c)$.
2. $S(a,b) = \alpha f(A \cap B) - \beta f(A - B) - \gamma f(A - B)$.

this result is what Tversky called the *Representation theorem*.

Condition 1 claims that similarity comparisons between pairs of objects could be represented by an interval scale. The classical interval measurement scales are such that one unit on them represents the same magnitude on the trait or characteristic being measured across the whole range of the scale. A classic example of it is the

Fahrenheit scale for temperature. Condition 2 claims that similarity between a and b could be represented as a linear combination of the scales (i.e., the measures) of the common features and the distinctive features of a and b; α, β, and γ are three numbers by which the relevance of the common or the distinctive features could be weighted. If, for example, α is 1 and β, and γ are 0, then the similarity between a and b is the measure of their common features. If, on the other hand, α is 0 and β and γ are equal to 1, then the similarity between a and b is a measure of their respective distinctive features (also called "symmetric difference"). The gist of condition 2 is therefore that similarity is expressed as a "contrast" between the measures of the common and distinctive features; it is for this reason that the feature matching approach is sometimes called the *contrast model*.

As observed (see on this Decock & Douven, 2011, 65) the FMA overcomes many (if not all) of the shortcomings of the geometrical model.

Symmetry of similarity relation would hold in the FMA only by requiring, in condition 2 of the Representation Theorem, that β, and γ be equal. Cases where similarity is non-symmetrical are cases where the distinctive features of the first object are weighted more heavily than the distinctive features of the second one (see Tversky, 1977, 334).

The FMA also deals with the contextual dependence of similarity relations. The basic idea of the FMA is that similarity relations are determined in terms of common and distinctive features. The weight of such features is not absolute: they are doubly relativized to α, β, and γ and to the salience scale f. By means of such a scale, the salience of features may change from context to context.

Contextual dependence is managed in the FMA also by a certain elasticity in the choice of the domain of objects to consider when assessing certain similarity relations. Similarity relations tend to vary when the object set is changed (Tversky, 1977, 343).

As it will be shown below, contextual elements play an important part in grouping patients. In the following section we will present issues on grouping patients stemming from current biomedicine, and a solution presented therein, to then return to Tversky's account in Sect. 4.

3 Clusters via Similarity

The present work is novel insofar as it extracts relevant points of similarity from biomedical research and applies them. In this section, we will draw upon existing empirical research in biomedicine to discover relevant properties which can ground similarity. judgments of the sort Tversky's approach requires. As well-known, over the centuries, biomedical research has devised ways of grouping people with the same set of features, for diagnostic, prognostic and therapeutic purposes. In recent decades, clinical trials and evidence-based medicine have embraced this taxonomic approach, producing indications for drugs and clinical practice guidelines, each adapted to a distinct group of patients identified as homogeneous on the basis of a specific set of biomarkers (be them at tissue level, at cellular level or at molecular level). Guidelines are collectively produced documents defining a set of

recommendations, together with eligibility criteria restricting their applicability to a specific class of patients. Each new patient is allocated to one of the guideline-defined subgroups on the basis of certain biomarkers, and treatment is planned accordingly. This way of identifying classes of patients and placing individuals in the proper groups has continued to be implemented even with the advent of molecular medicine (see e.g. Boniolo and Nathan 2017; Barilan et al. 2021), where these groups were based on genes, proteins, metabolites, etc. However, with the progress of molecular medicine, new sequencing technologies, molecular imaging technologies and, above all, the major impact of computational and informational technologies—that is with the advent of precision medicine—new issues have been raised.[2] Soon the promises of this approach, designed to provide highly personalized and highly effective care, faced a substantial challenge, due to the fact that each patient is unique both from the molecular and the biographical point of view, and that an increasing amount of data is available on him/her. The more data (concerning the molecular profile and the biography) are collected, the more the set of the collected features of the patient is unique. How can then medicine provide a diagnostic or therapeutic account that works for many people if it is acknowledged that every single patient is molecularly (and clinically) unique? This conundrum is drastically evident with tumor heterogeneity, which shows not only that each cancer is individualized in a specific patient, but, more importantly, that each cancer affecting a given individual is actually composed of a set of different cancer subpopulations with heterogeneous features (see Boniolo, 2017).

An enormous number of individualizing features is potentially disruptive for the usual clinical trial process and evidence-based medicine paradigm, that rely on the possibility to group a statistically significant number of diseased individuals on the basis of their being carriers of a precise set of biomarkers. The uniqueness of conditions is recognized as a distinctive feature of some diseases, as kinds of tumors, but only by means of some sort of proper grouping would an adequate testing of medical hypotheses be possible, and findings of research conducted on a sample of the patient population be generalized to the whole population.

As recalled, classically we have an approach according to which we group people on the basis of *being* or *not being* carriers of given biomarkers. That is, something (an individual) either belongs or does not belong to a given set (group, stratum, cluster, class, cohort, reference class, etc.), depending on whether s/he exhibits or not

[2] We do not discuss here the limits and the potentialities of precision medicine, in particular if precision medicine is really precise or if it is always ethically praiseworthy (both at individual and global level). We do not even face the question whether a more proper definition of precision medicine exists, or which its historical roots are. This is not the right place to face these issues. For our sake, however, we pragmatically accept the well-known definition offered by the US National Research Council, according to which "precision medicine is 'an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person', [meant] […] to predict more accurately which treatment and prevention strategies for a particular disease will work in which groups of people": https://ghr.nlm.nih.gov/ Precision Medicine; on the relations between precision and personalized medicine, https://ghr.nlm.nih.gov/primer/precisionmedicine/precisionvspersonalized. See also https://www.nih.gov/research-training/allofus-research-program (Accessed 30 April 2017). See Barilan, Brusa and Ciechanover 2021.

a previously established set of biomarkers, at whatever level they could be. Nevertheless, given the complexity of many molecular diseases, the enormous amount of molecular and clinical information we have (see, for example, Leonelli, 2016, Strasser, 2019), and the myriad of unique features every single individual presents, this way of grouping has quickly become unconvincing, since, if driven to the limit, classes should ultimately be composed of only one member: a single individual patient. At the same time, it cannot be denied that medicine still needs, and will need, to group patients and strive to find drugs which would benefit many individuals, not only a single one. It seems currently unfeasible not to produce indications on how to treat groups of patients, and to deliberately limit the efficacy of research outcomes to just one patient. Furthermore, were we even willing to do so, we would be very unlikely to reach such a goal without starting from some sort of grouping of patients and analysis of some shared pathological features. This situation could, hence, create a dangerous impasse both in the search for new drugs and in the search for treatment protocols—as recognized even in the recent philosophical literature concerning the reference class problem, the narrowness of reference classes and the aim for precision (see, e.g., Fuller & Flores, 2015; Wallmann, 2017; Wallmann & Williamson, 2017).

Given their ultimate goal, i.e. to treat and cure, the biomedical sciences need to group patients. But how to do this, once the uniqueness of the molecular and clinical features of any individual is assessed? One solution that has encountered success in the scientific arena has been given in terms of computational technologies which utilize algorithms grouping patients on the basis of similarity relationships. That is, rather than grouping patients on the basis of them carrying certain markers, the idea is to group them on the basis of them being more or less *similar*.

To illustrate this epistemological shift, we wish to recall some works by Caldas and his team, who have opted for cluster analysis based on the notion of similarity. Their contribution constitutes a landmark in classification within cancer research (Ali et al., 2014; Bruna et al., 2016; Curtis et al., 2012; Pereira et al., 2016; Russnes et al., 2017). They had access to 997 samples from breast cancer patients stored in two biobanks (one in the UK and one in Canada) who were homogeneous for treatment and who were followed-up for about ten years. Utilizing new sequencing technologies, they undertook genomic and transcriptomic investigations, considering also the follow-ups. At the end of the computation process, they obtained ten different clusters of patients, which they called Integrative Clusters (iCluster, or IntClusters) and were also predictive. To be sure that the clusterisations properly did their job from a diagnostic and prognostic point of view, they applied the same grouping technique to a second cohort of about 1,000 breast cancer samples, and a third cohort of about 7500 samples. As illustrated below (Fig. 1), this technique allowed them to compare the clusterisations both with other molecular characterizations (e.g., PAM50[3]) and with the clinical outcomes. They successfully showed that

[3] PAM50 Prosigna® is a tumour-profiling test that helps determine the benefit of using chemotherapy in addition to hormone therapy for some estrogen receptor-positive (ER-positive) and HER2-negative breast cancers.

Table 1 Overview of the Integrative Cluster Subtypes and the Dominating Properties with Regard to Copy Number Driving Events, Biomarkers, Type of DNA Architecture,[46] Dominant PAM50 Subtype, and Clinical Outcome

Integrative cluster group	Copy number driver	Pathology biomarker class	DNA architecture	Dominant PAM50	Clinical characteristics (survival)
1	Chromosome 17/ chromosome 20	ER^+ ($HER2^+$)	Simplex/firestorm (chromosome 17q)	Luminal B	Intermediate
2	Chromosome 11	ER^+	Firestorm (chromosome 11q)	Luminal A and B	Poor
3	Very few	ER^+	Simplex/flat	Luminal A	Good
4	Very few	ER^+/ER^-	Sawtooth/flat	Luminal A (mixed)	Good (immune cells)
5	Chromosome 17 (*HER2* gene)	$ER^-(ER^+)/HER2^+$	Firestorm (chromosome 17q)	Luminal B and HER2	Extremely poor (in pre-Herceptin cohorts)
6	8p deletion	ER^+	Simplex/firestorm (chromosome 8p/ chromosome 11q)	Luminal B	Intermediate
7	Chromosome 16	ER^+	Simplex (chromosome 8q/chromosome 16q)	Luminal A	Good
8	Chromosome 1, Chromosome 16	ER^+	Simplex (chromosome 1q/chromosome 16q)	Luminal A	Good
9	Chromosome 8/ Chromosome 20	ER^+ (ER^-)	Simplex/firestorm (chromosome 8q/ chromosome 20q)	Luminal B (mixed)	Intermediate
10	Chromosome 5, Chromosome 8, Chromosome 10, Chromosome 12	TNBC	Complex/sawtooth	Basal-like	Poor 5-year, good long-term if survival

ER, estrogen receptor; TNBC, triple-negative breast carcinoma.

Fig. 1 Overview of the integrative cluster subtypes and the dominating properties. From Curtis et al. (2012)

their integrative classification reflected differences in chemotherapy. This might be seen as an unprecedented way to link molecular classification to clinical treatment, and to treatment outcomes. To achieve this result, Caldas et al. used a collection of breast cancer studies on patients who received chemotherapy adjuvants and whose data concerning the pathological complete response (pCR) were available.[4]

At the end, they were able to gather patients together, grouping them in clusters, on the basis of having features which are similar to features possessed by the other patients of the same cluster. This idea of grouping diseased individuals on the basis of similarity is gaining importance at research and clinical level, as witnessed by the fact that it has been increasingly used in the last few years to classify many kinds of cancer (Ross-Adams et al., 2015; Weddell et al., 2015; Guinney et al., 2015; Robertson et al., 2017; Cancer Genome Atlas Network, 2015), and to cope with tumor heterogeneity (Morganella et al., 2016; Nik-Zainal et al., 2012, 2016).

The same idea has led to a new approach called *Patient Similarity* (Brown, 2016; Pai & Bader, 2018; Parimbelli et al., 2018; Sánchez-Valle et al., 2020; for a review of the state of the art, see Dai et al. 2020). Whether or not this approach succeeds in the long run, it is an interesting case-study for re-discussing the different ways of grouping individuals in the light of shift in medical paradigms. It jointly addresses three aspects we have already recalled: (1) the vast amount of available data, thanks to the new sequencing and imaging technologies, from the "omics" levels of up to thousands of healthy and diseased individuals; (2) the bulk of clinical data (diagnoses,

[4] A tumour is said to have had a pCR if, after surgery, no residual cancer cells remain.

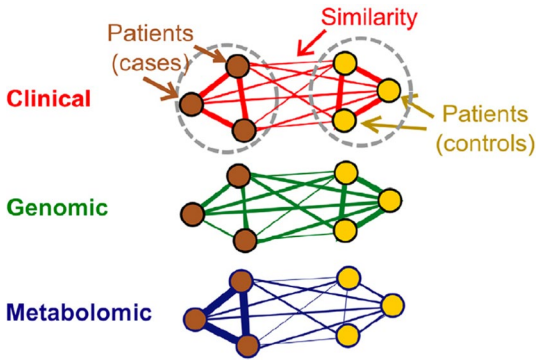
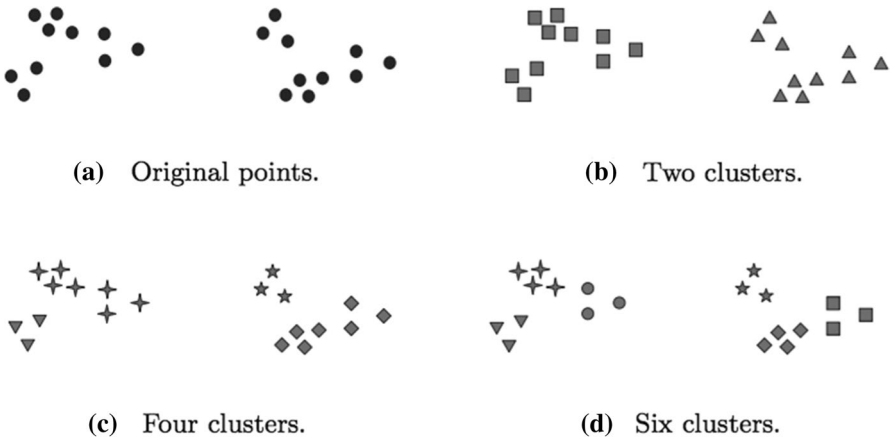


Fig. 2 How patients (nodes) are grouped by similarity (edges) considering three kinds of data (clinical, genomic and metabolomic data)



(a) Original points. **(b)** Two clusters.

(c) Four clusters. **(d)** Six clusters.

Fig. 3 Three different ways of clustering the same set of points (From Tan et al., 2017, 529)

laboratory results, prescriptions, therapies, response to treatment, disease progression, follow-up information, etc.) that electronic health records have allowed us to store and retrieve; iii) data concerning lifestyles and environments.

For example, Fig. 2 (from Pai & Bader, 2018) shows how similar patients (at the nodes) are linked together by edges (representing similarities) at different levels (clinical, genomic and metabolomic) and with other individuals serving as the control group. Whenever a new patient is considered, his/her data are inserted to find clinical, genomic and metabolomic similarities, and hence to decide in which group to include the patient, in order to propose a treatment and establish a prognosis.

Our concern here is with the notion of similarity: what do we *exactly* mean when we talk of *similarity* in this context? We remarked that an individual belongs to a group not because s/he possesses certain features per se, but because s/he possesses certain features which make him/her more or less similar to a certain group of patients already considered mutually similar. Here the notion of similarity has to be intended in term of distance. Thus, being more or less similar to a given patient means being more or less distant from him/her. Of course, if we use a different way

of implementing the distance, then patients will be grouped in different ways, as intuitively illustrated in the figure below (Fig. 3).

To illustrate how cluster analysis works, and to exemplify what a distance is, let us consider an example taken from Brown (2016),[5] which can help us grasp in which sense two patients (i.e., the two sets of data representing their "omics" and/or clinical features) are similar. In this formalization, a patient is represented by a vector defined in a multidimensional metric space, where each dimension represents a particular "omic" or item of clinical information. Thus, given two patients, represented by two vectors, their degree of similarity—and therefore the ground to establish if they belong to the same cluster—can be given, for example, by the so-called *cosine similarity*:

$$d(a,b) = \frac{a \cdot b}{\|a\|\|b\|} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}}$$

where a and b are the two vectors (representing the two patients), $a \bullet b$ is their scalar product, $\|a\|$ the module of the vector a and a_i its i-component (representing a molecular or clinical data). Hence the distance, that is, the similarity, is given in terms of the cosine of the angle between the two vectors (the two patients). That is to say, if the two patients are completely dissimilar, their vectors are opposite, thus the angle is 180° and the cos180° = −1. Instead, if the two patients are totally similar, they are represented by two equal vectors, thus the angle between them is 0° and $\cos 0° = 1$. It follows that, given a benchmark patient a, this approach enables us to grasp the similarity between him/her and any other patient b, by calculating $s(a, b)$. If we fix the degree of similarity, for example, between 0 (not included) and 1 (and therefore of the dissimilarity between 0 and − 1), for any new patient we can evaluate how similar (dissimilar) s/he is to the benchmark patient, and hence act accordingly in terms of treatment.

Both this example and the cases above show how this kind of similarity does not work in an abstract space (as we have seen above, the geometrical similarity discussed by Carnap and criticized by Goodman works), but in a well-defined statistical space whose points are given by the molecular and clinical information.[6]

At this point, to make even clearer the process leading to the grouping of the patients via similarity, it could be useful to show the flowchart of a standard patient similarity analysis (see Fig. 4; from Dai et al., 2020). It is relevant to note that the outcome depends also on the integration strategies chosen (graphically illustrated in Fig. 5; from Dai et al., 2020). For example, we could choose either the early data integration strategy (different data types are converted and standardized into the same format before calculating similarities) or the late data integration strategy (first

[5] For a more technical approach, see Zhu et al. (2016).

[6] This is also the reason why here the distance can be negative, while one of the conditions in the original metric space introduced in geometry and discussed by Carnap was that it has to be positive: a statistical space realized with biomedical data is different from an abstract topological space endowed with a metric.

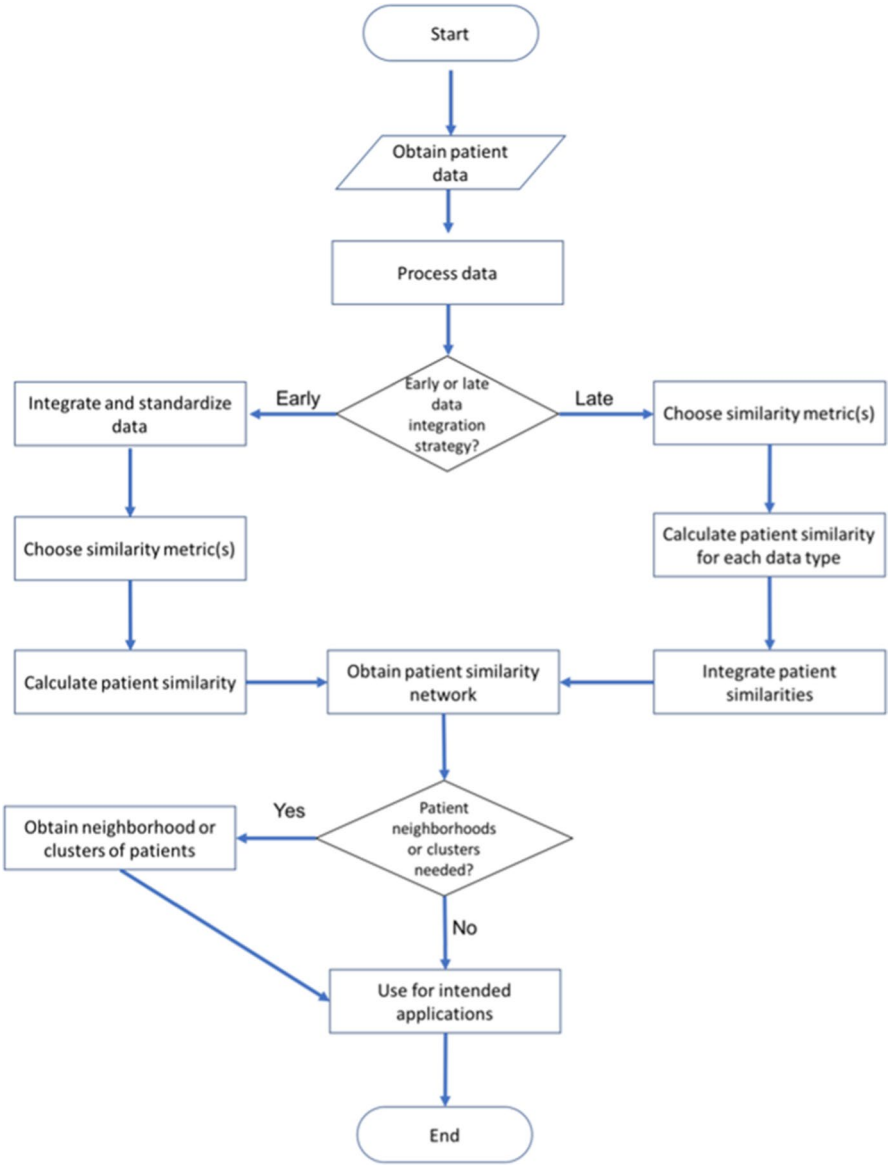


Fig. 4 Flowchart of a similarity analysis showing two different integration strategies

the similarity network is built for each data type, then networks are merged into a single one).

Before getting back to Tversky's view, let us also recall that other works on kinds might provide some interesting hints to enhance our understanding of similarity in medicine. In particular, Slater's work on stable property clusters intends to account for the joint presence of range of properties (see e.g. Slater, 2015). However, while

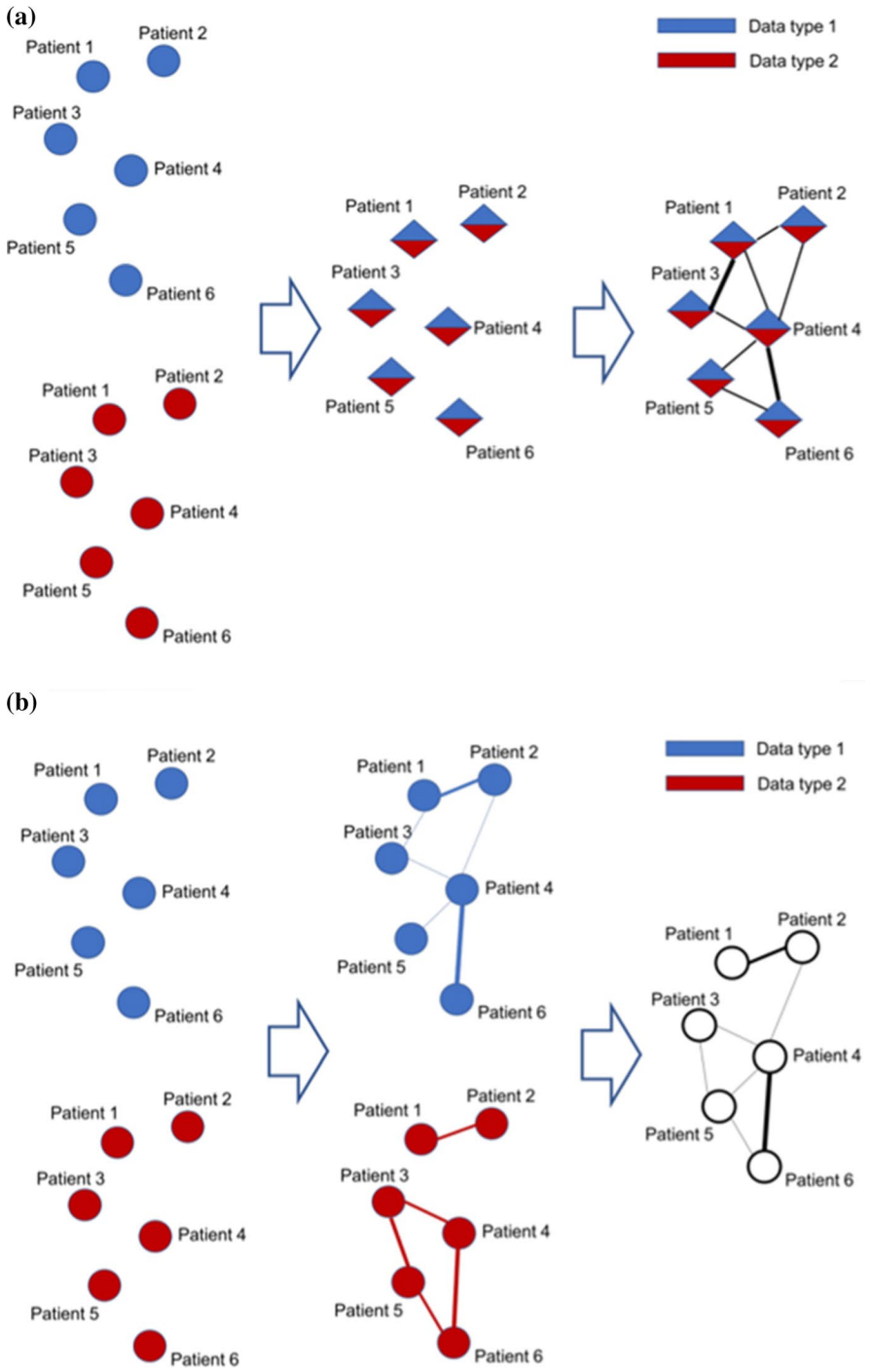


Fig. 5 **a** Early data integration strategy. **b** Late data integration strategy

relevant in terms of its aim as a contribution to the philosophy of biology, Slater's general view (see Slater, 2013) does not match the overall spirit of our analysis of similarity, insofar as it is meant as a contribution to the *metaphysics* of biology, rather than as an investigation on the actual role of similarity in scientific practice.

Summing up, what we have considered above illustrates the reasons why similarity has a role in contemporary biomedicine. More specifically, we have shown how degrees of similarity can provide the bases of classificatory procedures, and related uses, in personalized medicine. In the following section, we move back to philosophy in order to show how FMA can be applied to conceptually grasp what similarity is. We will argue that there is a role for Tversky's account in current biomedical contexts, insofar as it can foster some better conceptual understanding of a biomedical notion of similarity.

4 Tversky's Approach and Grouping Patient via Similarity

Let us begin from Tversky's general claim: "the representation of an object as a collection of features is viewed as a product of a prior process of extraction and compilation" (Tversky, 1977, 329). The main problem, in our case, is to understand what kinds of features could be associated with a given group of patients to represent it via FMA. In the psychological context, which is the standard context of application of the feature matching approach, stimuli associated with the perception of objects are the common way to extract features from a specific given domain of objects. Of course, we cannot adopt the same strategy for grouping cancer patients: it is a completely different kind of application. Why, then, should we apply the FMA in our context, and how should we do so? The basic idea is to extract, for example, the five features adopted in the Integrative Cluster approach to group breast cancer patients (Fig. 1), that is,

- Copy number driver,
- Pathology biomarker class,
- DNA architecture,
- Dominant PAM50,
- Clinical Characteristics (survival).
  adding a sixth feature, i.e.:
- TNBC (±) (Triple negative breast cancer).

The sixth feature is not part of the integrative cluster subtypes and the dominating properties listed in Table 1. It is just mentioned at the bottom of the table. We insert it to obtain a more perspicuous example of how our approach works.

Thus, for example, the 10 integrative clusters there indicated will be represented by a feature set like:

- Group (1) = {Having Chromosome 17/chromosome 20, ER+(HER2+), Simplex/firestorm, Luminal B, intermediate, TNBC−};

Table 1 Overview of the integrative cluster subtypes and the domination properties with regard to copy number driving events, biomarkers, type if DNA architecture,[46] dominant PAM50 subtype, and clinical outcome

Integrative cluster group	Copy number driver	Pathology biomarker class	DNA architecture	Dominant PAM50	Clinical characteristics (survival)
1	Chromosome 17/20	ER^+ ($HER2^+$)	Simplex/firestorm (Chromosome 17q)	Luminal B	Intermediate
2	Chromosome 11	ER^+	Firestorm (Chromosome 11q)	Luminal A and Luminal B	Poor
3	Very few	ER^+	Simplex/flat	Luminal A	Good
4	Very few	ER^+/ER^-	Sawtooth/flat	Luminal A (mixed)	Good (immune cells)
5	Chromosome 17 (HER2 gene)	$ER^-(ER^+)/(HER2^+)$	Firestorm (Chromosome 17q)	Luminal B and HER2	Extremely poor (in pre-Herceptin cohorts)
6	8p deletion	ER^+	Simplex/firestorm (Chromosome 8p/Chromosome 11q)	Luminal B	Intermediate
7	Chromosome 16	ER^+	Simplex (Chromosome 8q/Chromosome 16q)	Luminal A	Good
8	Chromosome 1, Chromosome 16	ER^+	Simplex (Chromosome 1q/Chromosome 16q)	Luminal A	Good
9	Chromosome 8/Chromosome 20	ER^+ (ER^-)	Simplex (Chromosome 8q/Chromosome 20q)	Luminal B (mixed)	Intermediate
10	Chromosome 5, Chromosome 8, Chromosome 10, Chromosome 12	TNBC	Complex/Sawtooth	Basel-like	Poor 5-year, good long-term if survival

- Group (2) = {Having Chromosome 11/chromosome 20, ER+, Firestorm, Luminal B, intermediate, TNBC+}.
- Group (3) = {Very few, ER+, Simplex/flat, Luminal A, Good, TNBC−}.
- Group (4) = {Very few, ER+/ER−, Sawtooth/flat, Luminal A, Good, TNBC+}.
- Group (5) = {Having Chromosome 17, ER-, Firestorm, Luminal B and HER2, extremely poor, TNBC+}.
- Group 6 = {8p deletion, ER+, Simplex/firestorm, Luminal B, intermediate, TNBC}.
- Group 7 = {Having Chromosome 16, ER+, Simplex, Luminal A, good, TNBC−}.
- Group 8 = {Having Chromosome 1/Chromosome 16, ER+, Simplex, Luminal A, good, TNBC+}.
- Group 9 = {Having Chromosome 8/Chromosome 20, ER+, Simplex/firestorm, Luminal B, intermediate, TNBC+}.
- Group 10 = {Having Chromosome 5/Chromosome 8/Chromosome 10/Chromosome 12, TNBC, complex, Basal like, poor 5 year, TNBC+}.

Consider, now, the conditions that Tversky's similarity should satisfy: (1) the Matching condition, the (2) the Monotonicity condition, and (3) the Independence condition.

According to the matching condition the degree of similarity (d) between two objects a and b has to be thought of as a function of three sets: (1) the set of their common features, and (2) the two sets of their distinctive features. Formally:

$$d(a, b) = F(AB, A-B, B-A)$$

Let "a" be Group (1) and "b" be Group (2). Just remember that the feature set of Group (1) is {Having Chromosome 17/chromosome 20, ER+(HER2+), Simplex/firestorm, Luminal B, intermediate, TNBC−} and the feature set of Group (2) is {Having Chromosome 11/chromosome 20, ER+, Firestorm, Luminal B, intermediate, TNBC+}. The function F is given by the set of their common features i.e. {Luminal B, intermediate} as first element; the set of the features that are in Group (1) and are not in Group (2): {Having Chromosome 17/chromosome 20, ER+(HER2+), Simplex/firestorm} and the set of features that are in Group (2) and are not in Group (1): {Having Chromosome 11/chromosome 20, ER+, Firestorm}. To resume, the similarity of Group (1), i.e. 'a' and Group (2), i.e. 'b' is given by the following function F:

$$d(a, b) = F(\{\text{Luminal B, intermediate}\}, \{\text{Having Chromosome 17/chromosome 20, ER + (HER2+), Simplex/firestorm}\}, \{\text{Having Chromosome 11/chromosome 20, ER+, Firestorm}\})$$

What does this very simple case show? It shows that in order to capture similarity of features in patient groups you should count common *and* distinctive features of the two groups. Moreover, it is easy to obtain a metric of similarity and dissimilarity among different patient groups, simply ordering the obtained results. Repeating the same operation with much more data for many patient groups,

one can get a very rich series of similarity results helping researchers to group patients in a more appropriate way.

According to the monotonicity condition monotonicity constraints similarity comparisons among objects, given a certain domain, as follows: an object a is more similar to an object b than it is to an object c iff the common features of a and c are a subset of the common features of a and b and the distinctive features of a and c are subsets of the distinctive features of those of a and b. Formally:

$$d(a, b) \geq d(a, c) \text{ whenever } (A \cap B) \text{ is subset of } (A \cap C),$$
$$(A-C) \text{ is subset of } (A-B),\ (C-A) \text{ is subset of } (B-A)$$

By the Monotonicity condition, in order to determine whether Group (1) = a is more similar to Group (2) = b than to a Group (3) = c, it is sufficient to check if the common and distinctive features of the pair (Group (1) & Group (3)) are subsets of the common and distinctive features of the pair (Group (1) & Group (2)).

The common and distinctive features of the former pair (Group (1) & Group (2)) are:

- Common: {Luminal B, intermediate, TNBC+}.
- Distinctive: {Having Chromosome 17/chromosome 20, ER+ (HER2+), Simplex/firestorm, Having Chromosome 11/chromosome 20, ER+, Firestorm};

  whereas the common and distinctive features of the latter pair (Group (1) & Group (3)) are:
- Common: {TNBC−}.
- Distinctive: {Having Chromosome 17/chromosome 20, ER+ (HER2+), Simplex/firestorm, Luminal B, intermediate, very few, ER+, Simplex/flat, Luminal A, Good}.

We could see that while the common features of the pair (Group (1) & Group (3)) are not a subset of the common features of the (Group (1) & Group (2)), the distinctive features of the former pair are a subset of the distinctive features of the latter pair. It follows that the pair (Group (1) & Group (2)) is more similar than the pair (Group (1) & Group (3)).

The above sketched second condition strengthens the idea that the FMA, applied to our topic, help us to obtain a more refined metric for patient groups. Specifically, the monotonicity condition is a sharp way to introduce, step by step a metric in the different groups comparing them two by two. To summarize: adopting FMA we have a way to conceptualize the biostatistical similarity among patient groups, in particular we have seen that the latter satisfies the condition of matching and monotonicity.

Finally, as said before, the Independence condition states the following. Assume that the pairs of objects (a, b) and (c, d) as well as the pairs (a′, b′) and (c′, d′) agree on the same two features, while the pairs (a, b) and (a′, b′) as well the pairs (c, d) and (c′, d′) agree on a third feature. If this is the case, Independence predicts that a is more similar to b than a′ to b′ if and only if c is more similar to d than c′ is to d′. Consider, again, our example in the independence case.

$$\begin{aligned}
&\text{Group 1} \cap \text{Group 2} = \text{Group 5} \cap \text{Group 5} = \text{Luminal B} = \text{X}\\
&\text{Group 7} \cap \text{Group 9} = \text{group 8} = \text{Group 3} = \text{Simplex} = \text{X}'\\
&\text{Group 1}-\text{Group 2} = \text{Group 5}-\text{Group 6} = \text{Having Chromosome 17} = \text{Y}\\
&\text{Group 7}-\text{Group 9} = \text{Group 8}-\text{Group 3} = \text{Having Chromosome 16} = \text{Y}'\\
&\text{Group 2}-\text{Group 1} = \text{Group 9}-\text{Group 7} = \text{TBNC+} = \text{Z}\\
&\text{Group 6}-\text{Group 5} = \text{Group 3}-\text{Group 8} = \text{TNBC}- = \text{Z}'
\end{aligned}$$

By independence we obtain that:

$$\begin{aligned}
\text{d(Group 1, Group 2)} &= \text{F (Group 1} \cap \text{Group 2, Group 1}-\text{Group 2,}\\
&\quad \text{Group 2}-\text{Group 1)}\\
&= \text{F(X, Y, Z)} \geq \text{F}\left(\text{X}', \text{Y}', \text{Z}'\right)\\
&= \text{F(Group 7} \cap \text{Group 9, Group 7} - \text{Group 9,}\\
&\quad \text{Group 9}-\text{Group 7)}\\
&= \text{d(Group 7, Group 9)}
\end{aligned}$$

If and only if

$$\begin{aligned}
\text{d(Group 5, Group 6)} &= \text{F (Group 5} \cap \text{Group 6, Group 5}-\text{Group 6,}\\
&\quad \text{Group 6}-\text{Group 7)}\\
&= \text{F(X, Y, Z)} \geq \text{F}\left(\text{X}', \text{Y}', \text{Z}'\right)\\
&= \text{F(Group 8} \cap \text{Group 3, Group 8}-\text{Group 3,}\\
&\quad \text{Group 3}-\text{Group 8)}\\
&= \text{d(Group 8, Group 3)}
\end{aligned}$$

Following Tversky we obtain that the ordering of the joint effect of any two components (X,Y vs X', Y') is independent of the fixed level of the third factor (Z or Z') (Tversky, 1977, 331).

Let us now wonder whether the cosine similarity (as any other particularization similarity used in cluster theory) satisfies the matching and the monotonicity condition in the FMA.[7]

Firstly, let us consider the *matching condition*. As remarked above, it is formulated in terms of the degree of similarity between two objects a and b and it is a function of three sets: the set of their common features, and the two sets of their distinctive features. In terms of the *cosine similarity*, as mentioned above, to say that two patients are totally similar is represented by two equal vectors and is equivalent to say, in terms of the FMA, that there is no difference among the features representing the set of features of *a* and those representing the set of features of *b*. Secondly, let us consider the monotonicity condition in the FMA. It implies that an object a is more similar to an object b than it is to an object c iff the common features of a and c are a subset of the common features of a and b and the distinctive features of a and

[7] Due to space limits, we do not show here that also the independence conditions is satisfied.

c are subsets of the distinctive features of those of a and b. In terms of the *cosine similarity* the condition is satisfied if and only if given three patients a, b, and c, a is more similar to b than to c if the difference between the vector angle of a with the vector angle of b is minor of that of the vector angle of a with the vector angle of c.

5 Conclusions

In the paper we have shown how philosophical reflections can provide relevant conceptual and formal tools to address some current issues in medicine more precisely and effectively. Specifically, aim of the first sections of this paper was to promote a change of attitude in the philosophy of biomedical studies, arguing for similarity of features as a way of grouping patients. In the second part of the paper we have shown how Tversky's FMA could be used to offer a philosophically detailed analysis of the notion of *similarity*.

FMA is a very simple tool, handy and useful. If markers are features, the idea to group patients on the basis of their being more or less *similar* to other groups is intuitive and immediately applicable via the model proposed. FMA simplicity and adaptability to different contexts of analysis gives us a simple way to measure similarity among patients grouping them. Considering whether and to what extent this way of conceiving similarity can be applied in other areas of the philosophy of biomedical studies can encourage some rethinking of classificatory practices and stimulate further epistemological reflections in new directions.[8]

Acknowledgements We wish to thank the referees of Erkenntnis for their comments and suggestions on preliminary versions of the paper.

References

Ali, R. H., Rueda, O., Chin, S.-F., et al. (2014). Genome-driven integrated classification of breast cancer validated in over 7,500 samples. *Genome Biology, 15*, 431.

Barilan, M. Y., Brusa, M., & Ciechanover, A. (Eds.). (2021). *Can precision medicine be personal: Can personalized medicine be precise?* Oxford: Oxford University Press.

Boniolo, G. (2017). Patchwork narratives for tumour heterogeneity. In H. Leitgeb, I. Niiniluoto, E. Sober, & P. Seppälä (Eds.), *Logic, methodology and philosophy of science—Proceedings of the 15th international congress* (pp. 311–24). London: College Publications.

Boniolo, G., & Nathan, M. J. (Eds.). (2017). *Philosophy of molecular medicine*. London: Routledge.

Brown, S.-A. (2016). Patient similarity: Emerging concepts in systems and precision medicine. *Frontiers in Physiology, 7*, 561. https://doi.org/10.3389/fphys.2016.00561

Bruna, A., Rueda, O. M., Greenwood, W., et al. (2016). A biobank of breast cancer explants with preserved intra-tumor heterogeneity to screen anticancer compounds. *Cell, 167*, 260–274.

Carnap, R. (1928). *Der logische Aufbau der Welt*. Berlin: Weltkreisverlag. Repr. Hamburg: Meiner [1961] (and later). English translation by Rolf A. George: *The Logical Construction of the World*, London: Routledge and Kegan Paul [1967].

[8] For other philosophical applications of the FMA, and, specifically, on using similarity for vagueness and identity, see Douven and Decock (2011).

Carrara, M., & Morato, V. (2011). Toward a Formal Account of Similarity and Family Resemblance for Technical Functions. In P. E. Vermaas & V. Dignum (Eds.), *Formal ontologies meet industry* (pp. 63–74). Amsterdam: IOS Press.

Curtis, C., Shah, S., Chin, S.-F., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature, 486*, 346–52.

Dai, L., Zhu, H., & Liu, D. (2020). Patient similarity: methods and applications. arXiv:2012.01976 [cs. LG].

Decock, L., & Douven, I. (2011). Similarity after Goodman. *Review of Philosophy and Psychology, 2*, 61–75.

Fuller, J., & Flores, L. J. (2015). The risk GP model: The standard model of prediction in medicine. Studies in History and Philosophy of Biological and Biomedical. *Sciences, 54*, 49–61.

Gärdenfors, P. (2004). *Conceptual spaces: The geometry of thought*. MIT Press.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman (Ed.), *Problems and projects* (pp. 437–446). Indianapolis/New York: Bobbs-Merrill.

Guinney, J., Dienstmann, R., Wang, X., et al. (2015). The consensus molecular subtypes of colorectal cancer. *Nature Medicine, 21*, 1350–1416.

Leonelli, S. (2016). *Data-centric biology: A philosophical study*. Chicago, IL: The University of Chicago Press.

Morganella, S., Alexandrov, L. B., Glodzik, D., et al. (2016). The topography of mutational processes in breast cancer genomes. *Nature Communications, 7*, 11383.

Network, C. G. A. (2015). Genomic classification of cutaneous melanoma. *Cell, 16*, 1681–1696. https://doi.org/10.1016/j.cell.2015.05.044

Nik-Zainal, S., Van Loo, P., Wedge, D. C., et al. (2012). The life history of 21 breast cancers. *Cell, 149*, 994–1007.

Nik-Zainal, S., Davies, H., Staaf, J., et al. (2016). Landscape of somatic mutations in 560 Bbreast cancer whole-genome sequences. *Nature, 534*, 47–54.

Pai, S., & Bader, G. D. (2018). Patient similarity networks for precision medicine. *Journal of Molecular Biology*. https://doi.org/10.1016/j.jmb.2018.05.037

Parimbelli, E., Marini, S., Sacchi, L., & Bellazzi, R. (2018). Patient similarity for precision medicine: A systematic review. *Journal of Biomedical Informatics*. https://doi.org/10.1016/j.jbi.2018.06.001

Pereira, B., Chin, S.-F., Rueda, O. M., et al. (2016). The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nature Communications, 7*, 11479. https://doi.org/10.1038/ncomms11479

Robertson, G. A., Kim, J., Al-Ahmadie, H., et al. (2017). Comprehensive molecular characterization of muscle-invasive bladder cancer. *Cell, 171*, 540-556.e25.

Ross-Adams, H., et al. (2015). Integration of copy number and transcriptomics provides risk stratification in prostate cancer: A discovery and validation cohort study. *eBioMedicine, 2*, 1133–1144.

Russnes, H. G., Lingjærde, O. C., Anne-LiseBørresen-Dale, A. L., Caldas, C., et al. (2017). Breast cancer molecular stratification: From intrinsic subtypes to integrative clusters. *American Journal of Pathology, 187*, 2152–2162.

Sánchez-Valle, J., et al. (2020). Interpreting molecular similarity between patients as a determinant of disease comorbidity relationships. *Nature Communications, 11*, 2854. https://doi.org/10.1038/s41467-020-16540-x

Slater, M. (2013). *Are species real? An essay in the metaphysics of science*. Basingstoke: Palgrave MacMillan.

Slater, M. (2015). Natural kindness. *The British Journal of Philosophy of Science, 66*, 375–411.

Strasser, B. (2019). *Collecting experiments: Making big data biology*. Chicago, IL: The University of Chicago Press.

Tan, P.-N., et al. (2017). *Introduction to data mining* (2nd ed.). Reading: Addison-Wesley.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*(4), 327–352.

Wallmann, C. (2017). A Bayesian solution to the conflict of narrowness and precision in direct inference. *Journal for General Philosophy of Science, 48*, 485–500.

Wallmann, C., & Williamson, J. (2017). Four approaches to the reference class problem. In G. Hofer-Szabó & L. Wroński (Eds.), *Making it formally explicit: probability, causality and indeterminism* (pp. 61–81). Dordrecht: Springer.

Weddell, N., Pajic, M., Patch, A.-M., et al. (2015). Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature, 518*(26), 495–501.

Zhu, Z., et al. (2016). Measuring patient similarities via a deep architecture with medical concept embedding. In *2016 IEEE 16th international conference on data mining*. https://doi.org/10.1109/ICDM.2016.0086.