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

Patient similarity in the era of precision medicine. A philosophical analysis

Giovanni Boniolo (1) - Raffaella Campaner (2) – Massimiliano Carrara (3)

1.

Dipartimento di Scienze Biomediche e Chirurgico Specialistiche

Università di Ferrara

Via Fossato di Mortara, 64a

44121-Ferrara (Italy)

giovanni.boniolo@unife.it

tel. +39 0532 455553

ORCID: 0000-0003-1968-4249

2.

Dept. of Philosophy and Communication Studies

University of Bologna

Via Zamboni, 38

40126 Bologna (Italy)

raffaella.campaner@unibo.it

tel. +39 051 2098329

ORCID: 0000-0003-4642-0337

3.

FISPPA Department

University of Padua

P. zza Capitaniato 3,

35139 Padova (Italy)

massimiliano.carrara@unipd.it

ORCID: 0000-0002-3509-1585

Corresponding Author: Massimiliano Carrara

Affilliation: FISPPA Department

University of Padua

P. zza Capitaniato 3,

35139 Padova (Italy)

Email: massimiliano.carrara@unipd.it

Patient similarity in the era of precision medicine. A philosophical analysis

Keywords

Precision medicine, grouping, similarity, feature matching approach

Abstract

According to N. Goodman, the Carnapian notion of similarity is useless in science and without interest for philosophy. In our paper we argue that this drastic position has to be revised, especially given the current role that the notion has in managing biomedical big data and given its 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 unprecedented 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. In the paper, we illustrate this use, by discussing two examples, and analyze it from a philosophical standpoint by resorting to A. Tversky's Features Matching Approach. In this way, we also show that the latter can allow for a better understanding of the meaning and current use of similarity in the context of biomedical big data, and that, therefore, it is of interest also for reflections in the philosophy of science, in particular in the philosophy of biomedicine.

1. Introduction

According to N. Goodman (1972), the Carnapian notion of similarity is useless in science and without interest for philosophy. After fifty years this position has to be revised in the light of personalized medicine, where big data have a central role.

Over the centuries, biomedical research has devised ways of grouping people with the same set of features to provide conceptual tools to treat individuals with the same pathological features in the same way. The advent of precision medicine has introduced an unprecedented amount of data, forcing experts to rethink classificatory practices. Patients' molecular and clinical uniqueness and the overwhelming abundance of information on their lifestyles and on the environments in which they live are dealt with new biostatistical tools, in particular cluster theory. The notion of *similarity* is central to this effort: 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.

But what do we exactly mean by *similarity* in this specific context? Actually, this question is twofold: first of all, we should understand how it is used in biomedicine (Sec. 2), and then what its philosophical counterpart is (Sec. 3). In particular, in section 2, we start by briefly illustrating the Integrative Cluster approach and the Patient Similarity approach, stressing what has motivated their introduction in the medical context . Section 3 is meant to zoom on some crucial moments in the philosophical discussion on similarity which we believe can provide some relevant insights on the topic. We will show, in particular, how the *Feature Matching Approach* (since now, FMA) proposed by A. Tversky in the late Seventies (1977) offers some 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. Instead, the aim of the paper, much more specifically, is to show the relevance in contemporary biomedicine of the notion of similarity, and to highlight, through Tversky's approach, that it is worth of deep philosophical attention

insofar as philosophy itself can allow a better conceptual understanding of what is currently going on in biomedicine.

2. Clusters via similarity

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), where these groups were based on genes, proteins, metabolites, etc. However, with the spectacular 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.[1]

[1] 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).

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 a previously established set of biomarkers, at whatever level they could be. Having the property of exhibiting such and such biomarkers amounts to be a patient that can be inserted in a given group. In other words, in this perspective every patient either belongs or does not belong to a given group, if properly diagnosed. Nevertheless, given the complexity of many molecular diseases, the enormous amount of molecular and clinical information we have (see, for example, Leonelli 2016 and 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 and Flores 2015; Wallmann 2017; Wallmann and 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 (Curtis et al. 2012; Ali et al. 2014; Bruna et al. 2016; 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 of 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 7,500 samples. As illustrated below (Fig. 1), this technique allowed them to compare the clusterisations both with other molecular characterizations (e.g., PAM50[2]) and with the clinical outcomes. They successfully showed that 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[3].

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).

[2] 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.

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

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 (Nik-Zainal et al. 2012; Nik-Zainal et al. 2016; Morganella et al. 2016).

The same idea has led to a new approach called *Patient Similarity* (Brown 2016; Pai and Bader 2018; Parimbelli et al. 2018). 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: i) 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; ii) the terrifying bulk of clinical data (diagnoses, 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 and 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.

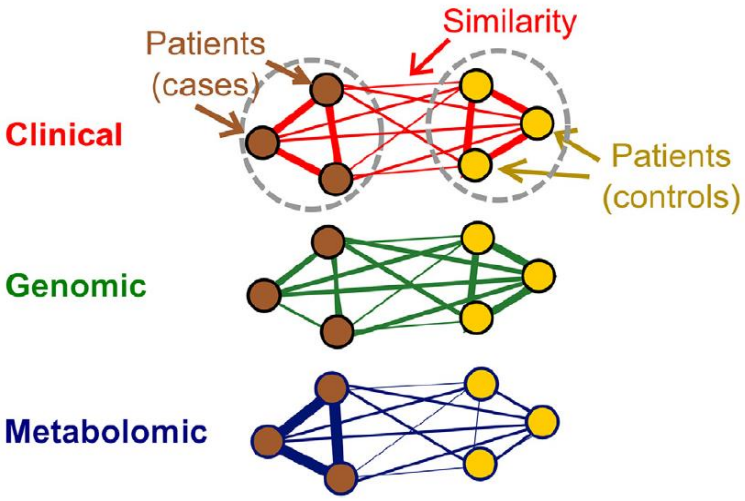

Fig. 2.

As anticipated, 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).

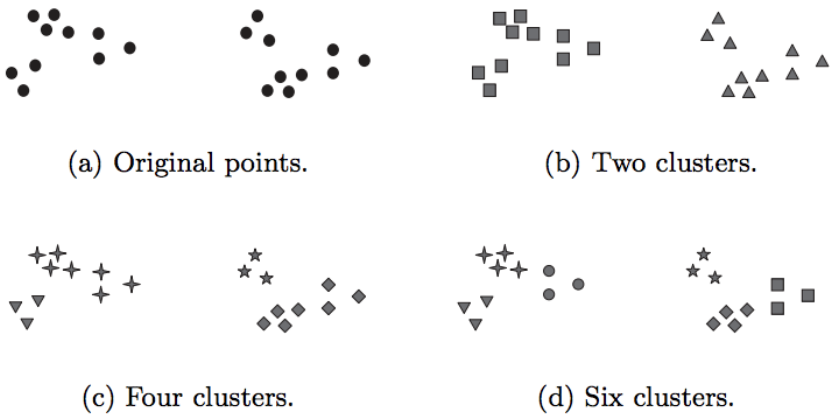

Fig. 3 Three different ways of clustering the same set of points (From Tan et al. 2017, 529).

To illustrate how cluster analysis works, and to exemplify what a distance is, let us consider an example taken from Brown (2016)[4], 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), $a \cdot b$ 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

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

patients are completely dissimilar, their vectors are opposite, thus the angle is 180° and the $\cos 180° = -1$. 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.

It is to note that both in this example and in the cases above, this kind of similarity does not work in an abstract space (as we will see, the geometrical similarity discussed by Carnap and criticized by Goodman works), but in a well-defined biomedical context given by the set of molecular and clinical information.

Summing up, what just seen 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 to philosophy in order to argue, through Tversky's account, that there is a role for it; in particular that this role allows for a better conceptual understanding of such a biomedical notion of similarity

3. Similarity and distance from a philosophical perspective

Let us enter more in depth into the notion of similarity. According to the original "geometrical model" (see, e.g., Carnap1928/1967) the notion of similarity is obtained *via* that of *similarity space*. 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.

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 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, and in relativity and quantum mechanics) adopt that geometric model. Unfortunately, as seen in the previous section, physics is not the only field where distance and similarity are used and we cannot forget the strong criticism of the geometrical model advanced by N. Goodman (1972). 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 and Douven 2011; and Carrara and 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 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.

There are two main different ways of bypassing such problems. The first one is the Tversky's FMA (Feature Matching Approach), the second one is the *Conceptual Space approach* proposed by P. 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. On the other hand, Gärdenfors' proposal is much more semantically and cognitively oriented than Tversky's and thus 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, 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.

Matching functions F are used to measure degree of similarity, that is, they are analogues to distances in the geo-metrical model. It could be shown, moreover, that the FMA solve the problems of the geometrical account, outside physics and mathematics (see Tversky, 1977), introducing some contextual elements. As it has been shown at the end of section 2, contextual elements play an important part in grouping patients. In the following section we will show how FMA can be applied to conceptually grasp what similarity is.

3.1. 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, to apply the FMA in our context, and how to 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).

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};
- Group (2) = {Having Chromosome 11 / chromosome 20, ER+, Firestorm, Luminal B, intermediate}.
- Group (3) = {Very few, ER+, Simplex/flat, Luminal A, Good}.
- Group (4) = {Very few, ER+/ER-, Sawtooth/flat, Luminal A, Good}.
- Etc.

Consider, now, the conditions that Tversky's similarity should satisfy. Actually, here, for reason of space, we just concentrate on: i) the Matching condition and the ii) the Monotonicity condition.

According to the matching condition the degree of similarity 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:

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

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} and the feature set of Group (2) is {Having Chromosome 11 / chromosome 20, ER+, Firestorm, Luminal B, intermediate}. 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({Luminal B, intermediate}, {Having Chromosome 17 / chromosome 20, ER+ (HER2+), Simplex/firestorm}, { Having Chromosome 11 / chromosome 20, ER+, Firestorm})

What does this very simple case show? Before and immediately it shown 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) ≥ d(a, c) whenever (A ∩ B) is subset of (A ∩ C), (A – C) is subset of (A – B), (C – A) 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}.
- 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: {∅}.
- 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.[5]

[5] We have omitted to show that also the independence conditions is satisfied for the sake of space.

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[6].

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 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.

4. 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 mind-changing attitude for 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

[6] As before, we leave to the reader the prove of the independence condition.

grouping them. Spreading this way of conceiving similarity in other areas of the philosophy of biomedical studies can be an epistemological turn for the whole topic.

References

Ali, Raza H., Oscar Rueda, Suet-Feung Chin, et al. 2014. "Genome-driven Integrated Classification of Breast Cancer Validated in over 7,500 Samples." *Genome Biology* 15: 431.

Boniolo, Giovanni. 2017. "Patchwork Narratives for Tumour Heterogeneity." In *Logic, Methodology and Philosophy of Science – Proceedings of the 15th International Congress*, ed. Hannes Leitgeb, Ilkka Niiniluoto, Elliot Sober, Päivi Seppälä, 311-24. London: College Publications.

Boniolo, Giovanni and Marco J. Nathan, eds. 2017. *Philosophy of Molecular Medicine*. London: Routledge.

Brown, Sherry-Ann. 2016. "Patient Similarity: Emerging Concepts in Systems and Precision Medicine." *Frontiers in Physiology* 7: 561. doi:10.3389/fphys.2016.00561

Bruna, Alejandra, Oscar M. Rueda, Wendy Greenwood, et al. (2016). "A Biobank of Breast Cancer Explants with Preserved Intra-tumor Heterogeneity to Screen Anticancer Compounds." *Cell* 167: 260–74.

Cancer Genome Atlas Network. 2015. "Genomic Classification of Cutaneous Melanoma." *Cell* 16: 1681-96. doi: 10.1016/j.cell.2015.05.044

Carnap, Rudolf. 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].

Carrara, Massimiliano and Vittorio Morato. 2011. "Toward a Formal Account of Similarity and Family Resemblance for Technical Functions." In *Formal Ontologies Meet Industry*, ed. Pieter E. Vermaas and Virginia Dignum, 63-74. Amsterdam: IOS Press.

Curtis, Christina, Sohrab Shah, Suet-Feung, et al. 2012. "The Genomic and Transcriptomic Architecture of 2,000 Breast Tumours Reveals Novel Subgroups." *Nature* Apr 18, 486: 346-52.

Decock, L. and Douven, I. 2011. Similarity After Goodman. *Rev Philos Psychol*. 2:61-75.

Fuller, Jonathan and Luis J. Flores. 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, Peter. 2004. *Conceptual Spaces: The Geometry of Thought*. Cambridge MA: MIT Press.

Goodman, Nelson. 1972. "Seven Strictures on Similarity." In *Problems and Projects*, 437-46Indianapolis/New York: Bobbs-Merrill.

Guinney, Justin, Rodrigo Dienstmann, Xin Wang. et al. 2015. "The Consensus Molecular Subtypes of Colorectal Cancer." *Nature Medicine* 21: 1350-16.

Leonelli, Sabina. 2016. *Data-Centric Biology: A Philosophical Study*. Chicago, IL: The University of Chicago Press

Morganella, Sandro, Ludmil B. Alexandrov, Dominik Glodzik, et al. 2016. "The Topography of Mutational Processes in Breast Cancer Genomes." *Nature Communications* 7:11383.

Nik-Zainal, Serena, Peter Van Loo, David C. Wedge, et al. 2012. "The Life History of 21 Breast Cancers." *Cell* 149: 994-1007.

Nik-Zainal, Serena, Helen Davies, Johan Staaf, et al. 2016. "Landscape of Somatic Mutations in 560 Bbreast Cancer Whole-Genome Sequences." *Nature* 534: 47-54.

Pai, Shraddha, Gary D. Bader. 2018. "Patient Similarity Networks for Precision Medicine." *Journal of Molecular Biology*. https://doi.org/10.1016/j.jmb.2018.05.037

Parimbelli, Elena, Simone Marini, Lucia Sacchi, Riccardo Bellazzi. 2018. "Patient Similarity for Precision Medicine: A Systematic Review." *Journal of Biomedical Informatics*, doi: https://doi.org/10.1016/j.jbi.2018.06.001

Pereira, Bernard, Suet-Feung Chin, Oscar M. Rueda, et al. (2016). "The Somatic Mutation Profiles of 2,433 Breast Cancers Refine their Genomic and Transcriptomic Landscapes." *Nature Communications* 7: 11479, doi:10.1038/ncomms11479.

Robertson, Gordon A., Jaegil Kim, Hikmat Al-Ahmadie, et al. 2017. "Comprehensive Molecular Characterization of Muscle-invasive Bladder Cancer." *Cell* 171: 540-556.e25.

Ross-Adams, Helen, et al. 2015. "Integration of Copy Number and Transcriptomics

Provides Risk Stratification in Prostate Cancer: a Discovery and Validation Cohort Study." *EBioMedicine* 2: 1133-44.

Russnes, Hege G., Ole Christian Lingjærde, Anne-Lise Børresen-Dale, A.L, Carlos Caldas, et al. 2017. "Breast Cancer Molecular Stratification: from Intrinsic Subtypes to Integrative Clusters." *American Journal of Pathology* 187: 2152-62.

Strasser, Bruno. 2019. *Collecting Experiments: Making Big Data Biology*. Chicago, IL: The University of Chicago Press

Tan, Pang-Ning, et al. 2017. *Introduction to Data Mining*. Addison-Wesley, Second Edition.

Tversky, Amos. 1977. "Features of Similarity." *Psychological Review* 84(4): 327–52.

Wallmann, Christian. 2017. "A Bayesian Solution to the Conflict of Narrowness and Precision in Direct Inference." *Journal for General Philosophy of Science* 48: 485-500.

Wallmann, Christian and Jon Williamson. 2017. "Four Approaches to the Reference Class Problem." In *Making It Formally Explicit: Probability, Causality and Indeterminism*, ed. Gábor Hofer-Szabó and Leszek Wroński, 61-81, Dordrecht: Springer.

Weddell, Nicola, Marina Pajic, Anne-Marie Patch, et al. 2015. "Whole Genomes Redefine the Mutational Landscape of Pancreatic Cancer." *Nature* 518, 26: 495–501.

Zhu, Zihao, et al. 2016. "Measuring Patient Similarities via a Deep Architecture with Medical Concept Embedding.", *2016 IEEE 16th International Conference on Data Mining*. Doi: 10.1109/ICDM.2016.0086.