Philosophy & Technology (2023) 36:8
https://doi.org/10.1007/s13347-023-00608-9

RESEARCH ARTICLE

Prediction via Similarity: Biomedical Big Data and the Case of Cancer Models

Fabio Boniolo[1] · Giovanni Boniolo[2] · Giovanni Valente[3]

Received: 29 June 2022 / Accepted: 8 January 2023 / Published online: 26 January 2023

Abstract
In recent years, the biomedical field has witnessed the emergence of novel tools and modelling techniques driven by the rise of the so-called Big Data. In this paper, we address the issue of predictability in biomedical Big Data models of cancer patients, with the aim of determining the extent to which computationally driven predictions can be implemented by medical doctors in their clinical practice. We show that for a specific class of approaches, called k-Nearest Neighbour algorithms, the ability to draw predictive inferences relies on a geometrical, or topological, notion of similarity encoded in a well-defined metric, which determines how close the characteristics of distinct patients are on average. We then discuss the conditions under which the relevant models can yield reliable and trustworthy predictive outcomes.

Keywords Biomedical Big Data · Cancer · Prediction · Models · Similarity · Distance

1 Introduction

The philosophical discussion on what models are is as old as philosophy. If we consider, for instance, the debates about the epistemological status of the mathematical representations constructed to capture the celestial bodies' motions in ancient Greece, since then, this topic has been a sort of Carsic River that, from time to time,

✉ Giovanni Boniolo
giovanni.boniolo@unife.it

Fabio Boniolo
fabio_boniolo@dfci.harvard.edu

Giovanni Valente
giovanni.valente@polimi.it

[1] Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA

[2] Dipartimento di Neuroscienze e Riabilitazione, Università di Ferrara, Ferrara, Italy

[3] Dipartimento di Matematica, Politecnico di Milano, Milan, Italy

has reappeared at the surface of the philosophers' attention (for an introduction, see Frigg & Hartmann, 2020). Nowadays, the unprecedented trust in formal modelling has been accelerated by the appearance of new computational approaches and new "objects" to model, the so-called Big Data[1]. The term Big Data modelling generally refers to a class of techniques designed to extract useful information from large datasets in a fully automated way. These modelling approaches, whose basic concepts were already introduced in the first half of the twentieth century, found fertile ground in different fields mainly thanks to the recent advancements in computational power (e.g. graphics processing units (GPUs), tensor processing units (TPU)), storage solutions, and analytical techniques. Particularly relevant for the scope of this paper, these methods have proven to be of paramount importance for the advancement of biomedicine, in part due to the amount of biomedical data being generated by new sequencing and imaging technologies, alongside the ever-growing collections of populational and individual clinical records.

This is a contemporary chapter of a long research tradition, which assumes particular importance considering the amount of money, resources, institutions, and researchers allocated to it, as well as of its positive results and fruitful prospects at both the theoretical and practical levels. As such, it provides philosophers of science and philosophers of technology an opportunity to contribute to current scientific practice by applying conceptual and analytical tools to concrete research problems, which have direct practical implications. In the biomedical field, the unprecedented computational ability of biomedical Big Data models to treat and classify huge amounts of clinical and molecular data enables scientists to infer important information about individual patients, which can then be offered to doctors for the sake of making actual clinical decisions about diagnosis, prognosis, and therapy for specific diseases. So, inasmuch as the predictions of such models are shown to be reliable, their use opens the prospect of creating interdisciplinary teams of medical doctors and data scientists, like the molecular tumour boards, that can design collaborative strategies for the effective treatment of cancer patients (Kato et al., 2020). The problem of evaluating the predictive power of biomedical Big Data models thus acquires an important ethical dimension too, as it has been emphasized by several authors (cfr. Vayena, Blasimme & Cohen, 2018; Basu, Engel-Wolf & Menzer, 2020; Heyen & Salloch, 2021;, Heilinger, 2022; Mittelstadt, 2019). In this paper, we take up the epistemological issue of whether, and to what extent, one can draw reliable clinical predictions about individual patients by means of computational techniques treating large amounts of biomedical data, with a specific focus on the application of a machine learning technique called k-Nearest Neighbours (kNN) to the classification of cancer patients.

Rather than surveying the nature and interpretation of the huge collection of data treated by such newly developed computational techniques, which one can already

[1] A review on the existing definitions of the term Big Data can be found in Chapter 6 of Durán (2018). Here, we are interested in biomedical Big Data and we adopt the definition proposed by Luo et al. (2016), according to which they are characterized by (i) high volumes, (ii) high dimensionality, (iii) high variety of types and structures, and (iv) high velocity of production.

find in Leonelli's recent account of biomedical Big Data models (2016, 2019, 2020), we provide a methodological and epistemological analysis of the structural components of the models that determine the predictive outcomes. As we are going to explain, a key aspect of the prediction process is the stratification of the large population of patients included in the initial dataset into well-defined groups, referred to as *clusters*, on the basis of a similarity relation to be intended, for reasons that we will discuss in detail, in terms of metric distance and not in terms of resemblance. In order to formulate predictions, the model assigns any individual target patient to the cluster of most similar patients, so that the typical properties of the latter can then be ascribed to the target patient too. Clearly, whether or not the thus-produced inferences can lead one to reliable biomedical predictions depends on the alleged similarities between the target patient and her relevant cluster. So, for the models to exhibit predictive power, it is of utmost importance to introduce an appropriate and well-defined measure of similarity. In fact, following the similarity conception of scientific representation (cfr. Weisberg, 2013, 2015), it is even tempting to ground predictability on the fact that a model adequately represents its target system by virtue of the similarities between their respective properties. But is that the case for the biomedical Big Data models under investigation here? That is, is there a sense in which an individual target patient is adequately represented by her similarity cluster? And how does it assure that computationally driven predictions may be trusted by medical doctors when it comes to formulating actual diagnoses, prognoses, and therapy in clinical practice?

In order to answer the questions at stake, we will proceed as follows. We begin in Section 2 by recalling some basic aspects of biomedical models, taking computational oncology as a case study. In particular, when reviewing the components of the models' formal structure that leads to the predictive outcomes, we stress that the relevant measure of similarity is intended to quantify the distance between sets of biomedical data associated with distinct patients. In the subsequent Section 3, we illustrate the use of the kNN technique by constructing a concrete example of a machine learning model for the classification of cancer subtypes. Our numerical analysis showcases how the methodological choice of different parameters, such as the number of clusters and the similarity metric, can yield rather different results. Moreover, we discuss internal statistical criteria that are standardly adopted to evaluate the worth of the resulting classifications of patients and the predictive performances of the model over the population of patients in the initial dataset. In Section 4, we address the issue of the predictability of biomedical Big Data models from a philosophical perspective. Specifically, we connect the question of whether, and how, the prediction process promises to yield empirically successful results for individual patients with the Problem of Surrogative Reasoning, namely the problem of determining to what extent a model enables one to make valid inferences about its target system. We then argue that the similarity conception of scientific representation, at least in its most elaborated version, namely the contrast approach *a là* Weisberg, does not really apply to the class of computational models we consider here. Indeed, in the contrast approach, the overall degree of similarity reflects how much the model resembles the target with respect to relevant properties; instead, the similarity metric adopted in our biomedical Big Data models is simply an index of the statistical correlations found between the properties of the target and the properties of the cluster she is assigned to. As a consequence, such a metric does not tell us whether the selected cluster resembles the target patient with respect to any single property regarded as relevant, but it only measures how close the values of all the properties of the target are *on average* to the values of the properties of the cluster she is assigned to. As we claim, this means that there is just a rather weak sense in which a similarity cluster may represent an individual patient about whom computationally driven clinical predictions are made. The upshot of our analysis is thus that predictability in biomedical Big Data models is mostly grounded on statistical correlations, and as such, it depends on the methodological choice of a specific similarity metric, which determines a classification of the population of patients satisfying internal statistical criteria for high performances of the model.

On this point, it should also be stressed that, even if one grants that biomedical Big Data models enable one to make valid inferences about the target patient, thereby offering medical doctors reliable and trustworthy predictions to implement in actual clinical practice, there remains the fact that we can only know a posteriori whether a certain prediction is correct, that is, whether what is predicted by the model is actually valid for a given patient. In fact, in the last analysis, the success or failure of a computationally driven clinical decision about diagnosis, prognosis, or therapy for a certain disease is adjudicated empirically on the basis of the follow-up of each individual patient.

2 Biomedical Modelling and Big Data

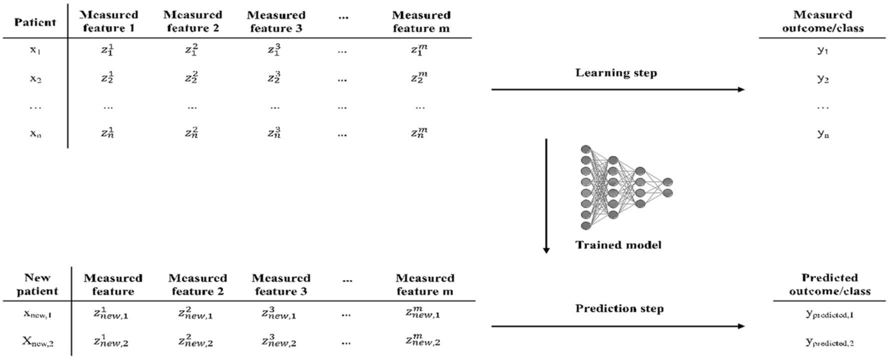
Modelling can take different forms in biology and medicine (see Benzekry, 2020). White-box models, typically referred to as mechanistic or hypothesis-driven, are based on a priori knowledge of the system under analysis and built from first principles. Systems of differential equations are a typical example, where most parameters have a purely physical and/or physiological significance, as in the case of models describing tumour growth (see Benzekry et al., 2014). On the other side of the spectrum, black-box models, typically referred to as data-driven, are purely determined by operational connections between system inputs and outputs and do not require a full understanding of the model's internal functioning. In this case, parameters do not necessarily have any physical or biological meaning. Approaches such as machine learning and deep learning are a good illustration of data-driven methods and are typically categorized as "black boxes", despite recent efforts to make them more interpretable by integrating prior knowledge (p. 3) (see Kelly et al., 2019; AlQuraishi & Sorger, 2021). In this paper, we focus on black-box models that handle cancer Big Data.

Cancer is an extremely complex and deadly disease characterized by high inter-patient and inter-tissue heterogeneity in terms of molecular and pathological features, treatment response levels, and overall survival times (see Boniolo, 2017). Despite its complexity, recent progress in the understanding of the multistep processes that transform healthy cells into neoplastic ones via the progressive accumulation of genetic alterations made it possible to identify well-defined characteristics of cancer biology (Hanahan & Weinberg, 2011). In parallel to these discoveries, advancements in sequencing technologies allowed for the molecular profiling of patients down to the level of the genome, transcriptome, epigenome, proteome, etc. (i.e. *omics*). As a result, there are now available large-scale molecular datasets spanning different human cancer types, e.g., the Cancer Genome Atlas (Tomczak, Czerwińska & Wiznerowicz, 2015) or the Pan-Cancer Analysis of Whole Genomes (Gerstung et al., 2020), as well as experimental models, as in the Genomics of Drugs Sensitivity in Cancer (Iorio et al., 2016), and the Cancer Cell Line Encyclopedia (Ghandi et al., 2019). Machine learning techniques can leverage the growing availability of biomedical data at different resolution scales (e.g., bulk tissue or single cell level), so as to disentangle the intrinsic complexity of cancer (Eraslan et al., 2019). The integration of these computational methods with wet-lab experiments and clinical information now offers the possibility of designing new medical approaches, such as precision medicine (Boniolo et al., 2021a).

Despite this, biomedical Big Data models are seldomly implemented in current medical practice. This is in part due to the long validation procedure that is required to regularly deploy them in the pre-clinical and clinical protocols (Bekisz & Geris, 2020), but it also reflects the reluctance of many physicians to adopt formal methods at the patient's bedside. In addition, crucial issues regarding the interpretability, fairness, and security of these tools are nowadays subject to discussion both at the technical and regulatory levels. The development of models that are explainable (Holzinger et al., 2017), bias-free (Chen et al., 2020), and privacy preserving (Kaissis et al., 2020) has the potential to pave the way for the systematic use of computationally driven diagnoses, prognoses, and treatment decision in new formulations of molecular tumour boards, where interdisciplinary teams of medical doctors and wet and dry-lab scientists work together to evaluate patients' molecular profiles (Kato et al., 2020). In light of the growing importance of biomedical Big Data models, it thus seems that a philosophical analysis of their predictive processes is in order.

2.1 Classification of Patients into Similarity Clusters

Let us review the formal structure of Big Data models to begin with. In this work, we address examples of supervised learning, a modelling strategy that tries to define a relation between measurements of patients' features (inputs) and patients' outcomes (outputs), as opposed to unsupervised learning, which attempts to uncover patterns of interest only by relying on patients' features. In particular, we describe classifier models, algorithms able to define associations between a set of omics and/or clinical data and a categorical outcome label (e.g. health status or disease group) for group of patients under analysis. To put it technically, let us suppose there is a population of N patients characterized by M numerical data, belonging to a discrete set of classes $C = \{0,1, \ldots, R\}$. Throughout the paper, we will adopt the following notation: each of the N patients is indicated by x_i (with $i = 1, \ldots, N$), whereas the $N \times M$ feature matrix is denoted by Z and the $N \times 1$ output matrix by Y. Specifically, each element z_i^j of the matrix Z is written with a subscript indicating the patient i (column) and a superscript j indicating the measured feature (row). In this form, any



Patient	Measured feature 1	Measured feature 2	Measured feature 3	...	Measured feature m	Measured outcome/class
x_1	z_1^1	z_1^2	z_1^3	...	z_1^m	y_1
x_2	z_2^1	z_2^2	z_2^3	...	z_2^m	y_2
...	...	...	...	...	...	...
x_n	z_n^1	z_n^2	z_n^3	...	z_n^m	y_n

New patient	Measured feature	Measured feature 2	Measured feature 3	...	Measured feature m	Predicted outcome/class
$x_{new,1}$	$z_{new,1}^1$	$z_{new,1}^2$	$z_{new,1}^3$	...	$z_{new,1}^m$	$y_{predicted,1}$
$X_{new,2}$	$z_{new,2}^1$	$z_{new,2}^2$	$z_{new,2}^3$	...	$z_{new,2}^m$	$y_{predicted,2}$

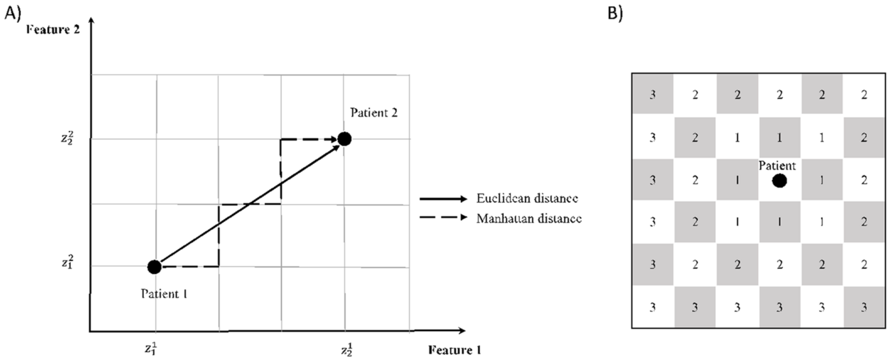
Fig. 1 Schematic view of the steps involved in building a machine learning classification model. Top layer: the training set, composed of an input matrix Z and an outcome vector Y, is used to learn a classification model able to associate each patient with the corresponding class or outcome. Bottom layer: new patients are predicted to belong to one of the existing classes by the trained model

patient x_i is characterized by a pair ($\mathbf{z_i}$, y_i), where $\mathbf{z_i} \in R^M$ is the M-dimensional feature vector whose components are the real-valued elements z_i^j for each feature j and $y_i \in \mathrm{C}$ is the class label associated to patient x_i.

The set of N pairs ($\mathbf{z_i}$, y_i), called *training set*, is used to build a function mapping each patient x_i to the corresponding class, that is, $f : R^M \rightarrow C$. The process of defining such a function f, namely the model, is referred to as *learning step* and typically consists in finding the function that leads to the optimal division of the input (or feature) space. In general, the goal of the learning step is to maximize or minimize predefined criteria, typically a cost function, evaluated on the training set[2]. Once the optimal model is trained, it is possible to use it to classify a new patient x_{N+1} that does not belong to the original training set. Based on the corresponding M-dimensional feature vector $\mathbf{z_{N+1}}$' the classifier will output a hypothesis about which one of the available classes the patient x_{N+1} might belong to. The assignment of new patients to one of these classes via the trained model is called *prediction step*. Figure 1 here below exemplifies the learning and prediction steps we have just described.

Let us stress that the computational process leading to a clinical prediction is highly dependent on how one defines the relevant classes, from here on referred to as *similarity clusters*, into which the patient population is partitioned. Indeed, these are determined by the total number of available clusters, defined *a priori*, as well as by the specific *similarity metric* one adopts in order to group the patients. We offer an overview of a class of routinely adopted similarity metrics in the next section.

[2] In classification, the learning process may consist in repeatedly modifying the model to reduce the number of times a patient x_i is associated with the wrong class (also referred to as mis-classification error).



3	2	2	2	2	2
3	2	1	1	1	2
3	2	1	Patient ●	1	2
3	2	1	1	1	2
3	2	2	2	2	2
3	3	3	3	3	3

Fig. 2 Schematic representation of the three distance metrics mentioned and used to calculate similarities between patients. **A** The solid line represents the Euclidean distance and the dashed line the Manhattan distance between two patients. **B** Representation of the Chebyshev distance, also known as "Chessboard distance"

2.2 Similarity as Distance Between Points

Formally, the notion of similarity adopted in biomedical Big Data models can be expressed in terms of the distance d of two M-dimensional feature vectors $\mathbf{z_i}$ and $\mathbf{z_{i\prime}}$, associated to the i-th and i'-th patients, respectively: that is:

$$d\left(z_i, z_{i\prime}\right) = L\left(z_i - z_{i\prime}\right).$$

where L is a *distance function*, which can be further specified by introducing different metrics. One of the most common metrics is the Minkowski distance, defined as:

$$d\left(z_i, z_{i\prime}\right) \ = \sum\nolimits_{j=1}^{m}\left(\left|z_i^j - z_{i\prime}^j\right|\right)^{1/p}.$$

This formula encodes the idea that, in order to compute the distance between the two M-dimensional vectors associated to patients' sets of data x_i and $x_{i'}$, respectively, one first calculates the difference between the values of the generic j-th component of the vectors (namely the generic j-th feature of each patient); then, one elevates it to the exponent $1/p$; and finally one sums over all the $j = 1, \ldots, M$ components. Note that the exponent $1/p$ can be set to define different types of distances depending on the value of p. Typical values given to p are 1, 2, or $p \rightarrow \infty$, resulting in formulations of Manhattan distance, Euclidean distance, and Chebyshev distance, respectively.

Below we give an intuitive depiction of how the distance between two 2-dimensional points associated with patient 1, i.e. $(z_1^1,\ z_1^2)$, and patient 2, i.e. (z_2^1, z_2^2), varies when taking $p = 1$ and $p = 2$ (Fig. 2A) or $p \rightarrow \infty$ (Fig. 2B).

Let us stress that this is a *geometric*, or *topological*, notion of similarity, whereby the chosen metric quantifies how close the biomedical properties of different patients are. As such, it measures the degrees of statistical correlations between the patients' data, described as elements of a metrical space[3]. The geometrical notion of similarity has been widely adopted in statistics, in particular cluster theory, where it is defined in terms of measurement, distance, and metric, and it is now at the core of basic classification techniques in biomedical Big Data modelling (see Boniolo, Campaner & Carrara, 2021b for a more detailed discussion). To illustrate how cluster analysis works, let us consider an example, taken from Brown (2016), which can help one grasp the sense in which two sets of biomedical data can be regarded as being similar. In the proposed formalization, any patient is associated to a vector defined in a multidimensional metric space, where each dimension is in turn associated, e.g. to a specific biomedical datum. Given two vectors expressing the sets of data relative to two distinct patients, their degree of similarity (i.e. their distance in the statistical space) can be given, for example, by the so-called cosine similarity:

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

where a and b are the two vectors, $a \bullet b$ is their scalar product, $\|a\|$ is the module of the vector a, and a_i its i-component (representing a molecular or clinical datum). Hence, the distance, that is, the similarity, is given in terms of the cosine of the angle between the two vectors. This means that, if the two sets of data are completely dissimilar, their vectors are opposite; thus, the angle is 180°, and the $cos\ 180^\circ = -1$. Instead, if the two sets of data are totally similar, they are associated to two equal vectors; thus, the angle between them is 0° and $cos\ 0^\circ = 1$. In general, given a benchmark set of data a associated with a patient, one captures its similarity with any set b associated with another patient by calculating $d(a,b)$. What is important to note is that the resulting degrees of similarity are evaluated just in terms of how close the respective sets of clinical features of the two patients are when they are taken *on average*.

When the comparison class grows, namely the number of patients is much higher than two, as in the models of biomedical Big Data under consideration, the cosine distance measure used in Brown's example becomes less adequate, and hence, one ought to adopt the variants of the Minkowski distance we listed above for the sake of enacting a partition of the patient's population into similarity clusters. As we will show by way of a concrete cancer model in the next section, the choice of one similarity measure over the others can determine rather different outcome predictions.

3 A Machine Learning Classification Model: Predicting Cancer Subtypes

Given the complexity reached nowadays by computational models (especially in the machine learning and artificial intelligence space), whose description is out of the scope of the present paper, we focus on a basic classification technique called

[3] Historically, this concept emerged in geometry around 1906, thanks to the work of the French mathematicians René Fréchet (even though the name "metrical space" is due to Felix Hausdorff), who introduced the notion of distance between two points in a topological space.

k-Nearest Neighbours (kNN). Nevertheless, we expect the concepts discussed here to hold for any supervised technique based on a well-defined similarity metric. In kNN, the training phase simply consists in storing the input matrix Z and the output vector Y in memory (lazy learning). In the prediction step, the unseen data, associated with the $N+1$-th patient, are used to identify the "closest", or "most similar", patients in the input matrix Z based on a predefined similarity metric d. Once the population of patients has been subdivided into similarity clusters, the class relevant for the new observation is selected through a *majority vote*, that is, as a function of the most frequent cluster label. In this framework, k is a parameter indicating the number of patients in the training set, i.e. the size of the neighbourhood, that contributes to assigning the new patient to one of the clusters (in the simplest, trivial case in which $k=1$, each new patient is assigned to the cluster of its closest neighbour). In concrete applications, determining the optimal combination of parameters, namely the value of k together with the similarity metric d, is one of the main computational challenges that have been shown to heavily drive the prediction performances of the learning algorithm. In order to better illustrate this, we illustrate the implementation of a kNN model for the classification of different cancer types based on a specific type of omics data, i.e., gene expression measured via RNA-sequencing (RNA-seq)[4].

Omics data are commonly represented in tabular format, having genes, transcripts, or genomic locations as rows and patients as columns, so that each element is an integer representing the number of RNA fragments measured for each gene in each patient at the time of the sampling. We used data from three different cancer types: lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), and oesophageal carcinoma (ESCA). The datasets, already normalized and log-transformed and containing data from 576, 553, and 196 patients, respectively, were obtained from the Cancer Genome Atlas (TCGA, see Tomczak, 2015) via the Xena Browser (Goldman et al., 2020). The data from the three cancer types were down-sampled to have comparable numbers of patients and subsequently concatenated after finding the common genes in the three datasets. The final matrix, used as input for the analysis, had dimensions 20.530 genes × 588 samples. The response vector was created by concatenating the cancer type labels, i.e. LUAD, LUSC, and ESCA, corresponding to the 588 patients.

As routine in machine learning applications, the dataset was further divided into a training set (containing 80% of the samples) and a validation set (containing the remaining 20%). This step, referred to as data splitting, enables one to consider the validation set as an external dataset and thus to test the generalization capabilities of the model calibrated on the training set. After splitting, genes (i.e. variables) in the training set showing zero variance across samples were removed, and the remaining ones were standardized using z-score normalization (i.e. all the features were

[4] RNA-seq is part of the so-called next-generation sequencing technologies, which allows for the quantification of the amount of RNA in a sample, or in a cell, at a given moment. Typical applications exploit RNA-seq data to study phenomena such as alternative splicing, changes in gene expression in different groups or treatment cohorts, and biomarker identification.

brought to a common scale with mean zero and standard deviation one). Principal component analysis (PCA), a linear dimensionality reduction technique typically used in these applications (see Huong & Holmes, 2019), was then performed to transform the original ~20k-dimensional dataset into a lower-dimensional dataset while retaining the original information content. The validation set was preprocessed by first removing the zero-variance features identified in the training set, followed by standardization and dimensionality reduction (using the model fitted on the training data).

The kNN model was implemented using the Scikit-learn python module (see Pedregosa et al., 2011). In order to evaluate the effect of the choice of the parameters on the final prediction, we selected three values for the size of the neighbourhood, that is, $k=2$, $k=22$, and $k=200$, and three different distance metrics, namely the Euclidean, Manhattan, and Chebyshev ones. As a result, we obtained 9 different combinations that we could compare. The whole process, comprising splitting, preprocessing, model calibration, and prediction, was repeated 100 times to obtain random samples for training and testing, thereby extracting robust estimates of model performance. The ability of the models to predict the correct tumour types in the test set was then evaluated based on *internal statistical criteria*, namely *accuracy*, *precision*, and *recall*, calculated as the average of the 100 model instances obtained for each combination of parameters. Let us describe the results we obtained in greater detail.

Accuracy is used to evaluate the overall performance of the model. In the 3-class classification problem described above, this results in calculating the proportion of samples correctly assigned to the corresponding cancer type. As represented in Fig. 3, the choice of different values of k and of distance metrics may heavily influence the overall performance of the models (with extremes that go from 0.9 down to 0.64) that reaches particularly disappointing results when Manhattan distance is combined with high values of k.

While accuracy is an indicator of the general performance of a model, precision and recall are more sensitive to the performances of the models for the individual cancer types. In a multi-class setting, *precision* (also defined as *positive predictive value*) is calculated for each class and represents the proportion of samples correctly assigned to a cancer type out of all the samples predicted to be part of that cancer type (Fig. 4A). On the other hand, *recall* (also referred to as *sensitivity*) is the proportion of samples correctly predicted to belong to a class out of the actual number of samples belonging to that type (Fig. 4B). Once again, different combinations lead to different performances of the models when predicting patients belonging to different cancer types. In particular, models based on Manhattan distance and medium/high values of k appear to reach low values (0.62) of precision for lung squamous cell carcinoma (LUSC) and low values of recall (0.21). When analysed together, the three scores highlight the generally low performances of the models based on Manhattan distance, especially when being in combination with high values of k.

To summarize our findings, we have shown how the choice of the relevant parameters, especially the size of closest neighbours and the similarity metric, can influence the final performance of computational techniques in terms of accuracy,

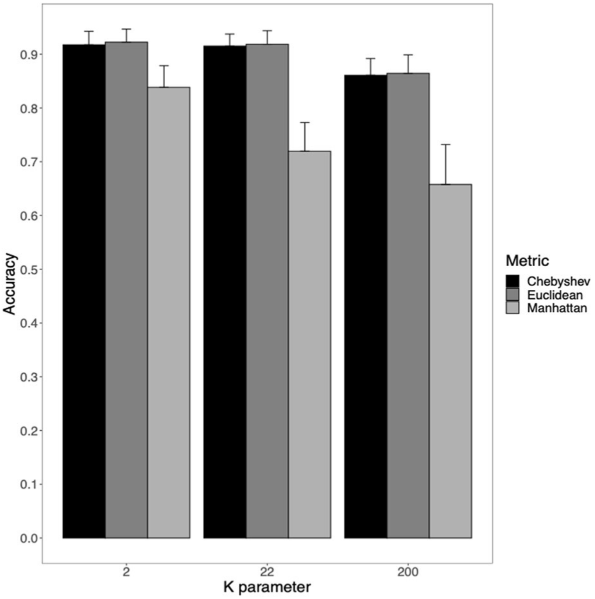


Fig. 3 Average accuracy values and standard deviations of the 9 classification models. Average accuracy for different distance metrics is represented as a function of the parameter *k*, defining the size of the neighbourhood in kNN

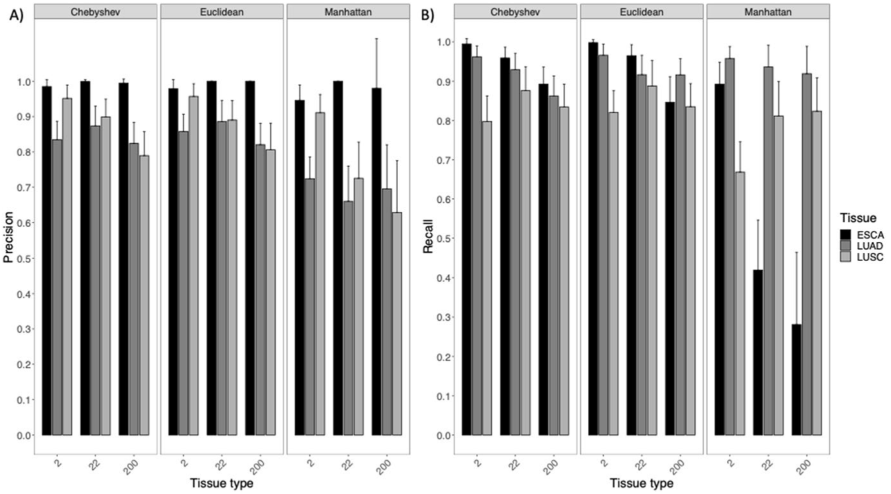


Fig. 4 Precision and recall for the 9 classification models. **A** Average precision values and standard deviations calculated for the 3 different classes as a function of the *k* parameter. Models based on the Manhattan metric tend to reach low values of precision for LUSC samples. **B** Average recall values and standard deviations visualized as for the precision ones. The lowest recall values are measured for ESCA samples when models are based on the Manhattan distance

precision, and recall, even in the framework of a simple classification problem[5]. As we will see in the next section, while the fact that the stratification of the population of patients into similarity clusters is sensitive to such parameters is a well-known

[5] In a typical setting, the choice of parameters is highly dataset-dependent and might vary based on the overall goals behind the calibration of a model (see Prasath et al., 2017).

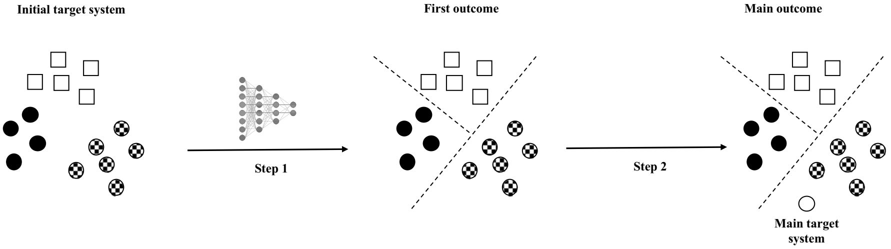


Fig. 5 The modelling and prediction process in the case of Big Data models

fact in the field, it assumes high relevance for the issue of predictability in biomedical Big Data models, at least for the class of models under investigation here.

4 Prediction via Similarity

In order to set the stage for our analysis of predictability in biomedical Big Data models, let us begin by offering a schematic reconstruction of the modelling and predictive process under consideration, which is illustrated in Fig. 5. In doing so, we will also fix the terminology we adopt in the ensuing philosophical analysis.

Before the training step, the cohort of N patients under analysis, namely the *initial target system*, is constituted by the results of numerous measurements of the relevant clinical/molecular features for each patient, displayed in the matrix Z. In the first step of the process, the adjustments of a set of parameters leads to the division of the original data space into different regions, which discriminate clusters of patients based on their mutual similarity and dissimilarity (first outcome). There are two crucial aspects worth highlighting in the process leading to the first outcome. For one, the definition of the clusters is highly sensitive to the methodological choice of a particular *similarity metric*. Moreover, the assessment of the validity of the first outcome, permitted by the machine learning model, is given in terms of *internal statistical criteria*, specifically accuracy, precision, and sensitivity, which we discussed in the previous section. Maximizing such criteria assures that the model has high predictive performances. Arguably, when the resulting clusterisation of the dataset is validated, the model is expected to give an adequate representation of the initial target system, in that it shows how the population of patients can be subdivided into actual groups based on their relevant features.

In the second step, the model is used to make predictions, such as diagnosis, course of the pathology, or optimal treatment strategies, about an individual patient, what we refer to as the *main target system*. In order to produce the intended predictions, the new patient x_{N+1} is assigned to one of the clusters partitioning the population of N patients in the initial dataset, based again on the similarity metric adopted in the partition in the previous step. One can thereby formulate conjectures about the pathological status of the target patient, make estimations of the future course of the pathology, or gather indications for a possible therapeutical pathway, on the grounds of the typical behaviour of the patients grouped in the relevant cluster. That is the main outcome of the process, which completes the prediction step. In terms of the technical terminology introduced in Section 2.1, a clinical prediction enabled by the kNN modelling technique thus takes the following form: the target patient x_{N+1} is predicted to have a certain value in the output set Y, which corresponds to the most frequent value of the model function f occurring for the cluster of k patients that are most similar to x_{N+1} among all the N patients in the initial dataset.

The philosophical issue at stake is to establish whether, and to what extent, the predictions of biomedical Big Data models can be taken seriously enough to be implemented by clinicians in medical practice. On this point, there are two important questions that are relevant. That is: To what extent are such computationally driven biomedical predictions reliable? And are they actually valid for the target patient?

The answer to the second question is entirely empirical. In fact, whether or not the decision on the prognosis, diagnosis, and treatment of a patient turns out to be correct can only be verified *a posteriori* by checking the effects it produced on the actual patient. Crudely put, given a cancer patient, if her physiological conditions do not improve, then we know that the prescribed therapy has failed, meaning that the biomedical prediction upon which the therapy relied was wrong or not good enough; else, if the patient's physiological conditions improve, then one has reasons to think that the prediction was empirically compatible with the outcome. The first question, instead, has to do with the theoretical structure of the model. Surely, it also has an empirical component, in that the prediction is made by comparing the clinical data of the target patient with those of the other patients in the initial dataset. However, such a prediction strongly depends on the formal mechanism by which the new patient is assigned to the relevant cluster. In particular, it is determined by the choice of the similarity metric, which selects the patients that are most similar to the target patient. So, the driver of the predictive inference is supposed to be the similarity between the new patient x_{N+1} and her relevant cluster. What one would like to know, then, is in what sense, if any, positing a well-defined similarity metric is sufficient to assure the trustworthiness of computationally driven clinical decisions.

4.1 Surrogative Reasoning and Representation via Geometrical Similarity

To begin with, let us stress that the issue of predictability in biomedical Big Data models is closely related to the outstanding *Problem of Surrogative Reasoning*, which philosophers of science have discussed for quite some time, often in connection with the ongoing debate on *Scientific Representation*. In the literature on scientific models, surrogative reasoning enables one to draw inferences about the target system based on some features of the model, which thus plays the role of the "surrogate system", as Swoyer (1991) pointed out. As such, it is a tool to generate hypotheses, as well as to formulate predictions, about the target. In the context of biomedical Big Data models, predictability relies on a form of surrogative reasoning, whereby one draws predictive inferences about the individual target patient on the basis of her similarity with a selected cluster of other patients. Arguably, an adequate model would entitle clinicians to ascribe to the new patient the typical features of the similarity cluster she is assigned to. Whether or not surrogative inferences of this kind are licensed is an epistemological issue that has practical implications in medical practice: for one would like to properly understand in what sense biomedical Big Data models can aid physicians to decide about diagnosis, prognosis, and treatment of individual patients. The Problem of Surrogative Reasoning thus becomes particularly pressing in a clinical context.

Resolving the Problem of Surrogative Reasoning requires one to establish under what conditions surrogative inferences are licensed. Let us refer to such inferences as *valid surrogative inferences*, by adopting a terminology proposed by Contessa (2007) that echoes our own remarks on the two above questions concerning biomedical predictions: more to the point, valid inferences further prove *sound* if their alleged conclusions about the target system turn out to be empirically valid. It seems reasonable to hold that a sufficient condition for the validity of surrogative inference is that the model gives an adequate representation of the target system[6]: indeed, one may then expect that features of the former are also shared by the latter. Although there are various different theories of scientific representation, the so-called similarity conception deserves particular attention in the connection with the models we are dealing with here. The underlying idea is that a model adequately represents a target just in case they are similar to each other in relevant respects to the appropriate degrees. Frigg and Nguyen (2020) review the main points of criticism raised against this view. Besides the fact that judging what properties count as relevant and the extent to which they should be similar across the model and the target is a context-dependent matter (Teller, 2001), one major problem already mentioned by Goodman (1972) concerns the logical structure of the similarity relation. In particular, the relation of being similar in relevant respects appears to be symmetrical (indeed, the model is similar to the target just as the target is similar to the model), whereas the notion of representation is not symmetrical at all: for it is only the model that is supposed to represent the target and not vice versa. On the positive side, though, Frigg and Nguyen concede that the similarity conception of scientific representation promises to offer an elegant account of surrogative reasoning, at least as long as the validity of surrogative inferences is grounded on a well-defined measure of similarity between the model and the target.

Contrast approaches to the similarity conception are designed to avoid the above objections. In fact, they rely on a similarity measure that is not symmetrical. In particular, in Weisberg's (2013, 2015) weighted feature matching account of model-world similarity, similarity is a measure of the relevant properties that are shared by the model and the target, which can be even assigned different weights depending

[6] Note that for Contessa, it is also a necessary condition: in fact, he claims that the model enables one to make valid surrogative inferences if and only if it gives an "epistemic representation" of the target. However, other authors disagree on this point. For instance, according to Suarez (2004) inferential conception of scientific representation, the ability to draw surrogative inferences does not presuppose that the model adequately represents the target (in fact, not even if one insists on the additional condition of denotation, that is, that the model's representational force points to the target).

on the context[7]. So, if computationally-driven predictions in biomedical Big Data models could be shown to fit into Weisberg's account, we would have an instance of valid surrogative inferences grounded on representation by similarity. However, we contend that there are cogent reasons why in the models we consider here the similarity measure cannot possibly correspond to Weisberg's measure. In fact, the relevant notion of similarity is rooted in the alternative geometrical, or topological, approach we described in Section 2.2.

Let us note that Frigg and Nguyen (2020) identified two ways in which a model *M* and a target *T* can be related by similarity: "[i]f the similarity between M and T is based on shared properties, then a property found in M would also have to be present in T ; and if the similarity holds between properties themselves, then T would have to instantiate properties similar to M.". In the first case, it is the fact that the model and target have a common set of properties that make them similar, whereas in the second case, some of the properties of the model are regarded as similar, rather than identical, to the corresponding properties of the target[8]. Weisberg's weighted feature-matching account of model-world similarity is an instance of the first case since it relies on the fact that the model and target share a set of identical properties. However, as objected by Parker (2015) and Khosrowi (2020), in scientific modelling, it is customary to assume that models have properties that are just similar to those of the target system, as prescribed in the second case. That is clearly what happens in biomedical Big Data models, wherein one does not presuppose that the target patient has exactly the same properties as the other patients in her cluster. Rather, similarity metrics, like the ones based on the Manhattan, the Euclidean, and the Chebyshev distance, just measure the distance between the values of a large number *M* of biomedical features of distinct patients: the target patient can thus be assigned to the group comprising the closest, i.e. most similar, patients in the initial dataset, but in general, the values of their respective features are different. It follows that Weisberg's account of representation via similarity cannot apply to the class of biomedical Big Data models under investigation here, such as the k-Nearest Neighbours model.

More to the point, underlying the contrast approach, there is a notion of representation as "resemblance", whereby the model is supposed to resemble its target system in the sense that it is similar or just like the latter with respect to the relevant properties they share. Instead, the geometric approach to similarity aims only to quantify the degrees of statistical correlations observed between the respective properties of distinct patients. Specifically, the kNN algorithm associates any new target

[7] For completeness, let us state Weisberg's measure of similarity here below:

$$S(m,t)=\frac{\theta f\left(M_a\cap T_a\right)+\rho f\left(M_m\cap T_m\right)}{\theta f\left(M_a\cap T_a\right)+\rho f\left(M_m\cap T_m\right)+\alpha f\left(M_a-T_a\right)+\beta f\left(M_m-T_m\right)+\gamma f\left(T_a-M_a\right)+\delta f\left(T_m-M_m\right)}.$$

Accordingly, as he explains, "m and t correspond to the model and target, M and T the sets of features possessed by the model and target that are members of the feature set D, f a weighting function, and the additional Greek letters correspond to weights on each term" (2015, 300).

[8] The difference between these two ways of relating model and target by similarity actually traces back to Niiniluoto's (1988) distinction between "Partial Identity" and "Likeliness".

patient $x_{N+1}$with the cluster of patients x_i whose features are the closest on average to the features of x_{N+1}. By adopting a given two-place similarity measure $d(\bullet,\ \bullet)$, one can rank the patients' data in the dataset on the basis of how similar they are to x_{N+1}, depending on the comparative values $d(z_{N+1}, z_i)$, for all M-dimensional feature vectors z_i with $i=1,\ \ldots,\ N$. Once the parameter k is fixed, one can then determine the similarity cluster for x_{N+1}, which includes the k patients' data with the lowest values for the metric d. Differently from Weisberg's contrast approach, the metric does not take into account the extent to which a similarity cluster may resemble the target with respect to any specific feature. Indeed, even small values of the distance $d(z_{N+1}, z_i)$ do not indicate that the empirical value of some particular omics feature (e.g. the j-th component of the corresponding feature vector) for patient x_{N+1} is close to the empirical value of the same omics feature (e.g. the j-th component of the corresponding feature vector) for patient x_i. In the geometrical approach, one just measures the average difference in values over all components of the respective M-dimensional feature vectors z_{N+1} and z_i.

Since Weisberg's weighted feature matching account of model-world similarity does not apply here, one is left with a more impoverished sense in which there is representation via similarity. For one, the similarity distance $d(z_{N+1}, z_i)$ is clearly symmetrical with respect to patient x_{N+1} and patient x_i, thereby lending itself to one of the standard objections to the similarity conception of scientific representation. Moreover, in the case of the kNN, how exactly the model partitions the dataset into clusters depends on the choice of methodological parameters, most notably the number k of patients in the relevant cluster as well as a particular similarity measure $d(\bullet, \bullet)$. As we showed in Section 3, changes in such parameters will result in different classifications into clusters, which may as well yield different biomedical predictions about the target patient. That is, even by keeping k fixed, the target patient may as well be represented by a different cluster depending on whether one chooses the Manhattan distance, or the Euclidean distance, or the Chebyshev distance (or any other admissible measure of similarity).

The only way to constrain the choice of similarity metric seems to be by imposing internal statistical criteria, in the first step of the prediction process, namely accuracy, precision, and sensitivity, so as to determine the optimal combination of parameters. The sense in which the target patient x_{N+1} is adequately represented by the similarity cluster she is assigned thus bears on how well the resulting classification into clusters represents the population of N patients in the initial dataset.

To underscore the weaker sense of representation retained by biomedical Big Data models, let us point out a curious aspect of the prediction process that can occur in the basic kNN models. There, the criterion to draw surrogative inferences depends on *the majority vote rule*. So, if the model function f defined over the M-dimensional real space R^M of the omics features of the patients maps the majority of the patients in the cluster onto a certain predictive outcome y in the output class Y, then one can infer that the target patient x_{N+1} has the same predictive outcome y, possibly with a probability given by the proportion of members of the cluster having this outcome. For the sake of simplicity, let us suppose that $k=3$, so that the cluster comprises the three patients in the dataset that are most similar to x_{n+1} according to a given metric d, say x_1, x_5, and x_{10}, to which the model assigns the outcome values

$f(Z_1) = y_1, f(z_5) = y_2$, and $f(z_{10}) = y_2$, respectively. Now, if it turns out that $d(z_{N+1}, z_1) < d(z_{N+1}, z_5) < d(z_{N+1}, z_{10})$, then x_1 should be identified as the most similar patient to x_{N+1}. Yet, based on the majority vote procedure, one ascribes the predictive outcome y_2 to x_{N+1} assigned to the other patients x_5 and x_{10}, instead of the outcome y_1 assigned to patient x_1. This shows that, perhaps counter-intuitively, the biomedical prediction about the target system is not even made on the basis of the patient that best represents her.

In the last analysis, we have shown that predictability in biomedical Big Data models is grounded on a geometrical account of similarity as distance between the clinical data of distinct patients. Surrogative inferences about the target patient are drawn on the basis of her similarity to a cluster of other patients in the initial dataset. In kNN models, such predictive inferences are supposed to be valid if the similarity metric (together with the number of patients in the cluster) determines a partition of the dataset into clusters that satisfy internal statistical criteria for the optimal performances of the model, thereby giving an adequate representation of the population of patients. In the face of such a weak notion of representation driving surrogative inferences about the target patient, one may thus wonder whether computationally driven biomedical predictions based on statistical correlations are effectively reliable and trustworthy for medical doctors to be implemented in clinical practice. Given that the empirical validation of clinical decisions on diagnosis, prognosis, and therapy of individual patients can only be made a *posteriori* by checking the follow-up of each patient, we suggest that a strategy to evaluate the trustworthiness of computationally driven biomedical predictions is to collect and examine future clinical data about the target patients themselves: accordingly, the modelling processes yielding trustworthy predictions are those permitting successful predictive outcomes that turn out to be robust under empirical validations.

5 Conclusions: Statistical Correlations as a Basis for Predictions

In this paper, we addressed the issue of predictability in biomedical Big Data models. That is an epistemological issue with direct practical and ethical implications: in fact, if computationally driven biomedical predictions prove to be reliable and trustworthy, they could be effectively implemented by medical doctors in actual clinical practice, so as to formulate diagnosis, prognosis, and therapy for individual patients. We argued that for the class of Big Data models, we considered, i.e. kNN classification models, predictability relies on the choice of a similarity measure, which determines a partition into clusters of the population of patients in the initial dataset: biomedical inferences about a new target patient are then drawn on the basis of the typical features of the cluster she is assigned to, namely the one comprising the most similar patients to her. Importantly, the relevant notion of similarity has a topological meaning, since it is introduced by a metric establishing the distance between the biomedical features of distinct patients taken on average.

We then contrasted this geometric account of similarity with Weisberg's similarity conception of scientific representation, wherein the similarity relation is intended to have a different meaning, that is, that a model resembles its target system in relevant respects. The upshot is that, in biomedical Big Data models, a target patient may be represented by her similarity cluster only in a weak sense, that is, merely on the basis of observed statistical correlations between omics features. To be sure, the emphasis on the role of statistical correlations in Big Data models is not new: in particular, Pietsch (2016, 2021) suggested that predictions should be made just when the observed correlations retain causal significance. However, in our view, causality is not even a necessary condition for predictability, as is demonstrated by the fact that it plays no role in the cancer model we constructed. Instead, computationally driven biomedical predictions can be regarded as reliable inasmuch as the clusterisation determined by the model satisfies internal statistical criteria, thereby giving an adequate representation of how the population in the initial dataset is classified into well-defined groups of patients. Whether clinical predictions are empirically valid or not, though, can be decided only a posteriori by evaluating the follow-up of each individual patient about whom one formulates diagnosis, prognosis, and therapy for specific diseases.

Acknowledgements We would like to thank Sabina Leonelli and Patricia Palacios who read previous versions of the manuscript and gave us useful suggestions. Moreover, we also wish to thank two anonymous referees for their comments on our paper and for allowing us to improve it.

Author Contributions FB, GB, and GV have equally contributed to the research and writing of the paper. All authors read and approved the final manuscript.

Funding Open access funding provided by Università degli Studi di Ferrara within the CRUI-CARE Agreement.

Data Availability Not applicable

Declarations

Ethics Approval and Consent to Participate Not applicable

Consent for Publication Not applicable

Competing Interests Not applicable

References

AlQuraishi, L. M., & Sorger, P. K. (2021). Differentiable biology: using deep learning for biophysics-based and data-driven modeling of molecular mechanisms. *Nature Methods, 18*(10), 1169–1180.

Basu, T., Engel-Wolf, S., & Menzer, O. (2020). The ethics of machine learning in medical sciences: Where do we stand today? *Indian Journal of Dermatology, 65*, 358–364.

Bekisz, S., & Geris, L. (2020). Cancer modeling: From mechanistic to data-driven approaches, and from fundamental insights to clinical applications. *Journal of Computational Science, 46*, 101198. https://doi.org/10.1016/j.jocs.2020.101198

Benzekry, S. (2020). Artificial Intelligence and mechanistic modeling for clinical decision making in oncology. *Clinical Pharmacology and Therapeutics, 108*, 471–486.

Benzekry, S., et al. (2014). Classical mathematical models for description and prediction of experimental tumor growth. *Plos Computational Biology, 10*(8), e1003800. https://doi.org/10.1371/journal.pcbi.1003800

Boniolo, F., et al. (2021a). Artificial intelligence in early drug discovery enabling precision medicine. *Expert Opinion on Drug Discovery, 2*, 1–17.

Boniolo, G. (2017). Patchwork narratives for tumour heterogeneity. In H. Leitgeb, I. Niiniluoto, E. Sober, P. Seppälä, Logic, Methodology and Philosophy of Science – Proceedings of the 15th International Congress, College Publications, pp. 311-324.

Boniolo, G., Campaner, R., & Carrara, M. (2021b). Patient similarity in the era of precision medicine: A philosophical analysis. *Erkentnis*, 1–22. https://doi.org/10.1007/s10670-021-00483-w

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

Chen, I. Y., et al. (2020). Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science, 4*.

Contessa, G. (2007). Scientific representation, interpretation, and surrogative reasoning. *Philosophy of Science, 74*(1), 48–68. https://doi.org/10.1086/519478

Durán, J. M. (2018). *Computer Simulations in Science and Engineering*. Springer.

Eraslan, G., et al. (2019). Deep learning: New computational modelling techniques for genomics. *Nature Reviews Genetics, 20*, 389–403.

Frigg, R., Hartmann, S. (2020). Models in science. The Stanford Encyclopedia of Philosophy.

Frigg, R., & Nguyen, J. (2020). *Modelling nature: an opinionated introduction to scientific representation*. Springer.

Gerstung, M., et al. (2020). The evolutionary history of 2,658 cancers. *Nature, 578*, 122–128.

Ghandi, M., et al. (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature, 569*, 503–508. https://doi.org/10.1038/s41586-019-1186-3

Goldman, M. J., et al. (2020). Visualizing and interpreting cancer genomics data via the Xena platform. *Nature Biotechnology, 38*, 675–678.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman (Ed.), *Problems and Projects* (pp. 437–446). Bobs-Merril.

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell, 4*, 646–674.

Heilinger, J. C. (2022). The ethics of AI ethics. A constructive critique. *Philosophy & Technology, 35*, 61.

Heyen, N. B., & Salloch, S. (2021). The ethics of machine learning-based clinical decision support: An analysis through the lens of professionalisation theory. *BMC Medical Ethics, 22*, 112.

Holzinger, A. et al. (2017). What do we need to build explainable AI systems for the medical domain?. arXiv preprint arXiv:1712.09923.

Huong, N.L, Holmes, S. (2019). Ten quick tips for effective dimensionality reduction. *PLoS Computational Biology* 15, 6. https://doi.org/10.1371/journal.pcbi.1006907

Iorio, F., et al. (2016). A landscape of pharmacogenomic interactions in cancer. *Cell, 166*, 740–754.

Kaissis, G. A., et al. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence, 2*, 305–311.

Kato, S., et al. (2020). Real-world data from a molecular tumor board demonstrates improved outcomes with a precision N-of-One strategy. *Nature Communications, 11*, 1–9.

Kelly, C. J., et al. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine, 17*, 195. https://doi.org/10.1186/s12916-019-1426-2

Khosrowi, D. (2020). Getting serious about shared features. *The British Journal for the Philosophy of Science, 71*(2), 523–546. https://doi.org/10.1093/bjps/axy029

Leonelli, S. (2016). *Data-Centric Biology: a Philosophical Study*. Chicago University Press.

Leonelli, S. (2019). What distinguishes data from models? *European Journal for Philosophy of Science, 9*. https://doi.org/10.1007/s13194-018-0246-0

Leonelli, S. (2020). Scientific research and big data. *The Stanford Encyclopedia of Philosophy* https://plato.stanford.edu/archives/sum2020/entries/science-big-data/

Luo, J., et al. (2016). Big data application in biomedical research and health care: A literature review. *Biomedical Informatics Insights, 8*, BII-S31559.

Mittelstadt, B. (2019). The ethics of biomedical ‘Big Data’ analytics. *Philosophy & Technology, 32*, 17–21.

Parker, W. S. (2015). Getting serious about similarity. *Biology and Philosophy, 30*(2), 267–276.

Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research, 12*, 2825–2830.

Pietsch, W. (2016). The causal nature of modeling with Big Data. *Philosophy & Technology, 29*, 137–171.

Pietsch, W. (2021). *Big Data*. Cambridge University Press.

Prasath, V. B. et al. (2017). Distance and similarity measures effect on the performance of K-Nearest Neighbor classifier--A review. arXiv preprint arXiv:1708.04321.

Suarez, M. (2004). Deflationary representation, inference, and practice. *Studies in History and Philosophy of Science, 49*, 36–47.

Swoyer, C. (1991). Structural representation and surrogative reasoning. *Synthese, 87*(3), 449–508. https://doi.org/10.1007/BF00499820

Teller, P. (2001). Twilight of the perfect model model. *Erkenntnis, 55*(3), 393–415. https://doi.org/10.1023/A:1013349314515

Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). *The Cancer Genome Atlas (TCGA):* An immeasurable source of knowledge. *Contemporary Oncology, 19*(1A), A68.

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine, 15*(11), e1002689.

Weisberg, M. (2013). *Simulation and Similarity*. Oxford University Press.

Weisberg, M. (2015). Response to critics. Biology and Philosophy symposium on simulation and similarity: Using models to understand the world. *Biology & Philosophy, 30*, 299–310. https://doi.org/10.1007/s10539-015-9475-1

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.