Frontiers in Nanoscience and Nanotechnology

Survey ISSN: 2397-6527

The use of data mining and machine learning in nanomedicine: A survey

Andrea Dimitri* and Maurizio Talamo

Management engineering Department, Tor Vergata University, Via della Ricerca Scientifica and INUIT Foundation Tor Vergata University, via dell'Archiginnasio snc, 00133 Rome, Italy

Abstract

The introduction of minute particles into the body to treat disease or repair damage may sound like something out of science fiction, but recent advances in nanomedicine leave researchers increasingly hopeful about the viability of medicinal opportunities on the nanoscale. Quantitative methods based on data mining and machine learning techniques have to strengthen this new branch of medicine. In this paper we analyze applications of supervised and unsupervised learning technique to better understand hot issues in nanomedicine regarding nanoparticles, molecules and cells behaviours and relationships.

Introduction

The goal of this paper is to explore the techniques of data mining and machine learning, used in nanomedicine researches, considering example articles in literature (see bibliography). Literature delivers a variety of definitions of nanotechnology which all have their advantages and limitations. While the prefix nano is often used just for a description of the length scale between 0.1 to 100 nanometer (1 nm=109 m), this size regime does not imply per se a new quality of materials or devices. A more specific definition has been given in 2000 by the US National Nanotechnology Initiative: "Nanotechnology is concerned with materials and systems whose structures and components exhibit novel and significantly improved physical, chemical and biological properties, phenomena and processes due to their nanoscale size". With the reduction of magnitude, apparently different, and qualitatively new and advantageous properties emerge from the respective material at the nanometer scale. Nanomedicine means essentially applying nanotechnology to medicine. Nanomedicine has not to be confused with nanobiotechnology: the former focus on the applications of nanotechnology concepts to medical applications, while the latter encloses all basic research at a nanoscopic level on biological systems, e.g. investigations on plants [1].

New nanomaterials are rapidly being developed for a wide range of biomedical applications. However, despite the breadth of applications centered on human health, relatively little is known about fundamental nanomaterial-biological interactions, therefore, even less is known about how to design nanoparticles to exhibit a desired effect in living organisms. A rational approach has to be employed to direct the safe development of novel nanotechnologies and to pro-vide accurate predictions of nanomaterial-biological interactions. Such an approach will inevitably require data mining and computer simulation to identify the most important design parameters in an almost infinite combinatorial space of nanoparticle formulations from global research efforts in nanoscience and nanotechnology. Thus, informatics has been largely recognized as an essential element of nanotech-nology and a rational approach to employ weight-of-the-evidence strategies that ensure its safe development. In fact, informatics methods that enable collaboration, data sharing, unambiguous representation of data, semantic (meaningful) search and integration of data, in nanomedicine, are impor-tant driving forces for successful mining of knowledge from existing nanotechnology and biomedical data resources. This knowledge is essential for the rational design and safe application of nanoparticle formulations in nanomedicine.

The steady growth of the field of nanomedicine has led to the development of nanoinformatics and subsequently the use of data mining and machine learning to develop methods to predict both functional and structural properties of nanoparticles and then to refine medical treatments. Research articles focusing on this area of research appear to be published in a wide variety of journals. The methods reported attempt to predict a large number of nanoparticle properties including, cellular uptake, cytotoxicity, molecular loading, molecular release, nanoparticle adherence, nanopar-ticle size, and polydispersity.

Correspondence to: Dimitri A, Management engineering Department, Tor Vergata University, Via della Ricerca Scientifica and INUIT Foundation Tor Vergata University, via dell'Archiginnasio snc, 00133 Rome, Italy, Email: andrea.dimitri@uniroma2.it

Special Issue: Nanotechnology: Challenges and Perspectives in Medicine

Dr. Federica Valentini
Department of Sciences and Chemical Technologies
Tor Vergata University
Italy

Maurizio Talamo
Professor
Department of Enterprise Engineering
Italy

Key words: *nanomedicine, nanoparticles, cellular uptake, cytotoxicity, QSAR, supervised vs unsupervised learning, linear regression, cluster analysis, bayesian approach, advanced neural networks, dimensional reduction*

Received: September 30, 2018; **Accepted:** October 14, 2018; **Published:** October 17, 2018

But also methodological as-pects and available tools and consequences of their choose and calibration are object of quantitative statistical analysis.

In this context, a set of issues have been addressed using data mining and machine learning techniques. Following, in a first step, the classification of [2], the first issue is cellular uptake. Nanoparticles are used to treat diseases. Cellular uptake has to be understood (related) with refer to the nanoparticle molecule, its quantitative and qualitative characteristics and with refer to the cells where nanoparticles are used. In this case cells are classified in target and non-target cells, with the goal to prevent negative effects of a treatment. The knowledge acquired using data mining in cellular uptake can help the preparation phases of a treatment, the selection of the target cells and to prevent negative effects of the treatment.

The second critical issue is cytotoxicity. The use of nanoparticles offers new potentialities and a new prospective in diagnostic and therapeutic applications. In spite of the rapid progress and the initial acceptation of the nanobiotech-nology, all negative effects on the human health related to the continued exposition to heterogeneous concentration of nano-materials have not yet been completely understood. In the nanometric dimension, particles could result in different physic-chemical characteristics if compared with particles of higher dimension but with the same composition. Some characteristics of these treatments like the area of the involved superficies, the high chemical reactivity, and many other variables like the capability to pass through the cellular membrane, can be viewed positively but, at the same time, could represent a problem if applied to untargeted cells and tissues.. To evaluate bad effects of treatment, the citotoxicity of nanoparticles is the main element of attention. Citotoxicity is quality of being tossic of a nanoparticle for a cell. Data analysis regarding citotoxicity varies for type of particles, delivered molecules, target and untargeted cells, methodologies used for the treatment. Classifications and relationships between the above elements are the goals of them.

Molecular loading level concerns the capability of nanoparticles to be a versatile molecular loading platform used as delivery device. For example, solid lipid nanopar-ticles (SLNs) are nanoparticulate drug delivery systems, which are considered very tempting as drug carriers es-pecially for lipophilic drugs. SLNs have the ability to protect these drugs and control their release. Moreover, they are used as innovative colloidal drug carriers for topical applications, especially in virtue of their interaction with the stratum corneum (SC) and other skin layers [3]. Accordingly, modelling drug-loading in these important nanoparticulate matrices was warranted in order to save researchers and formulators the efforts and time spent in the wet-laboratory experimentation and to provide them with initial and accurate estimations of the fate of their investigated drugs in the selected carrier (As a rule of thumb, better loading indicates better in-vitro and in-vivo stability of the prepared nanoparticles). Another related sce-nario: polydopamine nanoparticles could serve as a versatile molecular loading platform for magnetic resonance imaging guided combined chemo and photothermal therapy with minimal side effects, showing great potential for cancer theranostics. In these entire contexts two aspects have to be measured: the effects on target cells and systems (tissues or others) and negative effects on untargeted cells. Molecular release is another property of nanoparticles. For example the controlled release of an anticancer agent from drug nanoparticles could be achieved by varying the linker length of dimeric compounds as prodrug. The cytotoxicity of the cancer cells was closely related to the release rate of drug compounds. This strategy will lead to the establishment of the novel delivery system using drug nanoparticles. Nanopar-ticle adherence: In these classes the goal is to select the right nanoparticles for every group of cells object of a treatment. Morphologic characteristic of nanoparticles (for example size) and the adherence with target and untarget cells or systems (for example cell membrane) influences the cellular uptake efficiency. As can be seen above, the size of nanoparticles is a very important molecular property that can affect their usefulness in nanomedicine. For instance, the size of a nanoparticle has been found to be a very important factor determining the fate of the nanoparticle in vivo. Optimization of size is also important for the design and development of nanoparticles used to treat a variety of tumors, because the size of the nanoparticles affects their permeability and retention. Nanoparticles size can change based upon solution conditions, manufacturing, drug loading, and release of drugs [2]. Last, polidyspersity is a well known issue in particles, regarding the composition of the set of them used in a tretatment. For example, lipid-based drug delivery systems, or lipidic carriers, are being extensively employed to enhance the bioavailability of poorly-soluble drugs. They have the ability to incorporate both lipophilic and hydrophilic molecules and protecting them against degradation in vitro and in vivo. There is a number of physical attributes of lipid-based nanocarriers that determine their safety, stability, efficacy, as well as their in vitro and in vivo behaviour. These include average particle size/diameter and the polydispersity index (PDI), which is an indication of their quality with respect to the size distribution. The suitability of nanocarrier formulations for a particular route of drug administration depends on their average diameter, PDI and size stability, among other parameters. Controlling and validating these parameters are of key importance for the effective clinical applications of nanocarrier formulations.

Another class of issues is the analysis of the methodology behind a treatment. In [4], authors analyze Scanning Probe Microscopy where are not clear the parameters behind the building of an optimal probe for successive nanotechnology elaborations. The acquisition of large hyperspectral data sets bring on new challenges in data storage, dimensionality reduction, visualization and interpretation [1-5].

Finally, two things complete the scenarios that lead to the composition of the dataset used in the analysis. In all these issues often is used the so-called QSAR approach, where QSAR staying for quantitative structure-activity relationship. Typically, the first step to achieve a QSAR study is the identification of a molecule and to define the reference identification. Second step: identify the explanatory variable, what we want to understand. Third step is the definition of the quantitative characteristics of the molecule and last is the model selection and the calculus [5]. Second thing: one of the goal behind the use of nanotechnologies for medicine is to use in vitro and in silicio experiments to refine and improve successive in vivo treatments.

The goal of this introduction, a tentative to list issues of data mining and machine learning in nanomedicine, is an hard task. We started reporting the definition of nanotechnology first and nanomedicine after. Behind this decision there was a way to communicate this hardness. The radius of the problems is wide and its not easy to enclose them in a finite envelop. If we accept the above list of issues, there are about four phases that could be used to resume every research: the first one is the analysis of the tool (or the set of them) used for a treatment, the second one is the variables selection. Third: the output of the model that could be a classification, a relationship or a prediction. Fourth: the variables used in the model. The majority of models try to find a relationship between molecules of nanoparticles and human cells. Researches could

also be distinguished for types of nanoparticles, variables selected to characterize the nanoparticle or a class of variables. For example the QSAR methodology selects quantitative and morphologic variables about a nanoparticle. The same speech can be done for cells and tissues. Here the distinction is between target and untargeted cells. After, the selection regards special types of cells and special measurements on them.

Pre-processing of data, variables selection, model definition

We said that the goal of this paper is not a deeper investi-gation on the nature of nanomedicine. Nevertheless the full understanding of the recalled definitions of nanotechnology and nanomedicine give us a good starting point to introduce data mining and machine learning in nanomedicine. The first aspect to underline is the wide scope of the definitions above. They are suitable for a large class of issues, regarding many molecular concepts, many cellular concepts and many types of applications. This is the first obstacle in comparing applications of data mining in nanomedicine. Rarely more than three papers reported the use of data mining and machine learning techniques for the same context in terms of nanoparticles, cells and target of the research. And when this happen, used dataset have a low number of elements and the same variable has been collected with diverging techniques for every paper. For example, one of the most important reasearch question is understanding and prediction of cellular uptake of nanoparticles used to treat a desease (for example cancer). At first glance, the enquiry seems to be simple: understanding the uptake of nanoparticles in target and non target cells to better prepare a treatment and avoid misbehaviours. This is not the real case: all papers concluded the method of choice for the quantification of nanoparticles (NPs) uptake mainly depends on the research question, the available analytical devices as well as on the type of NPs

of interest. As of that, it is not possible to recommend one specific technique that could be used for quantification of all the different NPs types which exist nowadays. As well known from the convincing evidence from the literature, physicochemical properties of NPs such as their size, shape, core material and surface functionalization have a strong impact on NP cellular interaction including uptake, intracellular fate and induction of cell response but also require very different analytical methods [3].All these aspects impact in the task of comparing studies and then assessing data mining and learning techniques for the same inquiry.

Very little work has been reported on the use of data mining and machine learning methods to predict cytotoxicity of organic nanoparticles. One potential reason for this is the lack of databases or publications analyzing the cytotoxicity caused by a variety of organic nanoparticles. Another reason is the variability of biological models in different laborato-ries. Factors such as potential aggregation of nanoparticles, variations in the media used, cell origin and passage, among others further contribute to variability in the data obtained. Another commonality observed among many of the research articles presented in this review is the limited sample size related to the high-dimensionality of the dataset used for analysis. Several consequences can arise due to lack of data, including overfitting, difficulty in demonstrating reliability, generalizability, and applicability of the predictive models to other nanoparticles, and class imbalance. Validation of a predictive model can be problematic when the sample size is limited and the variables representing those samples have high-dimensionality. A simplistic and common method for overcoming the issue of high-dimensionality of a dataset is to utilize variable (feature) selection to reduce the number of variables analyzed in the predictive model. Other automated approaches, with defined properties, are PCA and bayesian variables selection, below recalled.

Class imbalance is a challenging problem for the data mining community. It occurs when the samples representing one class is much lower than those representing other classes. The simplest way to overcome this issue is to ensure that there is a balanced representation of the members of each class present in the dataset, but this is a significant challenge in nanoinformatics as the lack of large well curate datasets seriously limits the amount, quality, and variety of data available. We analyze this task in standardization. Class im-balance has to be considered also in model selection: linear regression based models will suffer whereas unsupervised cluster analysis methodologies have to be preferred, speaking about cancer nanomedicine, notes that considerable technological success has been achieved in the field, but the main obstacles to nanomedicine becoming a new paradigm in cancer therapy stem from the complexities and heterogeneity of tumour biology, an incomplete understand-ing of nanobio interactions and the challenges regarding chemistry, manufacturing and controls required for clinical translation and commercialization.

Three methodologies/tools to front these classes of problems are: standardization, variable reduction and meta-analysis [6-10].

Data sharing and standardization

The need of more coherence and structured paths that lead the conduct of nanotechnology research has been previously suggested. Paper [6] deals with this problem. It confirms that the lack of common reporting standards and non-uniformity of information reported are significant barriers to data shar-ing and re-use. And it suggests that the Nanotechnology Working Group (Nano WG) of the US National Institutes of Health National Cancer Informatics Program (NCIP) has been focused on addressing these issues.

The Nano WG, which includes representatives from over 20 organizations including government agencies, academia, industry, standards organizations and alliances - has de-veloped ISA-TAB-Nano4,5, a general framework for rep-resenting and integrating diverse types of data related to the description and characterization of nanomaterials using spreadsheet or TAB-delimited files. Nanoparticle character-ization studies have many of the same challenges as omics-based (metabolomics, genomics and functional genomics, for example) assays such as high data volume and variety, multiple experimental end points, and complex protocols and study samples. Therefore, ISA-TAB-Nano is based on the ISA-TAB format developed and used by the ISA Commons to share datasets in a diverse set of life sciences and in particular omics data. The ISA-TAB-Nano extension uses the three primary files of ISA-TAB investigation, study and assay (ISA) files as well as an additional file called the material file.

The paper underline that delivering a community-driven specification for nanotechnology data is the first phase of a proposed process. To be useful, ISA-TAB-Nano must be implemented in tools and by databases to assist researchers in reporting their data while shielding them from unneces-sary complexity. And new tools are to be developed.

To address the challenges of data sharing, efforts are un-derway, also by the National Cancer Institute (NCI) and collaborating organizations to define standards for repre-senting nanoparticles and their characterizations via the establishment of a Nanotechnology Working Group (Nano WG) and the development of nanoinformatics

resources, such as the cancer Nanotechnology Laboratory web por-tal (caNanoLab). The goal of caNanoLab is to provide a resource where primary nanotechnology research data are no longer disparate islands affiliated with their originators, but standardized and shared across the scientific and clinical community [7]. The paper confirms that progress in the field has been impeded by the lack of a knowledge-management infrastructure as well as the lack of standards to describe the complexity of nanoparticles and their highly diverse nature. Other institutions support standardization. The Nanoma-terial Biological Interactions (NBI) knowledgebase was de-veloped in 2008 to directly address the need for a compara-tive, integrative database information system, driven by the desire to promote the safe development of nanomaterials and nanotechnologies. NBI knowledgebase is functionally comprised of two components: a nanomaterial library and analysis tools. One of the focuses of NBI understands the risks associated with nanomaterial exposure. The Molecular Imaging and Contrast Agent Database (MICAD) is an on-line resource that provides information about imaging and contrast agents used with in vitro, animal or human studies that have been published in peer-reviewed scientific journals. MICAD also provides information about nanoparticles that are intended for use as imaging and contrast agents. InterNano is a web portal designed for sharing informa-tion on advances in applications, devices, metrology, and nanomaterials, in order to facilitate the commercial devel-opment and/or marketable applications of nanotechnology. InterNano gathers information from multiple sources, adds original commentaries on these sources, and provides news highlights, feature articles and assessments of the current state of practice in nanomanufacturing. A longer list of Organizations devoted to facilitate data sharing and stan-dardization can be found here [11-13] (Table 1).

Meta-analysis

Currently, most of the nanomedicine data are found in textual sources such as journal articles. It is inherently dif-ficult to process information from textual data sources. This difficulty is further exacerbated by several factors that are specific to the field of nanomedicine. First, the nanomedicine field lacks standard terminologies for describing elements of nanomedicine research and, in particular, does not have a systematic nomenclature for naming nanoparticle-based formulations. Second, there are substantial gaps in nanomedicine physical, chemical, and biological data due to inadequate characterization of nanomaterials. These gaps are directly related to the absence of minimum information standards for nanomedicine data reporting to ensure data quality, data completeness and data reliability in journal articles and databases. Third, the nanomedicine field suffers from data irreproducibility due to the poor availability of standardized protocols for preparation and characterization of nanomaterials. Fourth, the lack of standardized formats for exchanging data hinders efficient sharing and transfer of information about the chemical composition, synthesis, characterization, toxicity, and safe handling of nanomateri-als. Finally, there is a lack of raw data (versus analyzed data) which is necessary for renormalizing data from different sources for consistency for the successive analysis [13].

Meta-analysis could be one way to solve, at least partially, these problems. The goal of meta-analysis is to combine results from multiple scientific studies in a stronger one. Two aspects have to be considerated: the general lack of data in nanomedicine is not easy to solve with this approach. Second: doubts about this family of methodologies were recently expressed and regard the mechanisms of paper selection and, generally speaking, paper publication.

In 2016 shows a meta-analysis of pre-clinical studies comparing tumor accumulation of cancer nanomedicines. They used SciFinder and Google Scholar databases and the search term nanoparticle delivery, and identified 224 manuscripts. The data from 117 reports were tabulated and standardized to calculate the DE (nanoparticle delivery efficiency) based on a non-compartmental linear trapezoidal analysis model, a method to interpolate them. [9] shows a meta-analysis of clinical and preclinical studies comparing the anticancer efficacy of liposomal versus conventional non-liposomal doxorubicin. Results are controversial, demon-strating enhanced therapeutic efficacy of liposomal vs. free doxorubicin in pre-clinical studies but not in clinical studies [14-20].

Table 1. Summary of data available in caNanoLab

Nanomaterial type	Nanomaterial entities	Description
Biopolymer	13	A polymer fomed by a living organism
Carbon block	2	A material produced by the incomplete combustion of carbon-rich organic fuels in low oxygen condition
Carbon nanotubc	50	A nanotube comprised of one or more graphite sheers (graphene) of hexagonal arrays of carbon rolled into seamless cylinders with capped ends
Carbon particle	1	An amorphous nanopowder formed by laser techniques
Dendrimer	74	A polymeric molecule that has a highly branched, three-dimensional tree-like architecture, synthesized with monomers where shells of branched molecules are added in discrete steps to a central core
Emulsion	88	A colloid in which both liquids are immiscible with each other
Fullerence	16	Any cagelike, hollow molecule composed of hexagonal and/or pentagonal groups of carbon atoms
Liposome	34	A supramolecular structure which is a closed vesicle that forms on hydration of dry phospolipids above its transition temperature
Metal oxide	186	A nanomaterial composed of metal oxide
Metal	132	A nanomaterial composed of metal
Meralloid	36	A nanomaterial with properties between a metal or non-metal
Nanohorn	7	A single-walled carbon nanostructure with anirregular horn like shape
Nanorod	33	A nanoscale rod composed of either metallic or semiconductor material or a mixture of both
Nanoshell	1	A three-dimensional nanostructure that is composed of a spherical core surrounded by a few nanometers in thickness. If the shell is made of metal, then it is called a metallic nanoshell
Polymer	188	A nanomaterial composed of single or multiple rnonomers
Quantum dot	73	A nanometer site fragment of semiconductor material. whose excitons (electron-hole pairs) are confined in three spatial dimensions
Silica	43	A nanomaterial composed of a silicon oxide

Principal component analysis and dimensional reduction

One of the main problems in NPs dataset is their high dimensionality compared with the low number of statistical units. Principal component analysis is a method that can be used to reduce dimensionality of the dataset. The set of original variables, often with an high number of members with repetitions and redundancy, is replace with a lower number of new uncorrelated variables that can be view as variable profiles, latent in the original dataset. PCA components need to be interpreted: analysis is not automated and the new profiles need to be discovered considering relationships between old and new variables [14]. studied relationships between size and surface coverage change in nanomedicines and their behaviour in vivo. A principal component analysis (PCA) was carried out. A total of 11 variables measured for the different types of NPs were used for this analysis (discriminant for NP characterization in this study. Then, the method was able to provide direct information on the capacity of a NM surface to repel a series of proteins. Then, we can distinguish according to PC2 a difference between NPs A1, A2, R2 and R1, R3. This component is mainly driven by the size of NP and the tendency to adsorb aprotinin, which is globally weak but surprisingly a bit higher for NPs R1 and R3. These data showed that size and surface curvature are not sufficient by themselves to explain adsorption of proteins. However, the macromolecular grafting appears to be decisive for such interactions. Only 5 statistical units were used for ACP, that is atypical for this methodology. Units are types and not examples of types [4]. use PCA to reduce the high number of variables coming from a scanning probe. Authors criticize the fact that PCA components lack well-defined physical meaning and propose a decomposition based on Bayesian inference to front the problem. [15] make a risk analysis of cytotoxicity for two classes of nanomaterials. Twentyfour measurements from five different TiO2 features, and 18 measurements from six different ZnO features, were obtained from the experimental data sets. A QSAR analysis was done: a study to determine if some of the physical properties of TiO2 or ZnO strongly relate to each other. To this end, principal component analyses (PCA) and correlation analyses was used, which involved all of the input variables and the response variable. Because the different datatypes were measured in different units and show significant differences in their variances, the data were normalized by standardizing the individual variances. Subse-quent PCA showed that the first three principal components explain more than 90

As viewed above, Bayesian methods are an alternative to PCA for variables selection in a model. The issue is particularly studied in QSAR, where many quantitative characteristics could be collected for a nanoparticle but these are strongly related and only a subset could be used for predictions. [16] used a Bayesian approach to carry out this task. It employed a specialized sparse Bayesian feature reduction method based on an EM algorithm with a Laplacian prior to select a small set of the most relevant descriptors for modelling the response variables from a much larger pool of possibilities.

Linear models

Typically the nature of data variables and the hypothesis about relationships linking them determines the selection of the linear model. Classical multilinear regression model indicates that all variables are numerical and the explained variable (for example the level of cellular uptake) is a linear combination of the predictor variables (measurement about the particles). Linear discriminant analysis (LDA), normal discriminant analysis (NDA), or discriminant function anal-ysis are a generalization of Fisher's linear discriminant, a method used in statistics, pattern recognition and machine learning to find a linear combination of features that charac-terizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier. LDA is closely related to analysis of variance (ANOVA) and regression analysis, which also attempt to express one dependent variable as a linear combination of other fea-tures or measurements. However, ANOVA uses categorical independent variables and a continuous dependent variable, whereas discriminant analysis has continuous independent variables and a categorical dependent variable (i.e. the class label). Logistic regression and probit regression are more similar to LDA than ANOVA is, as they also explain a categorical variable by the values of continuous independent variables. These other methods are preferable in applications where it is not reasonable to assume that the independent variables are normally distributed, which is a fundamental assumption of the LDA method.

Using the logistic regression model, Liu et al. [20] classified cytotoxicity by examining the plasma membrane integrity when transformed bronchial epithelial cells (BEAS-2B) were submitted to nine different metal oxide nanoparticles. For the development of the model, a set of 10 nanoparticle descriptors was selected and measured experimentally. These descriptors include simple basic descriptors (number of oxygen atoms in the metal oxide, number of metal atoms in the metal oxide, metal oxide molecular weight, and atomic mass of the metal), stability and reactivity information (atomization energy), element group properties (periodic table group and period of the metal in the metal oxide), simple geometric descriptor (nanoparticle primary size), and indicators of surface charge and aggregation tendency (zeta potential and isoelectric point). Additional experimental con-ditions were taken into account by adding measured values for a set of four different concentrations as input parameters of the model. The paper does not explicitly indicates the number of samples used in the dataset, however art. 2 retraces that 83 samples were used. Considering logistic regression characteristics, All possible combinations of the descriptors and concentrations were analyzed for their nano-QSAR models, which generated accuracies ranging from 93 to 100Often papers get more than one model to consider fitting between data and the model and then to use the better one. [15] studies how selected metal oxide nanomaterial structural features perturb cytoplasmic leakage. They made a comparative study of two linear based mathematical models: multivariate linear regression and linear discriminant anal-yses (LDA) classification. The performance was evaluated with respect to the ability to predict a specific cellular response, i.e., lactate dehydrogenase (LDH) release after exposure to metal oxide nanomaterials. The multivariate linear regression represents the class of models attempting to detect trends in data, while LDA classification is an example of a method that aims at separating data based on the different levels of biological response.

Although well known and easy to understand in their results, linear models not always give good results. Linear models suffer limits listed above regarding the high dimensionality of the dataset compared with the low number of samples. Overfitting, multicollinearity, low levels of R often affect studies and often are not fully considered.

Clustering

Linear models applied with non quantitative variables act as classifier, but are considered supervised classifiers given the linear nature of the model in its causal variables. Unsupervised classifier use the concept of distance defined considering available explicative

variables and made a set of clusters. Next step is to understand the effective clustering profile and use it to learn.

In pattern recognition, the k-nearest neighbours algorithm (k-NN) is a non-parametric method used for classification. The input consists of a parameter k and a training set of objects for which the class or object property value is known. Given a distance definition and a new object, the k closest elements in the training set are considered (k-set) and the new object is included in the most frequent class in the k-set.

Liu et al. [21] used a variety of algorithms (IBK) in an effort to predict embryonic zebrafish post-fertilization toxic effects of several nanoparticles, including metal nanoparticles, dendrimers, metal oxides, and polymeric materials. IBK is a K-nearest neighbour predictor that assigns an input to the most common output label among its K nearest neighbours.

A nave Bayes (NB) classifier is an important classifier for data mining and applied in many real world classification problems because of its high classification performance. It is a simple probabilistic classifier based on the Bayes theorem, strong (naive) independence assumptions, and a preselected set of independent feature models. Nave Bayes classification with kernel density estimation, or so called flexible Bayes is an extension of the nave Bayes classifier which uses a kernel density estimation where the density of each continuous variable is estimated averaging over a large set of kernels. The method performs well in domains that violate the normality assumption and, in general, this flexible Bayesian classifier generalises better than the version that assumes a single Gaussian. In opto-magnetic Imaging Spectroscopy was applied in vitro and in vivo on cervical, colon, and skin samples. Research included 280 cervical samples, 112 colon samples and 96 skin samples. The opto-magnetic spectra showed a good differentiation between healthy and cancerous samples based on characteristic OMIS spectra intensities and peak positions. It is shown that spectra intensity decreases to-wards lower values in cases of precancerous and cancerous tissues in all three kinds of epithelial tissue. Classification results, using naive Bayes classifier, proved a high degree of accuracy in cancer detection (skin 91.67) [21-23] (Figure 1).

Artificial neural networks

The need of unsupervised classifiers and predictors leads to neural networks (ANNs). ANNs are computational sim-ulations of human neural networks for modeling highly nonlinear systems in which the relationship between the variables is unknown or very complex. Artificial neural networks (ANNs) become a widely used methodology in nanomedicine, often to create accurate predictions. A neural network is formed by a series of neurons (or nodes) that are organized in layers. Each neuron in a layer is connected with each neuron in the next layer through a weighted connection. The value of the weight wij indicates the strength of the connection between the ith neuron in a layer and the jth neuron in the next one. The mathematical process through which the network achieves learning can be principally ignored by the final user. In this way, the network can be viewed as a black box that receives a vector with m inputs and provides a vector with n outputs [24].

In spite of the enthusiasm that went with ANN, limits of this approach have to be kept in mind. First of one, ANN methods are hungry of data and this is a hard limit in medicine. Overfitting and associated limits in generalization processes are always to be considered. The black-box nature of ANN is another hard limit in assisted health, where physician needs quantitative methodologies to support their decisions but they cannot be used to substitute them [25,26].

In [16] the relationships between the descriptors and the response variable were derived using Bayesian regularized neural networks. These control the complexity of models to provide a balance between bias (model is too simple to capture the underlying relationships in the data) and variance (model is overly complex and fits the noise as well as the underlying relationships). Bayesian regularization provides a near-optimum method of regularizing nonlinear neural network regression models.

Final considerations

In this paper we explored examples of use of quantitative techniques (data mining and deep learning) in nanomedicine. The sparse nature of issues regarding many types of nanoparticles, molecules and cells, many types of measurements and used instruments and probes, defines

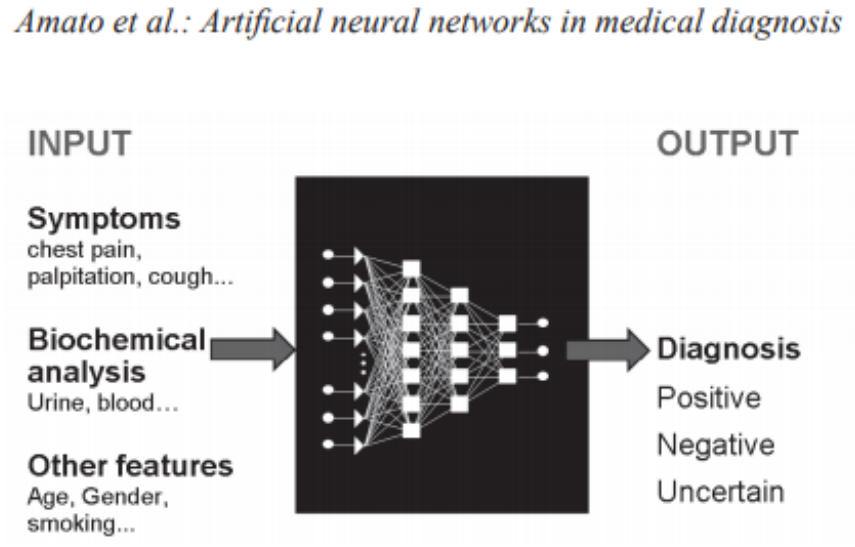

Figure 1. Source [24]

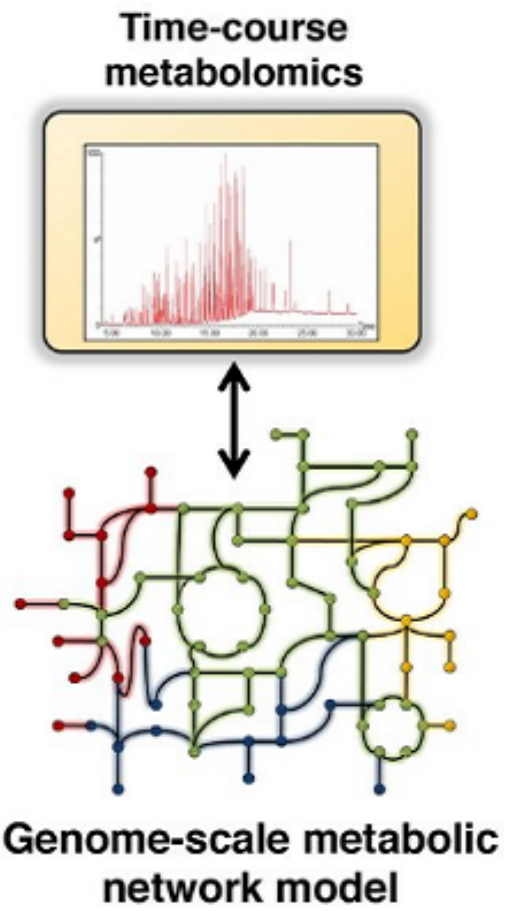

Figure 2. Source [26]

a starting phase for nanomedicine with a need of standards. QSAR methodologies need a strong data preparation, with attention to the number and the nature of used variables. Also the number of available samples has to be considered to select the model for the analysis. We analyzed PCA ad bayesian methodologies oriented to dimensional reduction. After we analyzed three clasess of methodologies: linear regression models, cluster analysis and artificial neural networks. The first class of models works with sparse data and for a limited set of contexts. Artificial neural networks also need large datasets. Cluster analysis is a general class of methodologies but need deeper exploration to understand effective relationships.

For every methodology there are contexts where it was successful in discovery a relationship or in simplify the adoption of a model, but we still aren't in a phase where successive studies reuse the same methodology, confirming its general validity in a context and for a class of problems. We thing there is a step that separates all these methodologies to this analytical phase: the biological process behind a treatment, also in its nano dimension, has a temporal evolu-tion that: 1) cannot be ignored, 2) cannot be considered with too complex models that quickly evolve in mathematically intractable problems. The gap to fill is the research of a new effective and feasible way between these two extreme approaches (Figure 2).

References

1. Riehemann K, Schneider SW, Luger TA, Godin B, Ferrari M (2009) Nanomedicine--challenge and perspectives. *Angew Chem Int Ed Engl* 48: 872-897. [Crossref]
2. Jones DE, Ghandehari H, Facelli J C (2016) A review of the applications of data mining and machine learning for the prediction of biomedical properties of nanoparticles. *Comput Methods Programs Biomed* 132: 93-103
3. Drasler B, Vanhecke D, (2017) Quantifying nanoparticle cellular uptake: which method is best. *Nanomedicine* 12: 1095-1099. [Crossref]
4. Kalinin SV, Strelcov E, Belianinov A, Somnath S, Vasudevan RK (2016) Big, Deep, and Smart Data in Scanning Probe Microscopy. *ACS Nano* [Crossref]
5. Liu R, Rallo R, Cohen Y (2013) Quantitative Structure-Activity-Relationships for Cellular Uptake of Nanoparticles, Proceed-ings of the 13th IEEE International Conference on Nanotechnology Beijing, China.
6. Baker NA, Klemm JD, Harper SL, Gaheen S, Heiskanen M (2013) Standardizing data. *Nat Nanotechnol* 8: 73-74. [Crossref]
7. Gaheen S, Hinkal GW (2013) caNanoLab: data sharing to expedite the use of nanotechnology in biomedicine. *Comput Sci Discov* 6: 014010. [Crossref]
8. Shi J, Kantoff PW, Wooster R, Farokhzad OC, (2017) Cancer nanomedicine: progress, challenges and opportunities. *Nat Rev Cancer* 17: 20-37. [Crossref]
9. Petersen GH, Alzghari SK (2016) Meta-analysis of clinical and preclinical studies comparing the anticancer efficacy of liposomal versus conventional non-liposomal doxorubicin. *J Control Release* 232: 266-264. [Crossref]
10. Gomes RHT, Moralles HF, Cominetti MR (2018) Global trends in nanomedicine research on triple negative breast cancer: a bibliometric analysis. *Int J Nanomedicine* 13: 2321-2336. [Crossref]
11. Papa E, Doucetb JB, Panaye AD (2016) Computational approaches for the prediction of the selective uptake of mag-netofluorescent nanoparticles into human cells. *RSC Advances*.
12. Rigon RB, Severino P, Santana MHA, Cho-rilli M (2018) Development of Solid Lipid Nanoparticles for Cutaneous Administration of Trans-resveratrol.
13. Thomas DG, Klaessig F (2011) Informatics and Standards for Nanomedicine Technology. *Wiley Interdiscip Rev Nanomed Nanobiotechnol* p. 3.
14. Cotya JB, Varenne F (2018) Characterization of nanomedicines surface coverage using molecular probes and capillary electrophoresis. *Eur J Pharm Biopharm* 130: 48-58.
15. Sayes C, Ivanov I (2010) Comparative Study of Predictive Com-putational Models for Nanoparticle-Induced Cytotoxicity. *Risk Anal* 30: 1723-1734.
16. Burden FR, Winkler DA (2009) Optimal Sparse Descriptor Selection for QSAR Using Bayesian Methods. *QSAR Comb Sci* 28: 6-7
17. Fourches D, Pu D, Tassa C, Weissleder R (2010) Quanti-tative nanostructure-activity relationship modelling. *ACS nano* p. 4.
18. Vrieze JD (2018) Meta-analyses were supposed to end scien-tific debates Often, they only cause more controversy.
19. Wilhelm S, Tavares AJ, Dai Q (2016) Analysis of nanoparticle delivery to tumours. *Nature Reviews Materials* p. 14.
20. Liu R, Rallo R, George S (2011) Classification NanoSAR development for cytotoxicity of metal oxide nanoparticles. *Small* 7: 1118-1126.
21. Liu X, Tang K, Harper S, Harper B, Steevens JA, et al. (2013) Pre-dictive modeling of nanomaterial exposure effects in biological systems. *Int J Nanomedicine* 8: 31-43.
22. Mel AD, Kalaskar DM (2014) Nanomedicine Ed By University College London, United Kingdom.
23. Soria D, Garibaldi JM, Ambrogi F, Biganzoli EM, Ellis IO (2011) A non-parametric version of the nave Bayes classifier. *Knowledge-Based Systems* 24: 775-784.
24. Amato F, Lpez A (2013) Artificial neural networks in medical diagnosis. *Journal of Applied Biomedicine* 11: 47-58
25. Gary Marcus (2018) Deep Learning: a critical appraisal arXiv:180100631.
26. https://metabolist.wordpress.com/2018/01/02/metabolist-december-2018/