BioTech

MDPI

Commentary

Predictive Modelling in Clinical Bioinformatics: Key Concepts for Startups

Ricardo J. Pais [1,2]

[1] Bioenhancer Systems, Office 63 182-184 High Street North, East Ham, London E6 2JA, UK; rjpais@bioenhancersystems.com
[2] Centro de investigação Interdisciplinar Egas Moniz (CiiEM), Instituto Universitário Egas Moniz, 2829-511 Caparica, Portugal

Abstract: Clinical bioinformatics is a newly emerging field that applies bioinformatics techniques for facilitating the identification of diseases, discovery of biomarkers, and therapy decision. Mathematical modelling is part of bioinformatics analysis pipelines and a fundamental step to extract clinical insights from genomes, transcriptomes and proteomes of patients. Often, the chosen modelling techniques relies on either statistical, machine learning or deterministic approaches. Research that combines bioinformatics with modelling techniques have been generating innovative biomedical technology, algorithms and models with biotech applications, attracting private investment to develop new business; however, startups that emerge from these technologies have been facing difficulties to implement clinical bioinformatics pipelines, protect their technology and generate profit. In this commentary, we discuss the main concepts that startups should know for enabling a successful application of predictive modelling in clinical bioinformatics. Here we will focus on key modelling concepts, provide some successful examples and briefly discuss the modelling framework choice. We also highlight some aspects to be taken into account for a successful implementation of cost-effective bioinformatics from a business perspective.

Keywords: predictive modelling; clinical bioinformatics; mathematical models; diagnostics; prognostics; clinical applications

check for updates

Citation: Pais, R.J. Predictive Modelling in Clinical Bioinformatics: Key Concepts for Startups. *BioTech* **2022**, *11*, 35. https://doi.org/10.3390/biotech11030035

Academic Editors: Pietro Pinoli and Anna Bernasconi

Received: 20 June 2022
Accepted: 3 August 2022
Published: 17 August 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Clinical Bioinformatics Role and its Dependency on Predictive Modelling

Clinicians consider access of patients' genetic information from genomes, transcriptomes, proteomes and metabolomes as advantageous for improving diagnostics and prognostics of diseases [1–5]. Accessing clinically relevant information from these 'omics' data is considered by many as precision medicine, which has the potential to enable more personalized and effective medicine [1,4,5]. Current advances in Next Generation Sequencing (NGS) and Mass Spectrometry (MS) technologies made possible the characterization of genomes and quantification of proteomes from patients' biological samples with reasonable accuracy and scalability, compatible with its application in the clinical point-of-care [6–11]; however, data from these technologies is too complex to be humanly handled and interpreted by clinicians. Bioinformatics is fundamental for providing humanly readable and clinically relevant genomics and proteomics interpretations from NGS and MS techniques [1,3,8,9]. For these reasons, bioinformatics is considered as a fundamental bridge between clinicians and 'omics' technology making this field quite attractive to the medical community. Another attractive feature of bioinformatics is also due to its potential to facilitate the automation of data analysis and opens the possibility for "big data" processing [12–15]; this will be advantageous when the new digital Era fully reaches the medical industry and clinical laboratories [7,16]. Although bioinformatics has its origins in evolutionary biology and intimately linked to genomics, a new clinically focused branch is growing fast and diversifying from the traditional bioinformatics [3,17]; this is called by many clinical bioinformatics and its objectives are focused on obtaining diagnostics, prognostics, or therapy assessment out of the data from an individual patient. The clinical bioinformatics branch is expected to play a central role in facilitating the identification of genetic diseases, the discovery of novel biomarkers, characterization of pathogens and enable a more informed decision for the therapeutical strategy to follow [3,18,19].

Predictive modelling on the other hand is a speciality commonly used in data science, computational biology and systems biology for more than six decades [5,20]. Recently, researchers have been proposing the combination of predictive modelling approaches with bioinformatics for improving current practices in disease identification, therapeutics and prognostics [13,18,21–23]; these have been shown advantageous to unlock the full potential of many high-throughput technologies as solutions for large population screening of multiple disorders and precision medicine; this is evident for high-throughput technologies such as mass spectrometry and next-generation sequencing which contains a huge and complex amount of information that require both high computer processing power and advanced mathematical modelling approaches for translating the complexity of the data into clinically relevant predictions [5,13,21]. Therefore, predictive modelling is set to play a key role in clinical bioinformatics, which should become part of the standard clinical bioinformatic pipelines as a downstream analysis step following a traditional bioinformatics pipeline; however, this step needs to be integrated with typical bioinformatics pipelines from genomics, proteomics and transcriptomics. In each of these cases, the predictive modelling step takes as input the bioinformatically curated "omics" data, integrates it with other sources of patient data (metadata) and generates an output that should be relevant phenotype information readable by a clinician; this integration is not an easy task and would depend on high-throughput data, the available metadata, the target disease and more important the choice of the modelling framework.

2. Key Concepts of Predictive Modelling in the Clinical Context

Despite the efforts for conveying the correct role of predictive modelling in the clinical context, there is still some misunderstanding in the medical and biotech community regarding key concepts underlying its application [18,24–26]; this often results in either undervaluing the modelling step or extrapolating it to a science fiction story. To fully understand the clinical applications of predictive modelling and their caveats, we should start by simply defining what is a predictive model in the first place. A predictive model is any mathematical abstraction of a system which generates a prediction of an unknown system component/property based on known system components [24,27,28]. In our case, the model is a conceptual description of a biomedical system of interest, where the components of the system are:

- Biological relevant and measurable or observational entities (dependent variables), which are the inputs of the model.
- Relational factors between variables with or without biological meaning (parameters), which can be estimated empirically or based on data fitting methodologies.
- Unknown clinical entities or properties of interest for prediction (dependent variables), which are the outputs of the model.

Models can be formally translated through mathematical equations, graphs containing processes with gate decisions or even more complex mathematical objects [24,25,28]. Models should be as accurate as possible in descriptions of the system and validated against enough data in an unbiased and independent manner [26,27,29]. In theory, we can always find a conceptual model behind any diagnostic methodology [18,27]; however, it is important to highlight that there is always a certain degree of uncertainty associated with any prediction generated by any model [29,30]. Thus, the clinical application of a predictive model depends on the evaluated performance during the validation process. For example, a mathematical model can be considered as part of a diagnostic test if the performance exceeds the minimum sensitivity and specificity required by medical authorities [18,29]. Usually, these values are around 97% but can change depending on the country and disease.

Otherwise, models fall into a predictive category and should be taken as an indication or tendency for such possibility, where the clinician's interpretation and judgment is absolutely necessary. In this scenario, the model prediction is useful as supporting information for the clinician, which is always better than relying on chance, intuition or based on the outcomes of previous patients. In general terms, predictive models should be seen as insights that enable a clinician to make a better informed and supported decision [17,18,20].

3. Examples of Clinical Applications of Predictive Modelling

There are several types of modelling techniques that can be used in bioinformatics for clinical diagnostics and prognostics. We listed the most frequently used in clinical contexts and describe their basic characteristics in Table 1. The most straightforward application of predictive models is the capacity of generating a prediction for the future [26,29]; this can be applied in the clinical context for the generation of disease prognostics for example for predicting; the emergence of developing a particular genetic disease, disease evolution, impact on life expectancy, impact on the society or even the success of a given treatment [24–27]. Predicting survival expectancy and response to treatments based on information from a given tumour biopsy characterization is an application that is often attempted by multiple predictive modelling approaches such as statistical, machine learning and logical network modelling [26,31–38]. Another interesting example is the modelling efforts conducted for predicting the impact and control of SARScov2 transmission effects and control during the COVID-19 outbreak [39–42]. In these works, statistical and Ordinary Differential Equations (ODE) based models have been successfully used for predicting expected peaks of infected, hospitalized and the timings by which the peaks occurred. Further, simulations from Susceptible and Infected ODE models also predicted useful information for the decision of implementing controlling measures that minimize the total of deaths in a certain region [39–41].

Predictive models are also useful for the detection of diseases in an early stage, in particular, if current diagnostic methods fail and the treatment efficiency benefits from early detection. One good example of this scenario is the poor detection rates of 40% observed during ovarian cancer screening programs [43,44]; this type of cancer does not show symptoms up to later stages and by then the treatment success is largely compromised. Some statistical and machine learning models have successfully combined multiple biomarkers resulting in surprisingly high sensitivities (>90%) and reasonable specificities (> 80%) which largely outcompete the sensitivities and specificities obtained under current screening programs [23,45]. The application of such modelling approaches at the point-of-care would definitely improve the identification rates at early stages potentially saving thousands of lives of women from ovarian cancer every year.

Another clinical application of predictive models is the generation of insights when the current diagnostic methodologies are too invasive and put at risk the health of the testing subjects [46,47]. One illustrative example of this scenario is the detection of genetic diseases in prenatal screening by amniocentesis and pre-implantation embryo testing using post-freezing PGT-A next-generation sequencing techniques [12,48–52]. In both cases, the procedures for conducting genetic testing are too invasive, compromising the viability of pregnancy and embryo survival during implantation. Machine learning models have been quite successful in predicting aneuploidies from indirect data such as embryo secretome in culture media and urine [48,49]. Impressively, predictive models from mass spectral patterns of secretome have rendered sensitivities very close to the diagnostic level with reasonably tolerable false positive rates, enabling affordable and non-invasive testing [48].

Table 1. Often used modelling techniques in clinical bioinformatics and their main characteristics.

Modelling Technique	Description	Application	Requirements
Statistical	Scoring and probability functions that assumes a distribution shape or behaviour.	Continuous Quantification	Data for parameter estimation. Depend on sample size.
Kinetic	Solving of systems of nonlinear differential equations. Do not assume any behaviour. Instead relies on rate laws of processes such as chemical reactions.	Binary Classification	Requires reported or estimated kinetic parameter. Do not depend on sample size.
Logical	Solving of logical equations based on predefined rules for each component. Assumes asynchronous or synchronous update schemes.	Binary Classification	Requires relational knowledge of its components. Do not depend on sample size.
Regression	Fitting of an assumed mathematical equation on data. Often are used models that describe a particular assumed data behaviour such as linear, polynomial, exponential, and logistic.	Binary Classification	Data for model fitting. Depend on sample size.
Random Forests	Supervised machine leaning algorithm based on averaging multiple generated decision trees.	Binary Classification	Data for model training and validation. Requires large datasets
Support Vector Machines	Supervised machine leaning algorithm based on clustering algorithms such as principal component analyses.	Binary Classification	Data for model training and validation. Requires large datasets
Neural Networks	Supervised machine leaning algorithm based on defining a set of neuron and layers as model components. Assumes all possible relational interactions between neurons.	Binary Classification	Data for model training and validation. Requires large datasets

Sequence-based prediction of pathogenic genetic variants (Single Nucleotide Polymorphisms, insertions or deletions) is becoming now a very important modelling application in clinical bioinformatics, in particular for the identification of rare genetic diseases [53–55]; these are based on predicting deleterious effects on protein function from gene sequence based on evolutionary conservation of sequence motifs or on machine learning approaches. There are multiple successful examples of models and tools with reasonable sensitivities over 85% such as SIFT, mutation taster, mutation accessor, Fathmn, Phanter and Polyphen-2 [54].

4. Choosing the Correct Modelling Framework

Choosing the correct modelling framework is a critical step in developing a suitable model and often is neglected from the beginning [5,56,57]. In most cases, researchers often start from their favourite modelling framework in an attempt to apply it in a given problem; this is not the best policy and resembles the usage of a hammer to perform all construction labour. Ideally, we should first gather the available data, available knowledge of the system we want to model and access which is the best suitable modelling framework for that particular case [27,58]; this is very tedious and theoretical research work but often pays off as it will save time later on by preventing reaching dead ends where models do not describe the systems, cannot be validated or their performances are simply not different from flipping a coin. Here, we briefly describe some advantages and disadvantages of the most promising modelling frameworks with clinical applications.

Machine-learning is a very powerful approach which is ideal for building disease classifiers with yes or no outcomes [8,21,59,60]. For the choice of this approach, it is mandatory to have sufficiently large datasets where the disease outcomes are known [21,60]; this is an absolute requirement for the training and validation of models. Using this approach, it is recommended to try multiple algorithms that have been quite success with biomedical data such as random forests, neural networks, regression models and support vector machines [21,60]. Neural networks (NN) and Recurrent Neural Networks (RNN) are particular important types of machine learning algorithms based on its high efficiency and robustness if well implemented and validated [61,62]; this is particularly important for modelling sequence-based phenotypes with clinical relevance. Hyperparameter exploratory analysis is also a necessary task in this approach, which can be an extensively time-consuming and computationally heavy [59,63]. Most of these algorithms are available in R packages, python libraries and even in Auto ML tools which is a huge advantage that facilitates the implementation of automated workflows for the model generation [63,64]; however, machine learning approaches are "black box" models which are prone to overfitting and artefactual models [21,60,65]. Thus, such modelling frameworks require additional and periodically checking of their reliability; moreover, the absence of knowing the exact rationale behind such prediction with some algorithms may cause difficulties in registration of diagnostic tests and patents.

Statistical models are the most conservative modelling approaches used in clinical contexts [21,66,67]. The development of these types of models relies on the choice of a theoretical statistical model and requires the estimation of its parameters with data. Often this approach is combined with machine learning algorithms for data fitting-based parameter estimation [21]. Statistical models can offer an estimated probability of having or not a particular disease; this brings an advantage over classification models in particular for the scenarios that best describe "gray" zones of uncertainty making them more realistic than the yes/no classification models This type of modelling depends on sample size numbers but often do not require large datasets as in machine learning approaches. Another advantage of this modelling framework is that is simple and has a straightforward implementation in laboratory software tools. A good example of these advantages was capture for the screening of multiple blood disorders using mass spectrometry, where we implemented a cumulative probability function in a laboratory software tool that enables automated estimating of the probability of a patient having a particular type of hemoglobinopathy on a large scale of analysis, applicable to population screenings [22].

Deterministic frameworks such as ODE-based (also called kinetic) and logical modelling are powerful quantitative (kinetic) and discrete modelling approaches (logical) [28,30,68]. Both are by far more descriptive in comparison to statistical and machine learning. The underlying principles of these frameworks rely on the laws of chemistry, physics, biochemical circuits and mathematics, making them more realistic and robust for finding drug targets and predicting therapeutic effects [24,27,28]. For example, exploring kinetic and logical models of signalling pathways in cancer growth and invasion results in predicted effects of drug targets for cell decisions that can be useful for therapy choices and predict cancer aggressiveness and progression [33,34,36,69–71]. In contrast with statistical and machine learning, this approach does not rely on sample sizes but requires extensive literature knowledge including knowing parameters and relational laws [27,28]. Developing such models is a huge investment of effort and time-consuming in comparison to the other frameworks as are more complex in terms of variables, and development and require an huge in-depth knowledge of the system; these models may take years to develop and depend on the availability of existing literature data or the capacity to estimate them experimentally [27,28]. In comparison, kinetic models are always preferable to logical as they are more accurate descriptions of the systems and provide a quantitative assessment [28,30,68]. The only advantage of choosing logical modelling is in the case where the kinetic parameters (e.g., rate constants of the processes) are unknown [28].

5. Challenges of Clinical Bioinformatics: The Business Perspective

Innovative technology coming from academic research related to clinical bioinformatics is attracting private investment and generate new startups. Most of these startups come from the academic labs which have developed during research projects attractive state-of-the-art bioinformatics algorithms and pipelines that can be applied to a new service or product that potentially can generate growth [3,8]. Upon investment, these startups face a paradigm change that constitutes a huge challenge for both academics that migrate to the industry and investors that need to guarantee the revenues of their investment. One of the main initial issues that most startups are facing is related with software tools and copywrite issues. In bioinformatics, the current way of thinking is based on the usage of command line tools which were developed for academic purposes and are restricted to commercial usage. Sometimes the developed technology utilizes such tools which causes software license issues; this forces startups to either develop their own "in house" workflows almost from scratch which is time-consuming, or buy the respective licenses which in most scenarios can compromise the business sustainability.

Another key issue is the patient's data; this is often a very sensitive issue which requires following an ethics protocol of personal data protection during data acquisition, storage and analysis [8,72]. Thus, it is absolutely mandatory to implement a secure database system for protecting patients' personal data and still enabling the bioinformatics pipelines to access some metadata of relevance for conducting the analysis [72]. A simple solution for this could be through using anonymized data pulling during the analysis and automated reporting that can be generated through the usage of secure relational databases.

Code protection is quite trivial in most informatics companies as a standard of best practices but in most academic bioinformatics groups the data is often saved in the postdoc personal computer and publicly available in multiple GitHub accounts. Although this is a severe data security breach it makes it impossible to get copyrights and patents leading to a business loss or a huge shift from original technology [72]. Ideally, the code and also the data should be stored in data centres for ensuring enough security, privacy police and maintenance with proper SSL certificates. Additionally, proprietary code should also be store into private GitHub or GitLab accounts for organizations with suitable ownership of the company and restrictive accessions of developers from both inside and outside of the organization. Both GitHub and GitLab enables such functionalities even for free as this is standard in commercial-based informatics projects and businesses. Ideally, these practices should be taken into account as early as possible, even during the phase of technology development.

Often, the transition from the academical environment to an industry startup environment must follow a huge change in the mind set of researchers; this includes the way to think and work also. From individually tacking a project to teams within compartmentalized projects, to following standard methods of software development like management frameworks like agile and the available implementations such as Kanban board; this last agile implementation is a very popular and flexible solution which is frequently available in many online software tools such as GitHub and Jira. The focus will be no longer, addressing a question and understanding a mechanism. Instead, the focus is finding solutions that are robust and meet company objectives with defined deadlines. One solution to facilitate this is to take advantage of available online courses in the field of informatics that can introduce the best practices to follow for project management and tools; this would help substantially researchers to optimize and adjust their way of working and tackling projects.

Scalability is also another issue to deal, as most of the technology is thought as an analysis service conducted by bioinformaticians as users; this would eventually become saturated because finding bioinformaticians as work force is limited and not an easy task [8,64]. Besides, the cost-effectiveness of the service is also compromised as well the competitiveness simply because bioinformaticians are not cheap labour. In this case, the ideal system would the implementation of automated pipelines that conduct the analysis without human intervention under a software as a service business model (SaaS); this would ensure both scalability and cost-effectiveness as there is only need for a bioinformatics team to maintain and improve the pipelines. Therefore, hiring the correct bioinformatics team is fundamental for the health and growth of the startup; this is often neglected and is indeed a difficult task to find such highly specialized professionals which sometime can be considered as rare unicorns.

6. Conclusions and Perspectives

Predictive modelling approaches have a fundamental role in ensuring the applicability of bioinformatics in the clinical context; it is also fundamental to invest in the correct modelling framework for each case and properly integrated with the bioinformatics pipeline and high-throughput technology to ensure the robustness of results given to a clinician and a technological gain in comparison to current methodologies available.

Clinical bioinformatics is still in its initial phase of growth and many startups are only now emerging from new born innovative technology coming from academia. Yet there is a long learning and adaptative process for successfully migrating from an academical mindset towards sustainable clinical bioinformatic services. There are still many challenges to overcome in the future to ensure a successful acceptancy of clinical bioinformatics and its generalization to the point-of-care. The future of clinical bioinformatic may depend on choosing a suitable and modern business model such as SaaS to ensure the sustainability of clinical bioinformatics services. In the future, this would be fundamental for keeping up with a possible scaling up of the demand for bioinformatic services from clinicians. Also keeping up with migration of clinical data to its digital form and becoming compatible with "big data" processing.

Funding: This research work received no external funding.

Data Availability Statement: No additional data is available for this commentary.

Conflicts of Interest: The author declares no confit of interest.

References

1. Denny, J.C.; Bastarache, L.; Roden, D.M. Phenome-Wide Association Studies as a Tool to Advance Precision Medicine. *Annu. Rev. Genomics Hum. Genet.* **2016**, *17*, 353–373. [CrossRef] [PubMed]
2. Bilder, R.M.; Sabb, F.W.; Cannon, T.D.; London, E.D.; Jentsch, J.D.; Parker, D.S.; Poldrack, R.A.; Evans, C.; Freimer, N.B. Phenomics: The Systematic Study of Phenotypes on a Genome-Wide Scale. *Neuroscience* **2009**, *164*, 30–42. [CrossRef] [PubMed]
3. Tsongalis, G.J.; Chao, E.; Hagenkord, J.M.; Hambuch, T.; Moore, J.H. Bioinformatics: What the Clinical Laboratorian Needs to Know and Prepare For. *Clin. Chem.* **2013**, *59*, 1301–1305. [CrossRef]
4. Mack, S.C.; Northcott, P.A. Genomic Analysis of Childhood Brain Tumors: Methods for Genome-Wide Discovery and Precision Medicine Become Mainstream. *J. Clin. Oncol.* **2017**, *35*, 2346–2354. [CrossRef] [PubMed]
5. Kholodenko, B.; Yaffe, M.B.; Kolch, W. Computational Approaches for Analyzing Information Flow in Biological Networks. *Sci. Signal.* **2012**, *5*, 1–14. [CrossRef] [PubMed]
6. McDermott, U. Next-Generation Sequencing and Empowering Personalised Cancer Medicine. *Drug Discov. Today* **2015**, *20*, 1470–1475. [CrossRef] [PubMed]
7. Pais, R.J. Bioinformatics and Predictive Modelling as Tools for Clinical Diagnostics. 2020, pp. 30–34. Available online: https://insights.omnia-health.com/laboratory/bioinformatics-and-predictive-modelling-tools-clinical-diagnostics (accessed on 1 August 2022).
8. Mann, M.; Kumar, C.; Zeng, W.F.; Strauss, M.T. Artificial Intelligence for Proteomics and Biomarker Discovery. *Cell Syst.* **2021**, *12*, 759–770. [CrossRef]
9. Khamis, M.M.; Adamko, D.J.; El-Aneed, A. Mass Spectrometric Based Approaches in Urine Metabolomics and Biomarker Discovery. *Mass Spectrom. Rev.* **2017**, *36*, 115–134. [CrossRef] [PubMed]
10. Morris, J.S.; Baggerly, K.A.; Gutstein, H.B.; Coombes, K.R. Statistical Contributions to Proteomic Research. *Methods Mol. Biol.* **2010**, *641*, 143–166. [CrossRef] [PubMed]
11. Zhao, Y. Whole Genome and Exome Sequencing Reference Datasets from a Multi-Center and Cross-Platform Benchmark Study. *Sci. Data* **2021**, *8*, 296. [CrossRef]

12. Pais, R.J.; Zmuidinaite, R.; Butler, S.A.; Iles, R.K. An Automated Workflow for MALDI-ToF Mass Spectra Pattern Identification on Large Data Sets: An Application to Detect Aneuploidies from Pregnancy Urine. *Inform. Med. Unlocked* **2019**, *16*, 100194. [CrossRef]
13. Pais, R.J.; Iles, R.K.; Zmuidinaite, R. MALDI-ToF Mass Spectra Phenomic Analysis for Human Disease Diagnosis Enabled by Cutting-Edge Data Processing Pipelines and Bioinformatic Tools. *Curr. Med. Chem.* **2021**, *28*, 6532–6547. [CrossRef] [PubMed]
14. Weisser, H.; Nahnsen, S.; Grossmann, J.; Nilse, L.; Quandt, A.; Brauer, H.; Sturm, M.; Kenar, E.; Kohlbacher, O.; Aebersold, R.; et al. An Automated Pipeline for High-Throughput Label-Free Quantitative Proteomics. *J. Proteome Res.* **2013**, *12*, 1628–1644. [CrossRef] [PubMed]
15. Malm, E.K.; Srivastava, V.; Sundqvist, G.; Bulone, V. APP: An Automated Proteomics Pipeline for the Analysis of Mass Spectrometry Data Based on Multiple Open Access Tools. *BMC Bioinform.* **2014**, *15*, 441. [CrossRef]
16. Hu, C.; Kumar, S.; Huang, J.; Ratnavelu, K. How to Better Satisfy Online Users? A Quantitative Study of Identity Reconstruction Based on Advanced Self-Discrepancy Theory. *J. Data Sci.* **2018**, *15*, 020081.
17. Belmont, J.W.; Shaw, C.A. Clinical Bioinformatics: Emergence of a New Laboratory Discipline. *Expert Rev. Mol. Diagn.* **2016**, *16*, 1139–1141. [CrossRef] [PubMed]
18. Simon, R. Genomic Biomarkers in Predictive Medicine: An Interim Analysis. *EMBO Mol. Med.* **2011**, *3*, 429–435. [CrossRef]
19. Ao Kong, A.; Gupta, C.; Ferrari, M.; Agostini, M.; Bedin, C.; Bouamrani, A.; Tasciotti, E.; Azencott, R. Biomarker Signature Discovery from Mass Spectrometry Data. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **2014**, *11*, 766–772. [CrossRef]
20. Chuang, H.-Y.; Hofree, M.; Ideker, T. A Decade of Systems Biology. *Annu. Rev. Cell Dev. Biol.* **2010**, *26*, 721–744. [CrossRef]
21. De Ridder, D.; De Ridder, J.; Reinders, M.J.T. Pattern Recognition in Bioinformatics. *Brief. Bioinform.* **2013**, *14*, 633–647. [CrossRef]
22. Pais, R.J.; Jardine, C.; Zmuidinaite, R.; Lacey, J.; Butler, S.; Iles, R. Rapid, Affordable and Efficient Screening of Multiple Blood Abnormalities Made Possible Using an Automated Tool for MALDI-ToF Spectrometry Analysis. *Appl. Sci.* **2019**, *9*, 4999. [CrossRef]
23. Pais, R.J.; Zmuidinaite, R.; Lacey, J.C.; Jardine, C.S.; Iles, R.K. A Rapid and Affordable Screening Tool for Early-Stage Ovarian Cancer Detection Based on MALDI-ToF MS of Blood Serum. *Appl. Sci.* **2022**, *12*, 3030. [CrossRef]
24. Ay, A.; Arnosti, D.N. Mathematical Modeling of Gene Expression: A Guide for the Perplexed Biologist. *Crit. Rev. Biochem. Mol. Biol.* **2011**, *46*, 137–151. [CrossRef]
25. Fisher, J.; Henzinger, T. A Executable Cell Biology. *Nat. Biotechnol.* **2007**, *25*, 1239–1249. [CrossRef] [PubMed]
26. Benson, N.; van der Graaf, P.H.; Peletier, L.A. Use of Mathematics to Guide Target Selection in Systems Pharmacology; Application to Receptor Tyrosine Kinase (RTK) Pathways. *Eur. J. Pharm. Sci.* **2017**, *109*, S140–S148. [CrossRef] [PubMed]
27. Somvanshi, P.R.; Venkatesh, K.V. A Conceptual Review on Systems Biology in Health and Diseases: From Biological Networks to Modern Therapeutics. *Syst. Synth. Biol.* **2014**, *8*, 99–116. [CrossRef]
28. Le Novère, N. Quantitative and Logic Modelling of Molecular and Gene Networks. *Nat. Rev. Genet.* **2015**, *16*, 146–158. [CrossRef]
29. Dankers, F.J.W.M.; Traverso, A.; Wee, L.; van Kuijk, S.M.J. Prediction Modeling Methodology. In *Fundamentals of Clinical Data Science*; Springer International Publishing: Cham, Switzerland, 2019; pp. 101–120.
30. Qian, G.; Mahdi, A. Sensitivity Analysis Methods in the Biomedical Sciences. *Math. Biosci.* **2020**, *323*, 108306. [CrossRef]
31. Swan, A.L.; Mobasheri, A.; Allaway, D.; Liddell, S.; Bacardit, J. Application of Machine Learning to Proteomics Data: Classification and Biomarker Identification in Postgenomics Biology. *Omi. A J. Integr. Biol.* **2013**, *17*, 595–610. [CrossRef]
32. Edwards, N.J.; Oberti, M.; Thangudu, R.R.; Cai, S.; McGarvey, P.B.; Jacob, S.; Madhavan, S.; Ketchum, K.A. The CPTAC Data Portal: A Resource for Cancer Proteomics Research. *J. Proteome Res.* **2015**, *14*, 2707–2713. [CrossRef] [PubMed]
33. Pais, R.J. Simulation of Multiple Microenvironments Shows a Pivot Role of RPTPs on the Control of Epithelial-to-Mesenchymal Transition. *Biosystems* **2020**, *198*, 104268. [CrossRef] [PubMed]
34. Lebedeva, G.; Sorokin, A.; Faratian, D.; Mullen, P.; Goltsov, A.; Langdon, S.P.; Harrison, D.J.; Goryanin, I. Model-Based Global Sensitivity Analysis as Applied to Identification of Anti-Cancer Drug Targets and Biomarkers of Drug Resistance in the ErbB2/3 Network. *Eur. J. Pharm. Sci.* **2012**, *46*, 244–258. [CrossRef] [PubMed]
35. Flobak, Å.; Baudot, A.; Remy, E.; Thommesen, L.; Thieffry, D.; Kuiper, M.; Lægreid, A. Discovery of Drug Synergies in Gastric Cancer Cells Predicted by Logical Modeling. *PLoS Comput. Biol.* **2015**, *11*, e1004426. [CrossRef] [PubMed]
36. Wynn, M.L.; Consul, N.; Merajver, S.D.; Schnell, S. Logic-Based Models in Systems Biology: A Predictive and Parameter-Free Network Analysis Method. *Integr. Biol.* **2012**, *4*, 1323–1337. [CrossRef]
37. Calzone, L.; Tournier, L.; Fourquet, S.; Thieffry, D.; Zhivotovsky, B.; Barillot, E.; Zinovyev, A. Mathematical Modelling of Cell-Fate Decision in Response to Death Receptor Engagement. *PLoS Comput. Biol.* **2010**, *6*, e1000702. [CrossRef]
38. Anderson, A.R.A.; Weaver, A.M.; Cummings, P.T.; Quaranta, V. Tumor Morphology and Phenotypic Evolution Driven by Selective Pressure from the Microenvironment. *Cell* **2006**, *127*, 905–915. [CrossRef]
39. Pais, R.J.; Taveira, N. Predicting the Evolution and Control of the COVID-19 Pandemic in Portugal. *F1000Research* **2020**, *9*, 283. [CrossRef]
40. IHME COVID-19 Health Service Utilization Forecasting Team; Murray, C.J.L. Forecasting COVID-19 Impact on Hospital Bed-Days, ICU-Days, Ventilator-Days and Deaths by US State in the next 4 Months. *medRxiv* **2020**. [CrossRef]
41. Kucharski, A.J.; Russell, T.W.; Diamond, C.; Liu, Y.; Edmunds, J.; Funk, S.; Eggo, R.M.; Sun, F.; Jit, M.; Munday, J.D.; et al. Early Dynamics of Transmission and Control of COVID-19: A Mathematical Modelling Study. *Lancet Infect. Dis.* **2020**, *3099*, 1–7. [CrossRef]

42. Chen, T.M.; Rui, J.; Wang, Q.P.; Zhao, Z.Y.; Cui, J.A.; Yin, L. A Mathematical Model for Simulating the Phase-Based Transmissibility of a Novel Coronavirus. *Infect. Dis. Poverty* **2020**, *9*, 1–8. [CrossRef]
43. Henderson, J.T.; Webber, E.M.; Sawaya, G.F. Screening for Ovarian Cancer. *JAMA* **2018**, *319*, 595. [CrossRef] [PubMed]
44. Jacobs, I.J.; Menon, U.; Ryan, A.; Gentry-Maharaj, A.; Burnell, M.; Kalsi, J.K.; Amso, N.N.; Apostolidou, S.; Benjamin, E.; Cruickshank, D.; et al. Ovarian Cancer Screening and Mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): A Randomised Controlled Trial. *Lancet* **2016**, *387*, 945–956. [CrossRef]
45. Whitwell, H.J.; Worthington, J.; Blyuss, O.; Gentry-Maharaj, A.; Ryan, A.; Gunu, R.; Kalsi, J.; Menon, U.; Jacobs, I.; Zaikin, A.; et al. Improved Early Detection of Ovarian Cancer Using Longitudinal Multimarker Models. *Br. J. Cancer* **2020**, *122*, 847–856. [CrossRef] [PubMed]
46. Rosenwaks, Z.; Handyside, A.H.; Fiorentino, F.; Gleicher, N.; Paulson, R.J.; Schattman, G.L.; Scott, R.T.; Summers, M.C.; Treff, N.R.; Xu, K. The Pros and Cons of Preimplantation Genetic Testing for Aneuploidy: Clinical and Laboratory Perspectives. *Fertil. Steril.* **2018**, *110*, 353–361. [CrossRef]
47. Cimadomo, D.; Capalbo, A.; Ubaldi, F.M.; Scarica, C.; Palagiano, A.; Canipari, R.; Rienzi, L. The Impact of Biopsy on Human Embryo Developmental Potential during Preimplantation Genetic Diagnosis. *Biomed Res. Int.* **2016**, *2016*, 7193075. [CrossRef]
48. Pais, R.J.; Sharara, F.; Zmuidinaite, R.; Butler, S.; Keshavarz, S.; Iles, R. Bioinformatic Identification of Euploid and Aneuploid Embryo Secretome Signatures in IVF Culture Media Based on MALDI-ToF Mass Spectrometry. *J. Assist. Reprod. Genet.* **2020**, *37*, 2189–2198. [CrossRef]
49. Ray, K.I.; Nicolaides, K.; Pais, R.; Zmuidinaite, R.; Keshavarz, S.; Poon, L.; Butler, S. The Importance of Gestational Age in First Trimester, Maternal Urine MALDI-Tof MS Screening Tests for Down Syndrome. *Ann. Proteomics Bioinforma.* **2019**, *3*, 10–17. [CrossRef]
50. Sharara, F.; Butler, S.A.; Pais, R.J.; Zmuidinaite, R.; Keshavarz, S.; Iles, R.K. BESST, a Non-Invasive Computational Tool for Embryo Selection Using Mass Spectral Profiling of Embryo Culture Media. *EMJ Repro Health* **2019**, *5*, 59–60.
51. Campbell, A.; Fishel, S.; Bowman, N.; Duffy, S.; Sedler, M.; Hickman, C.F.L. Modelling a Risk Classification of Aneuploidy in Human Embryos Using Non-Invasive Morphokinetics. *Reprod. Biomed. Online* **2013**, *26*, 477–485. [CrossRef]
52. Scriven, P.N. Towards a Better Understanding of Preimplantation Genetic Screening for Aneuploidy: Insights from a Virtual Trial for Women under the Age of 40 When Transferring Embryos One at a Time. *Reprod. Biol. Endocrinol.* **2017**, *15*, 49. [CrossRef]
53. Dong, C.; Wei, P.; Jian, X.; Gibbs, R.; Boerwinkle, E.; Wang, K.; Liu, X. Comparison and Integration of Deleteriousness Prediction Methods for Nonsynonymous SNVs in Whole Exome Sequencing Studies. *Hum. Mol. Genet.* **2015**, *24*, 2125–2137. [CrossRef] [PubMed]
54. Montenegro, L.R.; Lerário, A.M.; Nishi, M.Y.; Jorge, A.A.L.; Mendonca, B.B. Performance of Mutation Pathogenicity Prediction Tools on Missense Variants Associated with 46,XY Differences of Sex Development. *Clinics* **2021**, *76*, e2052. [CrossRef] [PubMed]
55. Seaby, E.G.; Pengelly, R.J.; Ennis, S. Exome Sequencing Explained: A Practical Guide to Its Clinical Application. *Brief. Funct. Genomics* **2016**, *15*, 374–384. [CrossRef] [PubMed]
56. Huppert, A.; Katriel, G. Mathematical Modelling and Prediction in Infectious Disease Epidemiology. *Clin. Microbiol. Infect.* **2013**, *19*, 999–1005. [CrossRef] [PubMed]
57. Paulson, R.J. Mathematics Should Clarify, Not Obfuscate: An Inaccurate and Misleading Calculation of the Cost-Effectiveness of Preimplantation Genetic Testing for Aneuploidy. *Fertil. Steril.* **2019**, *111*, 1113–1114. [CrossRef]
58. Cohen, D.P.A.; Martignetti, L.; Robine, S.; Barillot, E.; Zinovyev, A.; Calzone, L. Mathematical Modelling of Molecular Pathways Enabling Tumour Cell Invasion and Migration. *PLoS Comput. Biol.* **2015**, *11*, e1004571. [CrossRef]
59. Telikani, A.; Gandomi, A.H.; Tahmassebi, A.; Banzhaf, W. Evolutionary Machine Learning: A Survey. *ACM Comput. Surv* **2021**, *54*, 1–35. [CrossRef]
60. Kotsiantis, S.B. Supervised Machine Learning: A Review of Classification Techniques. *Informatica* **2007**, *31*, 249–268. [CrossRef]
61. Le, N.Q.K.; Ho, Q.-T. Deep Transformers and Convolutional Neural Network in Identifying DNA N6-Methyladenine Sites in Cross-Species Genomes. *Methods* **2022**, *204*, 199–206. [CrossRef]
62. Tng, S.S.; Le, N.Q.K.; Yeh, H.-Y.; Chua, M.C.H. Improved Prediction Model of Protein Lysine Crotonylation Sites Using Bidirectional Recurrent Neural Networks. *J. Proteome Res.* **2022**, *21*, 265–273. [CrossRef]
63. Olson, R.S.; Urbanowicz, R.J.; Andrews, P.C.; Lavender, N.A.; Kidd, L.C.; Moore, J.H. Automating Biomedical Data Science Through Tree-Based Pipeline Optimization. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Cham, Switzerland, 2016; Volume 9597, pp. 123–137. ISBN 9783319312033.
64. Le, T.T.; Fu, W.; Moore, J.H. Scaling Tree-Based Automated Machine Learning to Biomedical Big Data with a Feature Set Selector. *Bioinformatics* **2020**, *36*, 250–256. [CrossRef] [PubMed]
65. Matejka, J.; Fitzmaurice, G. Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, 6–11 May 2017; pp. 1–5. [CrossRef]
66. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer Statistics, 2019. *CA. Cancer J. Clin.* **2019**, *69*, 7–34. [CrossRef] [PubMed]
67. Morris, J.S.; Brown, P.J.; Herrick, R.C.; Baggerly, K.A.; Coombes, K.R.; Morris jeffmo, J.S. Bayesian Analysis of Mass Spectrometry Proteomics Data Using Wavelet Based Functional Mixed Models. *Biometrics* **2008**, *2*, 479–489. [CrossRef] [PubMed]
68. Eberhard, O. *Voit Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists*; Cambridge University Press: Cambridge, UK, 2000.

69. Schlatter, R.; Schmich, K.; Avalos Vizcarra, I.; Scheurich, P.; Sauter, T.; Borner, C.; Ederer, M.; Merfort, I.; Sawodny, O. ON/OFF and beyond—A Boolean Model of Apoptosis. *PLoS Comput. Biol.* **2009**, *5*, e1000595. [CrossRef]
70. Rateitschak, K.; Kaderali, L.; Wolkenhauer, O.; Jaster, R. Autocrine TGF-β/ZEB/MicroRNA-200 Signal Transduction Drives Epithelial-Mesenchymal Transition: Kinetic Models Predict Minimal Drug Dose to Inhibit Metastasis. *Cell. Signal.* **2016**, *28*, 861–870. [CrossRef]
71. Fumiã, H.F.; Martins, M.L. Boolean Network Model for Cancer Pathways: Predicting Carcinogenesis and Targeted Therapy Outcomes. *PLoS ONE* **2013**, *8*, e69008. [CrossRef]
72. Arellano, A.M.; Dai, W.; Wang, S.; Jiang, X.; Ohno-Machado, L. Privacy Policy and Technology in Biomedical Data Science. *Annu. Rev. Biomed. Data Sci.* **2018**, *1*, 115–129. [CrossRef]