PSL UNIVERSITÉ PARIS

**THÈSE DE DOCTORAT**

**DE L'UNIVERSITÉ PSL**

Préparée à l'Institut Curie

# From the mechanistic modeling of signaling pathways in cancer to the interpretation of models and their contributions: clinical applications and statistical evaluation

Soutenance prévue par

**Jonas BEAL**

Le 23/09/2020

École doctorale n°515

**Complexité du Vivant**

Spécialité

**Génomique**

Composition du jury :

| | |
|---|---|
| Adeline LECLERQ-SAMSON<br>Professeur, Université Grenoble Alpes | *Rapporteur* |
| Lodewyk WESSELS<br>Professeur, Netherlands Cancer Institute | *Rapporteur* |
| Émilie LANOY<br>Ingénieure de recherche, Gustave Roussy | *Examinateur* |
| Denis THIEFFRY<br>Professeur, ENS, PSL | *Examinateur* |
| Emmanuel BARILLOT<br>Directeur de Recherche, Institut Curie, PSL | *Directeur de thèse* |
| Aurélien LATOUCHE<br>Professeur, Institut Curie, Cnam | *Directeur de thèse* |
| Laurence CALZONE<br>Ingénieure de recherche, Institut Curie, PSL | *Membre invitée, co-encadrante de thèse* |

institut Curie | PSL

*Ce sont les sévères artistes*
*Que l'aube attire à ses blancheurs,*
*Les savants, les inventeurs tristes,*
*Les puiseurs d'ombre, les chercheurs,*
*Qui ramassent dans les ténèbres*
*Les faits, les chiffres, les algèbres,*
*Le nombre où tout est contenu,*
*Le doute où nos calculs succombent,*
*Et tous les morceaux noirs qui tombent*
*Du grand fronton de l'inconnu !*
Victor Hugo (*Les mages*)

# Abstract

Beyond its genetic mechanisms, cancer can be understood as a network disease that often results from the interactions between different perturbations in a cellular regulatory network. The dynamics of these networks and associated signaling pathways are complex and require integrated approaches. One approach is to design mechanistic models that translate the biological knowledge of networks in mathematical terms to simulate computationally the molecular features of cancers. However, these models only reflect the general mechanisms at work in cancers.

This thesis proposes to define personalized mechanistic models of cancer. A generic model is first defined in a logical (or Boolean) formalism, before using omics data (mutations, RNA, proteins) from patients or cell lines in order to make the model specific to each one profile. These personalized models can then be compared with the clinical data of patients in order to validate them. The response to treatment is investigated in particular in this thesis. The explicit representation of the molecular mechanisms by these models allows to simulate the effect of different treatments according to their targets and to verify if the sensitivity of a patient to a drug is well predicted by the corresponding personalized model. An example concerning the response to BRAF inhibitors in melanomas and colorectal cancers is thus presented.

The comparison of mechanistic models of cancer, those presented in this thesis and others, with clinical data also encourages a rigorous evaluation of their possible benefits in the context of medical use. The quantification and interpretation of the prognostic value of outputs of some mechanistic models is briefly presented before focusing on the particular case of models able to recommend the best treatment for each patient according to his molecular profile. A theoretical framework is defined to extend causal inference methods to the evaluation of such precision medicine algorithms. An illustration is provided using simulated data and patient derived xenografts.

All the methods and applications put forward a possible path from the design of mechanistic models of cancer to their evaluation using statistical models emulating clinical trials. As such, this thesis provides one framework for the implementation of precision medicine in oncology.

**Key-words**: Modeling, Cancer, Mechanistic model, Biostatistics, Causal inference, Precision medicine

# Résumé

Au delà de ses mécanismes génétiques, le cancer peut être compris comme une maladie de réseaux qui résulte souvent de l'interaction entre différentes perturbations dans un réseau de régulation cellulaire. La dynamique de ces réseaux et des voies de signalisation associées est complexe et requiert des approches intégrées. Une d'entre elles est la conception de modèles dits mécanistiques qui traduisent mathématiquement la connaissance biologique des réseaux afin de pouvoir simuler le fonctionnement moléculaire des cancers informatiquement. Ces modèles ne traduisent cependant que les mécanismes généraux à l'oeuvre dans certains cancers en particulier.

Cette thèse propose en premier lieu de définir des modèles mécanistiques personnalisés de cancer. Un modèle générique est d'abord défini dans un formalisme logique (ou Booléen), avant d'utiliser les données omiques (mutations, ARN, protéines) de patients ou de lignées cellulaires afin de rendre le modèle spécifique à chacun. Ces modèles personnalisés peuvent ensuite être confrontés aux données cliniques de patients pour vérifier leur validité. Le cas de la réponse clinique aux traitements est exploré en particulier dans ce travail. La représentation explicite des mécanismes moléculaires par ces modèles permet en effet de simuler l'effet de différents traitements suivant leur mode d'action et de vérifier si la sensibilité d'un patient à un traitement est bien prédite par le modèle personnalisé correspondant. Un exemple concernant la réponse aux inhibiteurs de BRAF dans les mélanomes et cancers colorectaux est ainsi proposé.

La confrontation des modèles mécanistiques de cancer, ceux présentés dans cette thèse et d'autres, aux données cliniques incite par ailleurs à évaluer rigoureusement leurs éventuels bénéfices dans la cadre d'une utilisation médicale. La quantification et l'interprétation de la valeur pronostique des biomarqueurs issus de certains modèles méchanistiques est brièvement présentée avant de se focaliser sur le cas particulier des modèles capables

de sélectionner le meilleur traitement pour chaque patient en fonction des ses caractéristiques moléculaires. Un cadre théorique est proposé pour étendre les méthodes d'inférence causale à l'évaluation de tels algorithmes de médecine de précision. Une illustration est fournie à l'aide de données simulées et de xénogreffes dérivées de patients.

L'ensemble des méthodes et applications décrites tracent donc un chemin, de la conception de modèles mécanistiques de cancer à leur évaluation grâce à des modèles statistiques émulant des essais cliniques, proposant ainsi un cadre pour la mise en oeuvre de la médecine de précision en oncologie.

**Mots-clés**: Modélisation, Cancer, Modèle mécanistique, Biostatistiques, Inférence causale, Médecine de précision

# Remerciements

On ne se lance pas dans une thèse dans le seul but d'en écrire les remerciements. C'est peut être un tort tant l'occasion est belle d'en profiter pour exprimer sa gratitude à tous ceux qui la méritent sans que l'on pense toujours à le leur dire. Je vais donc essayer de remercier tous ceux qui occupent mon esprit à l'heure où j'écris ces lignes, d'un bout à l'autre du spectre, de ceux sans qui cette thèse n'aurait pas été la même à ceux sans qui je serais bien différent.

Merci tout d'abord aux membres du jury, Denis Thieffry, Émilie Lanoy, et en particulier aux rapporteurs Adeline Leclerq-Samson et Lodewyk Wessels, d'avoir accepté d'évaluer ce travail, contribuant ainsi à le rendre plus enrichissant pour moi. Merci également aux membres de mes comités de suivi de thèse Benno Schwikowski, Mélanie Prague et Xavier Paoletti pour leur regard extérieur et leurs remarques pertinentes et constructives.

Mes remerciements suivants, à la fois professionnels et personnels vont à Laurence, Aurélien et Emmanuel pour avoir supervisé cette thèse en m'orientant sans me diriger, dans un juste équilibre d'encadrement et de liberté, chacun suivant ses spécificités. Merci en particulier à Laurence pour son inépuisable bienveillance, prodiguée avec générosité. Merci à Aurélien pour sa disponibilité et sa positivité, mot piégé s'il en est. Et enfin merci à Emmanuel pour sa capacité à prendre du recul.

Merci ensuite à tous les collègues de l'équipe de biologie des systèmes, présents ou passés, particulièrement ceux avec lesquels j'ai pu collaborer directement. Merci à Arnau pour m'avoir inoculé son souci excessif de certains détails, à Vincent pour m'avoir enseigné certaines bonnes pratiques, à Pauline pour m'avoir accueilli à l'Institut et à Lorenzo pour m'avoir supporté. Merci plus largement à tous les membres de l'équipe qui ont fait de cette thèse une période d'échanges, scientifiques ou non : Nicolas, Mihaly, Cristobal, Andrei, Urczula, Céline, Laura, Loredana, Jane, Om, Marianyela,

Christine, Maria, Inna. Mention spéciale à mes collègues de bureau successifs Loïc et Jonathan pour leur humour bienvenu et pour avoir toléré le mien. Merci de la même manière aux collègues de l'équipe STAMPM, tout spécialement à Bassirou et Alessandra pour m'avoir accueilli dans leur bureau mouvant de Saint-Cloud, mais aussi Christophe et Xavier (à nouveau) pour leur écoute et leurs conseils. Merci aux autres membres de l'U900, avec une pensée particulière pour Caroline, Kati et Yasmina dont le sens de l'organisation a contribué à me faciliter la vie.

Merci également à tous ces collègues inconnus qui partagent librement connaissances, tutoriels, code informatique ou modèles typographiques. Ils forment une communauté virtuelle à l'impact bien réel.

Merci aussi à ceux qui ont contribué à orienter mon parcours au fil des années, de Polytechnique à la biologie computationnelle. Je pense notamment à François et Fred chez Novadiscovery ainsi qu'à Vassily et Misko à l'EPFL.

De manière toujours plus personnelle, je me dois aussi de remercier tous les amis qui ont fait de ces trois années un voyage agréable. Au sein de l'Institut tout d'abord, pour quelques repas avec Élise et Anne, au Collège des Ingénieurs, pour des discussions intellectuellement variées et rafraichissantes. Merci ensuite à tous ces amis, connus à Saint-Étienne, Lyon, Palaiseau, Paris ou Lausanne, que se reconnaitront sans qu'il soit nécessaire de tous les nommer exhaustivement. Mention spéciale à Guillaume, Laurent, Reda, Michaël, Pierre, Etienne, Florian et Amélie pour m'avoir aidé ou réjoui plus souvent qu'à leur tour.

Une grande partie de la gratitude restante à ce stade revient à ma famille pour son soutien sans faille. Merci à mes parents, frère, soeur et grand-mère pour m'avoir accompagné tout au long de mes études de leur curiosité et de leur fierté, autant de vent dans mes voiles sans lequel je n'aurais pas tant avancé. Merci tout spécialement à mes parents, je dois à votre tranquille mais tenace confiance de n'avoir jamais douté de ma capacité à choisir ma voie et à la tracer avec succès.

Enfin merci à Honorine pour tout le reste, c'est à dire pour l'essentiel. Si "*on ne possède d'un être que ce qu'on change en lui*"[1], ne doute jamais qu'une large portion de ce travail t'appartient. Son auteur quant à lui ne s'appartient déjà plus tout à fait…

[1] André Malraux, *La condition humaine*

# Preface

The present thesis is structured in three parts, each subdivided into three chapters. Since the whole thesis is about cancer modeling, the first part aims at defining the type of model to be referred to, and in particular models that will be called mechanistic, as well as the object of the modeling, *i.e.*, the molecular networks involved in cancer. So the first part answers the question: **what is a cancer model and what is its purpose?**

The second part will be devoted to the methods developed during this thesis to transform qualitative models of molecular networks, known as logical models, into personalized models that can be interpreted clinically. In short, **how can a mathematical representation of biological knowledge be transformed into a tool that contributes to the understanding of the clinical manifestations of cancer?**

Finally, the third and last part will look at how the clinical relevance of all the above-mentioned models can be rigorously evaluated, both in their ability to predict the evolution of the disease and in their ability to recommend the most appropriate treatments for each patient. **How to quantify and interpret the value of the clinical information delivered by these models?**

Moreover, this thesis also exists in an online version that allows to take advantage of the interactivity of some graphs and applications: https://jonasbeal.github.io/files/PhdThesis/.

**Scientific content**

Except for the first part, essentially introductory and based on scientific literature, the different chapters are based on original scientific work done during this thesis (2017-2020) and mentioned at the beginning of each chapter in a box similar to the this one.
The main articles behind this thesis are indicated below with one published article and two pre-prints currently under review:

- Béal, Jonas, Arnau Montagud, Pauline Traynard, Emmanuel Barillot, and Laurence Calzone. "Personalization of logical models with multi-omics data allows clinical stratification of patients." Frontiers in physiology 9 (2019): 1965. Link.
- Béal, Jonas, Lorenzo Pantolini, Vincent Noël, Emmanuel Barillot, and Laurence Calzone. "Personalized logical models to investigate cancer response to BRAF treatments in melanomas and colorectal cancers." bioRxiv (2020). Link.
- Béal, Jonas, and Aurélien Latouche. "Causal inference with multiple versions of treatment and application to personalized medicine." arXiv preprint arXiv:2005.12427 (2020). Link.

These three articles were described or completed in oral presentations, respectively in International Conference of Systems Biology 2018, conference on Intelligent Systems for Molecular Biology (ISMB/ECCB 2019, Video) and conference of International Society of Clincial Biostatistics (ISCB41, coming in August 2020).

# Table of contents

# List of Tables

**Table** **Page**

# List of Figures

# Part I

# Cells and their models

Chapter 1

# Scientific modeling: abstract the complexity

*"Ce qui est simple est toujours faux. Ce qui ne l'est pas est inutilisable."*
*Paul Valéry (Mauvaises pensées et autres, 1942)*

The notion of modeling is embedded in science, to the point that it has sometimes been used to define the very nature of scientific research. What is called a model can, however, correspond to very different realities which need to be defined before addressing the object of this thesis which will consist, if one wants to be mischievous, in analyzing models with other models. This semantic elucidation is all the more necessary as this thesis is interdisciplinary, suspended between systems biology and biostatistics. In order to convince the reader of the need for such a preamble, he is invited to ask a statistician and a biologist how they would define what a model is.

Figure 1.1: **A scientist and his model.** Joseph Wright of Derby, *A Philosopher Giving a Lecture at the Orrery (in which a lamp is put in place of the sun)*, c. 1763-65, oil on canvas, Derby Museums and Art Gallery

## 1.1 What is a model?

### 1.1.1 In your own words

A model is first of all an ambiguous object and a polysemous word. It therefore seems necessary to start with a semantic study. Among the many meanings and synonymous proposed by the dictionary (Figure 1.2), while some definitions are more related to art, several find echoes in scientific practice. It is sometimes a question of the physical representation of an object, often on a reduced scale as in Figure 1.1, and sometimes of a theoretical description intended to facilitate the understanding of the way in which a system works [Collins, 2020]. It is even sometimes an ideal to be reached and therefore an ambitious prospect for an introduction.

The narrower perspective of the scientist does not reduce the completeness of the dictionary's description to an unambiguous object [Bailer-Jones, 2002]. In an attempt to approach these multi-faceted objects that are the models, Daniela Bailer-Jones interviewed different scientists and asked them the same question: what is a model? Across the different profiles and fields

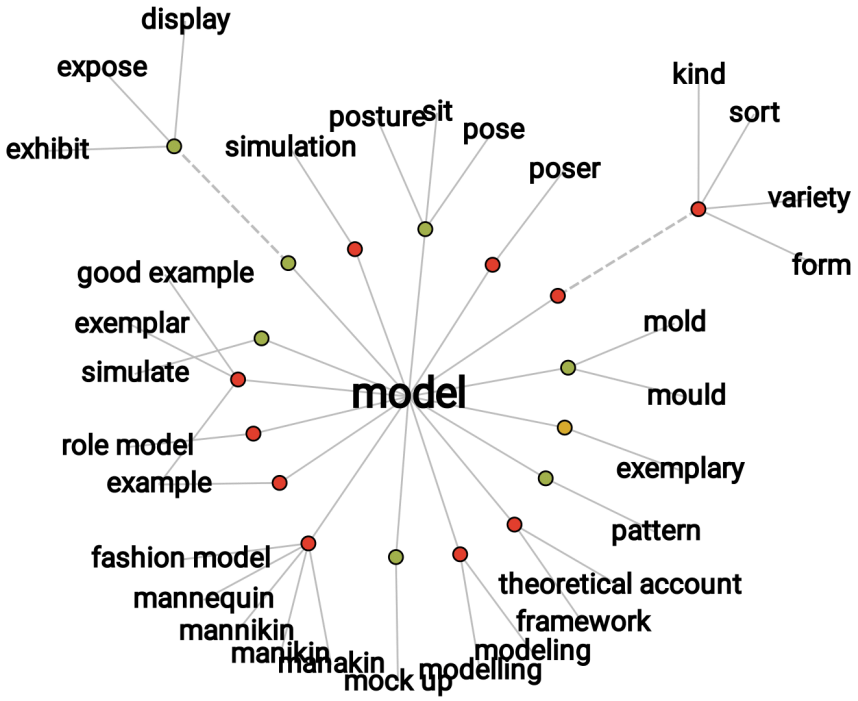



Figure 1.2: **Network visualization of *model* thesaurus entries.** Generated with the 'Visual Thesaurus' ressource

of study, the answers vary but some patterns begin to emerge (Figure 1.3). A model must capture the essence of the phenomenon being studied. Because it eludes, voluntarily or not, many details or complexity, it is by nature a simplification of the phenomenon. These limitations may restrict its validity to certain cases or suspend it to the fulfilment of some hypotheses. They are not necessarily predictive, but they must be able to generate new hypotheses, be tested and possibly questioned. Finally, and fundamentally, they **must provide insights about the object of study and contribute to its understanding**.

These definitions circumscribe the *model* object, its use and its objectives, but they do not in any way describe its nature. And for good reason, because even if we agree on the described contours, the biodiversity of the models remains overwhelming for taxonomists:

> *Probing models, phenomenological models, computational models, developmental models, explanatory models, impoverished*

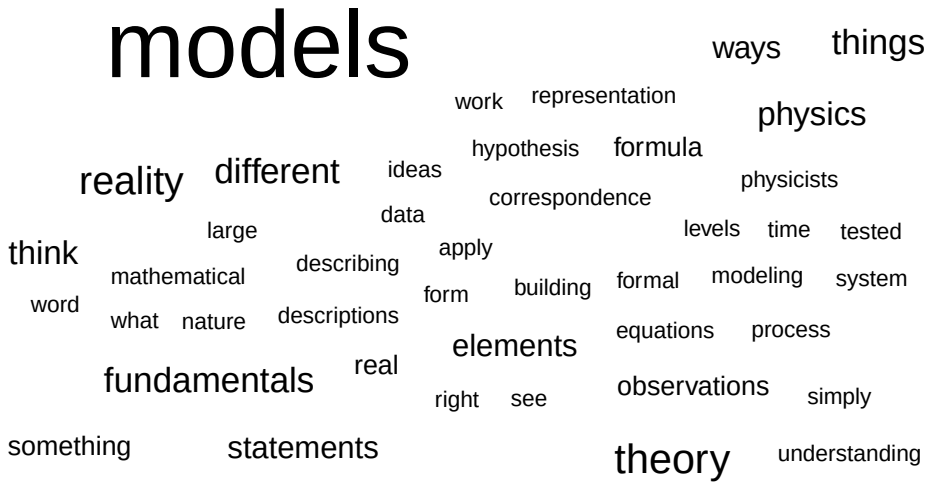



Figure 1.3: **Scientists talk about their models: words cloud.** Cloud of words summarizing the lexical fields used by scientists to talk about their models in dedicated interviews reported by Bailer-Jones [2002].

> *models, testing models, idealized models, theoretical models, scale models, heuristic models, caricature models, exploratory models, didactic models, fantasy models, minimal models, toy models, imaginary models, mathematical models, mechanistic models, substitute models, iconic models, formal models, analogue models, and instrumental models are but some of the notions that are used to categorize models.*
> [Frigg and Hartmann, 2020]

### 1.1.2 Physical world and world of ideas

Without claiming to be exhaustive, we can make a **first simple dichotomy between physical/material and formal/intellectual models** [Rosenblueth and Wiener, 1945]. The former consists in replacing the object of study by another object, just as physical but nevertheless simpler or better known. These may be models involving a change of scale such as the simple miniature replica placed in a wind tunnel, or the metal double helix model used by Watson and Crick to visualize DNA. In all these cases the model allows to visualize the object of study (Figure 1.4 A and B), to manipulate it and play with it to better understand or explain a phenomenon, just like the scientist with his orrery (Figure 1.1). In the case of biology, there are mainly model organisms such as drosophila, zebrafish or mice, for example. We then benefit from the relative simplicity of their genomes, a shorter

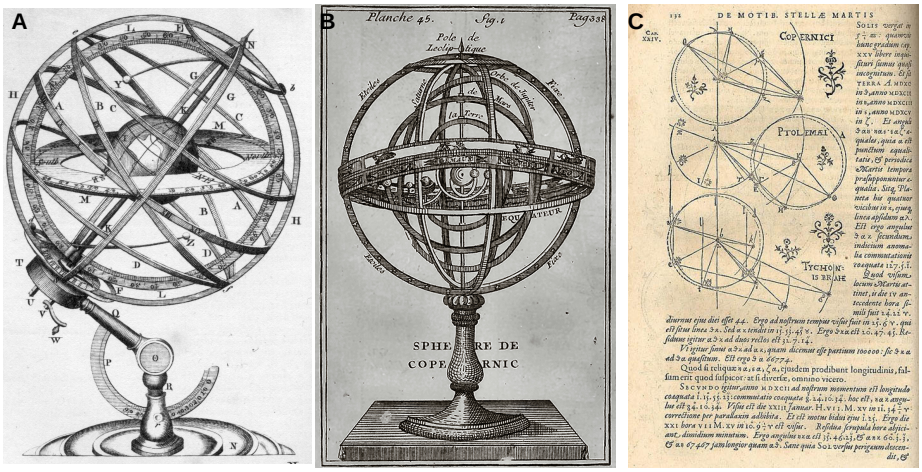



Figure 1.4: **Orrery, planets and models**. Physical models of planetary motion, either geocentric (Armillary sphere from *Plate LXXVII* in *Encyclopedia Britannica*, 1771) or heliocentric in panel B (Bion, 1751, catalogue Bnf) and some geometric representations by Johannes Kepler in panel C (in *Astronomia Nova*, 1609)

time scale or ethical differences, usually to elucidate mechanisms of interest in humans. Correspondence between the target system and its model can sometimes be more conceptual, such as that ones relying on mechanical–electrical analogies: a mechanical system (e.g. a spring-mass system) can sometimes be represented by an electric network (e.g. a RLC circuit with a resistor, a capacitor and an inductor).

The model is then no longer simply a mimetic replica but is based on an intellectual equivalence: we are gradually moving into the realm of formal models [Rosenblueth and Wiener, 1945]. These are of a more symbolic nature and they **represent the original system with a set of logical or mathematical terms**, describing the main driving forces or similar structural properties as geometrical models of planetary motions summarized by Kepler in Figure 1.4C. Historically these models have often been expressed by sets of mathematical equations or relationships. Increasingly, these have been implemented by computer. Despite their sometimes less analytical and more numerical nature, many so-called computational models could also belong to this category of formal models. There are then many formalisms, discrete or continuous, deterministic or stochastic, based on differential equations or Boolean algebra [Fowler et al., 1997]. Despite their more abstract nature, they offer similar scientific services: it is possi-

ble to play with their parameters, specifications or boundary conditions in order to better understand the phenomenon. One can also imagine these formal models from a different perspective, which starts from the data in a bottom-up approach instead of starting from the phenomenon in a top-down analysis. These models will then often be called statistical models or models of data [Frigg and Hartmann, 2020]. This distinction will be further clarified in section 1.2.

To summarize and continue a little longer with the astronomical metaphor, the study of a particularly complex system (the solar system) can be broken down into a variety of different models. Physical and mechanical models such as armillary spheres (1.4A and B) make it possible to touch the object of study. In addition, we can observe the evolution of models which, when confronted with data, have progressed from a geocentric to a heliocentric representation to get closer to the current state of knowledge. Sometimes, models with more formal representations are used to give substance to ideas and hypotheses (1.4C). One of the most conceptual forms is then the mathematical language and one can thus consider that the previously mentioned astronomical models find their culmination in Kepler's equations about orbits, areas and periods that describe the elliptical motion of the planets. We refer to them today as Kepler's laws. The model has become a law and therefore a paragon of mathematical modeling [Wan, 2018].

### 1.1.3 Preview about cancer models

As we get closer to the subject of our study, and in order to illustrate these definitions more concretely, we can take an interest in the meaning of the word *model* in the context of cancer research. For this, we restrict our corpus to scientific articles found when searching for "cancer model" in the PubMed article database. Among these, we look at the occurrences of the word *model* and the sentences in which it is included. This cancer-related context of model is represented as a tree in Figure 1.5. Some of the distinctions already mentioned can be found here. The *mouse* and *xenograft* models, which will be discussed later in this thesis, represent some of the most common physical models in cancer studies. These are animal models in which the occurrence and mechanisms of cancer, usually induced by the biologist, are studied. On the other hand, *prediction*, *prognostic* or *risk score* models refer to formal models and borrow from statistical language.

Another way to classify cancer models may be to group them into the fol-

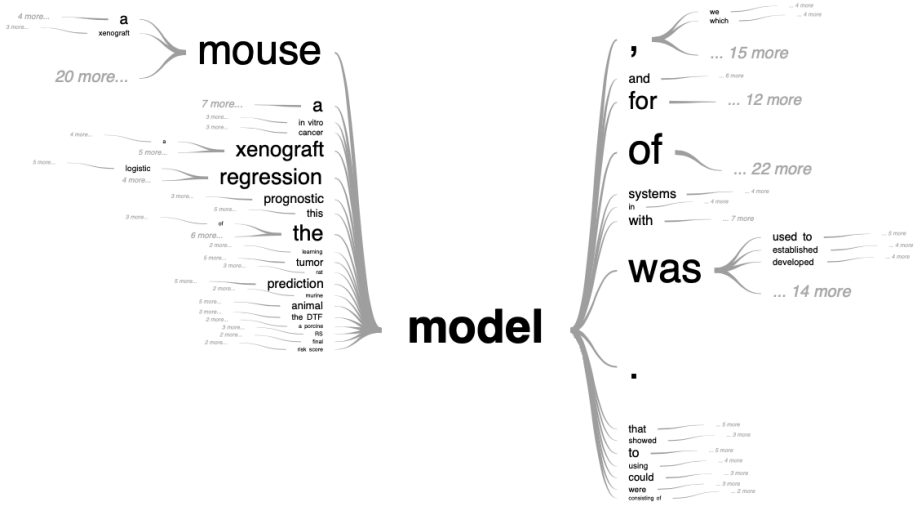



Figure 1.5: **Tree visualization of *model* semantic context in cancer-related literature** Generated with the 'PubTrees' tool by Ed Sperr, and based on most relevant PubMed entries for "cancer model" search.

lowing categories: *in vivo*, *in vitro* and *in silico*. The first two clearly belong to the physical models but one uses whole living organisms (e.g. a human tumor implanted in an immunodeficient mouse) and the other separates the living from its organism in order to place it in a controlled environment (e.g. tumor cells in growth medium in a Petri dish). **In the thesis, data from both *in vivo* and *in vitro* models will be used. However, unless otherwise stated, a model will always refer to a representation *in silico*.** This third category, however, contains a very wide variety of models [Deisboeck et al., 2009], to which we will come back in chapter 3. A final ambiguity about the nature of the formal models used in this thesis needs to be clarified beforehand.

## 1.2 Statistics or mechanistic

A rather frequent metaphor is to compare formal models to black boxes that take in input $X$ predictors, or independent variables, and output response variable(s) $Y$, also named dependent variables. The models then split into two categories (Figure 1.6) depending on the answer to the question: are you modeling the inside of the box or not?

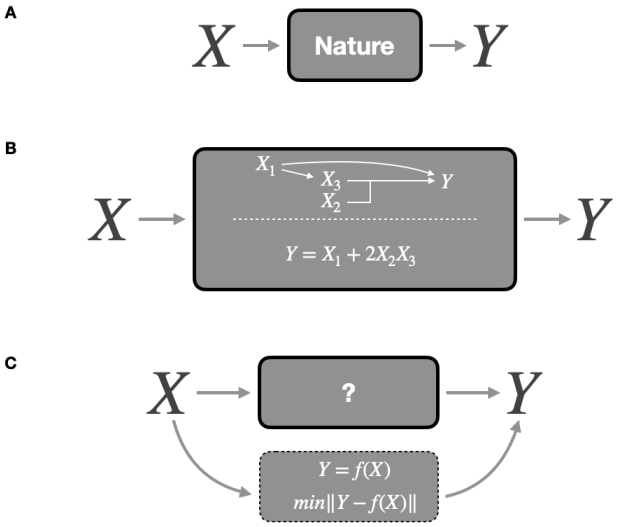



Figure 1.6: **Different modeling strategies.** (A) Data generation from predictors $X$ to response $Y$ in the natural phenomenon. (B) Mechanistic modeling defining mechanisms of data generation inside the box. (C) Statistical modeling finding the function $f$ that gives the best predictions. Adapted from Breiman [2001b].

### 1.2.1 The inside of the box

The purpose of this section is to present in a schematic, and therefore somewhat caricatural, manner the two competing formal modeling approaches that will be used in this thesis and that we will call mechanistic modeling and statistical modeling. Assuming the unambiguous nature of the predictors and outputs we can imagine that the natural process consists in defining the result Y from the inputs X according to a function of a completely unknown form (Figure 1.6A).

The first modeling approach, that we will call **mechanistic**, consists in **building the box by imitating what we think is the process of data generation**, or in other words, by representing the mechanisms at work (Figure 1.6B). This integration of prior knowledge can take different forms. In this thesis it will often come back to presupposing certain relations between entities according to what is known about their behaviour. $X_1$ which acts on $X_3$ may correspond to the action of one biological entity on another, supposedly unidirectional; just as the joint action of $X_2$ and $X_3$ may reflect a known synergy in the expression of genes or the action of proteins. Mathematically this is expressed here with a perfectly deterministic

model defined *a priori*. All in all, in a purely mechanistic approach, the nature of the relations between entities should be linked to biological processes and the parameters in the model all have biological definitions in such a way that it could even be considered to measure them directly. For example, the coefficient 2 multiplying $X_2X_3$ can correspond to a stoichiometric coefficient or a reaction constant which have a theoretical justification or are accessible by experimentation. In some fields of literature these models are sometimes called mathematical models because they propose a mathematical translation of a phenomenon, which does not start from the data in a bottom-up approach but rather from a top-down theoretical framework. In this thesis we will adhere to the *mechanistic model* name, which is more transparent and less ambiguous compared to other approaches also based on mathematics, without necessarily the other characteristics described above.

The second approach, often called **statistical modeling**, or sometimes machine learning depending on the precise context and objective, does not necessarily seek to reproduce the natural process of data generation but to **find the function allowing the best prediction of $Y$ from $X$** (Figure 1.6C). Pushed to the limit, they are an "idealized version of the data we gain from immediate observation" [Frigg and Hartmann, 2020], thus providing a phenomenological description. The methods and algorithms used are then intended to be sufficiently flexible and to make the fewest possible assumptions about the relationships between variables or the distribution of data. Without listing them exhaustively, the approaches such as simple linear regressions or more complex support vector machines [Cortes and Vapnik, 1995] or random forests [Breiman, 2001a], which will sometimes be mentioned in this thesis, fall into this category which contains many others [Hastie et al., 2009].

Several discrepancies result from this difference in nature between mechanistic and statistical models, some of which are summarized in the Table 1.1. In a somewhat schematic way, we can say that the mechanistic model first asks the question of *how* and then looks at the result for the output. The **notion of causality is intrinsic to the definition of the model**. Conversely, the statistical model first tries to approach the Y and then possibly analyses what can be deduced from it, regarding the importance of the variables or their relationships in a *post hoc* approach [Ishwaran, 2007, Manica et al., 2019]. The causality is then not a by-product of the algorithm and must be evaluated with dedicated frameworks [Hernán and Robins, 2020]. The greater flexibility of statistical methods makes it possible to better accept the heterogeneity of the variables, but this is generally done at the

Table 1.1: **Some pros and cons for mechanistic and statistical modeling**. Adapted from Baker et al. [2018].

| Mechanistic modeling | Statistical modeling |
|---|---|
| **Definition** | |
| Seeks to establish a mechanistic relationship between inputs and outputs | Seeks to establish statistical relationships between inputs and outputs |
| **Pros and cons** | |
| Presupposes and investigates causal links between the variables | Looks for patterns and establishes correlations between variables |
| Capable of handling small datasets | Requires large datasets |
| Once validated, can be used as a predictive tool in new situations possibly difficult to access through experimentation | Can only make predictions that relate to patterns within the data supplied |
| Difficult to accurately incorporate information from multiple space and time scales due to constrained specifications | Can tackle problems with multiple space and time scales thanks to flexible specifications |
| Evaluated on closeness to data and ability to make sense of it | Evaluated based on predictive performance |

cost of a larger number of parameters and therefore requires more data. Moreover, statistical models can be considered as inductive, since they are able to use already generated data to identify patterns in it. Conversely, mechanistic models are more deductive and they can theoretically allow to extrapolate beyond the original data or knowledge used to build the model [Baker et al., 2018]. Finally, the most relevant way of assessing the value or adequacy of these models may be quite different. A statistical model is measured by its ability to predict output in a validation dataset different from the one used to train its parameters. The mechanistic model will also be evaluated on its capacity to approach the data but also to order it, to give a meaning. If its pure predictive performance is generally inferior, **how can the value of understanding be assessed?** This question will be one of the threads of the dissertation.

Mechanistic and statistical models are not perfectly exclusive and rather form the two ends of a spectrum. The definitions and classification of some

examples is therefore still partly personal and arbitrary. For instance, the example in 1.6B can be transformed into a model with a more ambiguous status:

$$logit(P[Y = 1]) = \beta_1 X_1 + \beta_{23} X_2 X_3$$

This model is deliberately ambiguous. As a logistic model, it is therefore naturally defined as a statistical model. But the definition of the interaction between $X_2$ and $X_3$ denotes a mechanistic presupposition. The very choice of a logistic and therefore parametric model could also result from a knowledge of the phenomenon, even if in practice it is often a default choice for a binary output. Finally, the nature of the parameters $\beta_1$ and $\beta_{23}$ is likely to change the interpretation of the model. If they are deduced from the data and therefore optimized to fit Y as well as possible, one will think of a statistical model whose specification is nevertheless based on knowledge of the phenomenon. On the other hand, one could imagine that these parameters are taken from the biochemistry literature or other data. The model will then be more mechanistic. The boundary between these models is further blurred by the different possibilities of combining these approaches and making them complementary [Baker et al., 2018, Salvucci et al., 2019a].

### 1.2.2 A tale of prey and predators

The following is a final general illustration of the concepts and procedures introduced with respect to statistical and mechanistic models through a famous and characteristic example: the Lotka-Volterra model of interactions between prey and predators. This model was, like many students, my first encounter with what could be called mathematical biology. The Italian mathematician Vito Volterra states this system for the first time studying the unexpected characteristics of fish populations in the Adriatic Sea after the First World War. Interestingly, Alfred Lotka, an American physicist deduced the exact same system independantly, starting from very generic process of redistribution of matter among the several components derived from law of mass action [Knuuttila and Loettgers, 2017]. A detailed description of their works and historical formulation can be found in original articles [Lotka, 1925, Volterra, 1926] or dedicated reviews [Knuuttila and Loettgers, 2017].

The general objective is to understand the evolution of the populations

of a species of prey and its predator, reasonably isolated from outside intervention. Here we will use Canada lynx (*Lynx canadensis*) and snowshow hare (*Lepus americanus*) populations for which an illustrative data set exists [Hewitt, 1917]. In fact, commercial records listing the quantities of furs sold by trappers to the Canadian Hudson Bay Company may represent a proxy for the populations of these two species as represented in Figure 1.7A. Denoting the population of lynx $L(t)$ and the population of hare $H(t)$ it can be hypothesized that prey, in the absence of predators, would increase in population, while predators on their own would decline in the absence of prey. A prey/predator interaction term can then be added, which will positively impact predators and negatively impact prey. The system can then be formalized with the following differential questions with all coefficients $a_1, a_2, b_1, b_2 > 0$:

$$\frac{dH}{dt} = a_1 H - a_2 HL$$

$$\frac{dL}{dt} = -b_1 L + b_2 HL$$

$a_1 H$ represents the growth rate of the hare population (prey), i.e., the population grows in proportion to the population itself according to usual birth modeling. The main losses of hares are due to predation by lynx, as represented with a negative coefficient in the $-a_2 HT$ term. It is therefore assumed that a fixed percentage of prey-predator encounters will result in the death of the prey. Conversely, it is assumed that the growth of the lynx population depends primarily on the availability of food for all lynxes, summarized in the $b_2 HL$ term. In the absence of hares, the lynx population decreases, as denoted by the coefficient $-b_1 L$. Important features of mechanistic models are illustrated here: the equations are based on *a priori* knowledge or assumptions about the structure of the problem and the parameters of the model can be interpreted. $a_1$, for example, could correspond to the frequency of litters among hares and the number of offspring per litter.

This being said, the structure of the model having been defined *a priori*, it remains to determine its parameters. Two options would theoretically be possible: to propose values based on the interpretation of the parameters and ecological knowledge, or to fit the model to the data in order to find the best parameters. For the sake of simplicity, and because this example

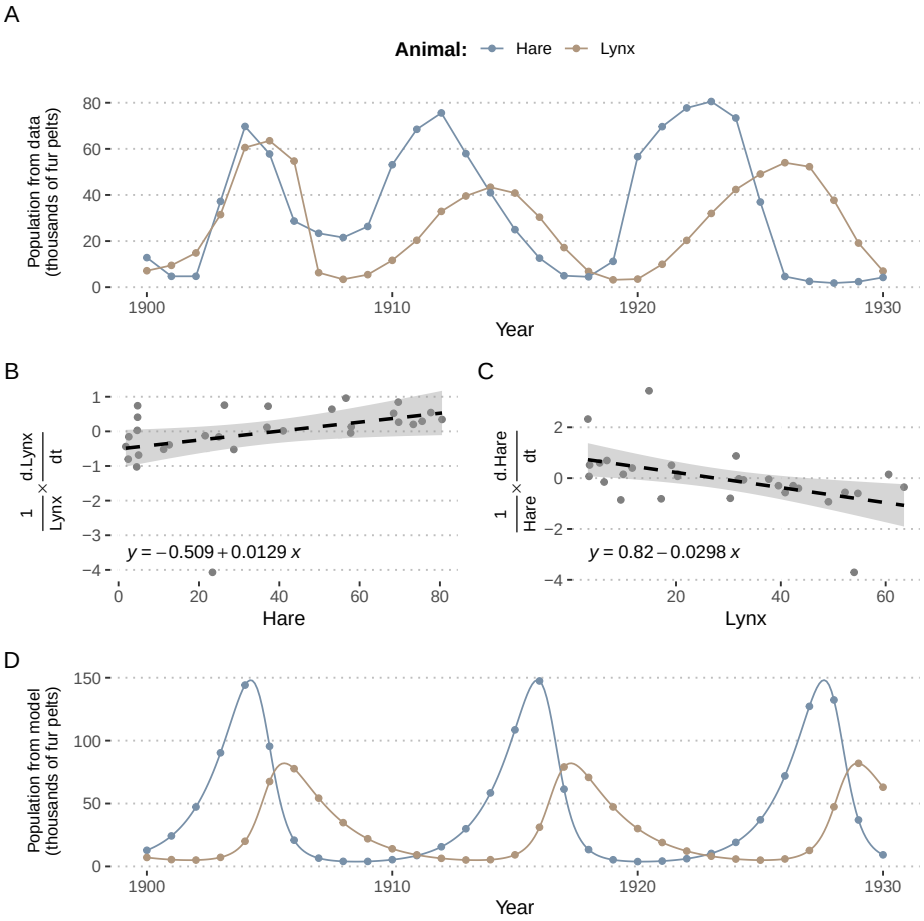



Figure 1.7: **Some analyses around Lotka-Volterra model of a prey-predator system**. (A) Evolution of lynx and hares populations based on Hudson Bay Company data about fur pelts. (B) and (C) Linear regression for estimation of parameters. (D) Evolution of lynx and hare populations as predicted by the model based on inferred parameters and initial conditions.

has only a pedagogical value in this presentation, we propose to determine them approximately using the following Taylor-based approximation:

$$\frac{1}{y(t)}\frac{dy}{dt} \simeq \frac{1}{y(t)}\frac{y(t+1)-y(t-1)}{2}$$

By applying this approximation to the two equations of the differential system and plotting the corresponding linear regressions (Figures 1.7B and C), we can obtain an evaluation of the parameters such as $a_1 = 0.82$, $a_2 = 0.0298$, $b_1 = 0.509$, $b_2 = 0.0129$. By matching the initial conditions to the data, the differential system can then be fully determined and solved numerically (Figures 1.7D). Comparison of data and modeling provides a good illustration of the virtues and weaknesses of a mechanistic model. Firstly, based on explicit and interpretable hypotheses, the model was able to recover the cyclical behaviour and dependencies between the two species: the increase in the lynx population always seems to be preceded by the increase in the hare population. However, the amplitude of the oscillations and their periods are not exactly those observed in the data. This may be related to approximations in the evaluation of parameters, random variation in the data or, of course, simplifications or errors in the structure of the model itself.

Besides, if one tries to carry out a statistical modeling of these data, it is very likely that it is possible to approach the curve of populations evolution much closer, especially for the hares. But should it be expressed simply as a function of time or should a joint modeling be proposed? The nature of the causal link between prey and predators will be extremely difficult to establish without strong hypotheses such as those of the mechanistic model. On the other hand, if populations in later years had to be predicted as accurately as possible, it is likely that a sufficiently well-trained statistical model would perform better. Finally, and this is a fundamental difference, the **mechanistic model enables to test cases or hypotheses that go beyond the scope of the data**. Quite simply, by playing with the variables or parameters of the model, we can predict the exponential decrease of predators in the absence of prey and the exponential growth of prey in the absence of predator. More generally, it is also possible to study analytically or numerically the bifurcation points of the system in order to determine the families of behaviours according to the relative values of the parameters [Flake, 1998]. It is not possible to infer these new or hypothetical behaviours directly from the data of the statistical model. This is

theoretically possible on the basis of the mechanistic model, provided that it is sufficiently relevant and that its operating hypotheses cover the cases under investigation. Now that the value of mechanistic models has been illustrated in a fairly theoretical example, all that remains is to explore in the next chapters how they can be built and used in the context of cancer.

## 1.3 Simplicity is the ultimate sophistication

Before concluding this modeling introduction, it is important to highlight one of the most important points already introduced in a concise manner by the poet Paul Valéry at the beginning of this chapter. **Whatever its nature, a model is always a simplified representation of reality and by extension is always wrong to a certain extent**. This is a generally well-accepted fact, but it is crucial to understand the implications for the modeler. This simplification is not a collateral effect but an intrinsic feature of any model:

> *No substantial part of the universe is so simple that it can be grasped and controlled without abstraction. Abstraction consists in replacing the part of the universe under consideration by a model of similar but simpler structure. Models, formal and intellectual on the one hand, or material on the other, are thus a central necessity of scientific procedure.*
> [Rosenblueth and Wiener, 1945]

Therefore, a model exists only because we are not able to deal directly with the phenomenon and simplification is a necessity to make it more tractable [Potochnik, 2017]. This simplification appeared many times in the studies of frictionless planes or theoretically isolated systems, in a totally deliberate strategy. However, this idealization can be viewed in several ways [Weisberg, 2007]. One of them, called Aristotelian or minimal idealization, is to eliminate all the properties of an object that we think are not relevant to the problem in question. This amounts to lying by omission or making assumptions of insignificance by focusing on key causal factors only [Frigg and Hartmann, 2020]. We therefore refer to the *a priori* idea that we have of the phenomenon. The other idealization, called Galilean, is to deliberately distort the theory to make it tractable as explicited by Galileo himself:

> *We are trying to investigate what would happen to moveables very diverse in weight, in a medium quite devoid of resistance, so that the whole difference of speed existing between these moveables would have to be referred to inequality of weight alone. Since we lack such a space, let us (instead) observe what happens in the thinnest and least resistant media, comparing this with what happens in others less thin and more resistant.*

This fairly pragmatic approach should make it possible to evolve iteratively, reducing distortions as and when possible. This could involve the addition of other species or human intervention into the Lotka-Volterra system described above. A three-species Lotka-Volterra model can however become chaotic [Flake, 1998], and therefore extremely difficult to use and interpret, thus underlining the importance of simplifying the model.

We will have the opportunity to come back to the idealizations made in the course of the cancer models but it is already possible to give some orientations. The biologist who seeks to study cancer using cell lines or animal models is clearly part of Galileo's lineage. The mathematical or *in silico* modeler has a more balanced profile. The design of qualitative mechanistic models based on prior knowledge, which is the core of the second part of the thesis, is more akin to minimal idealization, which seeks to highlight the salient features of a system. The Galilean consistins in studying mathematically tractable systems was also important. To take the example of prey-predator interactions, a differential system with more variables quickly becomes impossible to solve by hand. The development of more and more powerful computers has apparently pushed back the limits of the computationally tractable systems and thus of Galilean idealization. However, this is always necessary, for example in high-dimensional statistical approaches (thousands of variables) where the modelers decide to consider the variables independently while neglecting their interactions.

Because of the complexity of the phenomena, simplification is therefore a necessity. The objective then should not necessarily be to make the model more complex, but to **match its level of simplification with its assumptions and objectives**. Faced with the temptation of the author of the model, or his reviewer, to always extend and complicate the model, it could be replied with Lewis Carrol words[1]:

[1]More concisely stated by Rosenblueth and Wiener [1945]: "best material model for a cat is another cat, or preferably the same cat."

*"That's another thing we've learned from your Nation," said Mein Herr, "map-making. But we've carried it much further than you. What do you consider the largest map that would be really useful?"*
*"About six inches to the mile."*
*"Only six inches!" exclaimed Mein Herr. "We very soon got to six yards to the mile. Then we tried a hundred yards to the mile. And then came the grandest idea of all! We actually made a map of the country, on the scale of a mile to the mile!"*
*"Have you used it much?" I enquired.*
*"It has never been spread out, yet," said Mein Herr: "the farmers objected: they said it would cover the whole country, and shut out the sunlight! So we now use the country itself, as its own map, and I assure you it does nearly as well."*
Lewis Carroll, *Sylvie and Bruno* (1893)

CHAPTER 2

# Cancer as deregulation of complex machinery

*"Does not the entireness of the complex hint at the perfection of the simple?"*

*Edgar Allan Poe (Eureka, 1848)*

Armed with all these models, whether statistical or mechanistic, we are going to look at cancer, a particularly complex system that fully justifies their use. Since the first chapter recalled how important prior knowledge of the phenomenon under study is for designing models, whatever their nature, this chapter will briefly summarize some of the most important characteristics of this disease before returning to the models themselves in the next chapter. Without aiming for exhaustiveness, and after an epidemiological and statistical description, we will focus on the most useful information for the modeler, i.e., the underlying biological mechanisms and available data.

Figure 2.1: **Cancer is an old disease.** Rembrandt, *Bathsheba at Her Bath*, c. 1654, oil on canvas, Louvre Museum, Paris

## 2.1 What is cancer?

Cancer can be described as a group of diseases characterized by **uncontrolled cell divisions and growth which can spread to surrounding tissues**. Descriptions of this disease, especially when associated with solid tumors, have been found as far back as ancient Egyptian documents, at least 1600 BC and we know from the first century A.D. with Aulus Celsus that it is better to remove the tumors and this as soon as possible [Hajdu, 2011a]. Progress will accelerate during the Renaissance with the renewed interest in medicine, and anatomy in particular, which will advance the knowledge of tumor pathology and surgery [Hajdu, 2011b]. The progress of anatomical knowledge has also left brilliant testimonies in the field of painting, which

make the renown of the Renaissance today. The precision of these artists' traits has also allowed some retrospective medical analyses, some of them going so far as to identify the signs of a tumor in some of the subjects of these paintings [Bianucci et al., 2018]. Such is the bluish stain on the left breast of the Bathsheba painted by Rembrandt (Figure 2.1) which has been subject to controversial interpretations, sometimes described as an example of "skin discolouration, distortion of symmetry with axillary fullness and peau d'orange" [Braithwaite and Shugg, 1983] and sometimes spared by photonic and computationnal analyses [Heijblom et al., 2014]. The mechanisms of the disease only began to be elucidated with the appearance of the microscope in the 19th century, which revealed its cellular origin [Hajdu, 2012a]. The classification and description of cancers is then gradually refined and the first non-surgical treatments appear with the discovery of ionising radiation by the Curies [Hajdu, 2012b]. The 20th century is then the century of understanding the causes of cancer [Hajdu and Darvishian, 2013, Hajdu and Vadmal, 2013]. Some environmental exposures are characterized as asbestos or tobacco. Finally, the biological mechanisms become clearer with the identification of tumor-causing viruses and especially with the discovery of DNA [Watson and Crick, 1953]. The foundations of our current understanding of cancer date back to this period, which marks the beginning of the molecular biology of cancer. It is this branch of biology that contains the bulk of the knowledge that will be used to build our mechanistic models, and it will be later detailed in Section 2.3.

One of the ways to read this brief history of cancer is to see that theoretical and clinical progresses have not followed the same timeframes. The medical and clinical management of cancers initially progressed slowly but surely, and this in the absence of an understanding of the mechanisms of cancer. Conversely, the theoretical progress of the last century has not always led to parallel medical progress, except on certain specific points. The interaction between the two is therefore not always obvious. The **transformation of fundamental knowledge into clinical impact is therefore of particular importance**. This is what is called *translational medicine*, the aim of which is to go from laboratory bench to bedside [Cohrs et al., 2015]. It is in this perspective that we will analyze the mechanistic models studied in this thesis. Their objective is to integrate biological knowledge, or at least a synthesis of this knowledge, in order to transform it into a relevant clinical information.

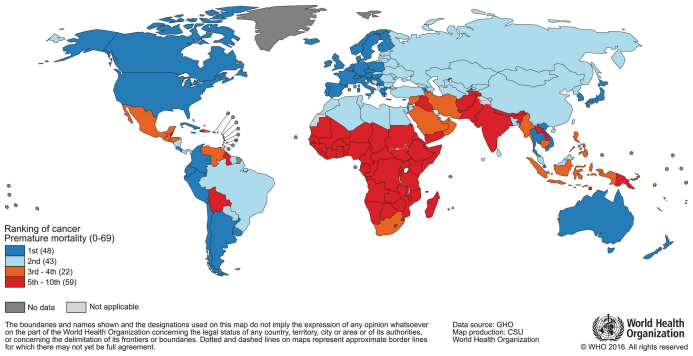



Figure 2.2: **World map and national rankings of cancer as a cause of premature death.** Classification of cancer as a cause of death before the age of 70, based on data for the year 2015. Original Figure, data and methods from Bray et al. [2018].

## 2.2 Cancer from a distance: epidemiology and main figures

Before going down to the molecular level, it is important to detail some figures and trends in the epidemiology of cancer today. Following the description in the previous section, cancer is first and foremost defined as a disease. Considered to be a unique disease, it caused 18.1 million new cancer cases and 9.6 million cancer deaths in 2018 according to the Global Cancer Observatory affiliated to World Health Organization [Bray et al., 2018]. However, these aggregated data conceal disparities of various kinds. The first one is geographical. Indeed, mortality figures make cancer one of the leading causes of premature death in most countries of the world but its importance relative to other causes of death is even greater in the more developed countries (Figure 2.2). All in all, cancer is the first or second cause of premature death in almost 100 countries worldwide [Bray et al., 2018]. These differences call for careful consideration of the impact of population age structures and health-related covariates.

A second disparity lies in the different types of cancer. If we classify tumors solely according to their location, i.e., the organ affected first, we already obtain very wide differences. First of all, the incidence varies considerably (Figure 2.3A)). Cancers do not occur randomly anywhere in the body and certain environments or cell types appear to be more favourable [Tomasetti and Vogelstein, 2015]. Mortality is also highly variable but is not directly inferred from incidence. Not all types of cancer have the same

prognosis (Figure 2.3A and B) and survival rates [Liu et al., 2018]. Although breast cancer is much more common than lung cancer, it causes fewer deaths because its prognosis is, on average, much better. The mechanisms at work in the emergence of cancer are therefore not necessarily the same as those that will govern its evolution or its response to treatment. And still on the response to treatment, Figure 2.3B highlights another disparity: not only are the survival prognosis associated with each cancer very different, but the evolution (and generally the improvement) of these prognoses has been very uneven over the last few decades. This means that theoretical and therapeutic advances have not been applied to all types of cancer with the same success. It is one more indication of the **diversity of cancer mechanisms in different tissues and biological contexts**, which make it impossible to find a panacea, and which, on the contrary, encourage us to carefully consider the particularities of each tumor, both to understand them and to treat them. Under a generic name and in spite of common characteristics, the cancers thus appear as extremely heterogeneous. And to understand the sources of this heterogeneity, it is necessary to consider the disease on a smaller scale.

## 2.3 Basic molecular biology and cancer

If it is not possible and desirable to summarize here the state of knowledge about the biology of cancer, we are going to give a very partial vision focused on the main elements used in this thesis, thus aiming to make it a self-sufficient document. The details necessary for a finer and more general understanding can be found in dedicated textbooks such as Alberts et al. [2007] and Weinberg [2013].

### 2.3.1 Central dogma and core principles

Some of the principles that govern biology can be described at the level of one of its simplest element, the cell. Let us consider for the moment a perfectly healthy cell. It must ensure a certain number of functions necessary for its survival and, sometimes, for its division/reproduction. These functions are encoded to a large extent in its genetic information in the form of DNA, which is stable and shared by the different cells since it is defined at the level of the individual. Most biological functions, however, are not performed by DNA itself which remains in the nucleus of the cell. The DNA is thus transcribed into RNA, another nucleic acid which, in addition to performing some biological functions, becomes the support of

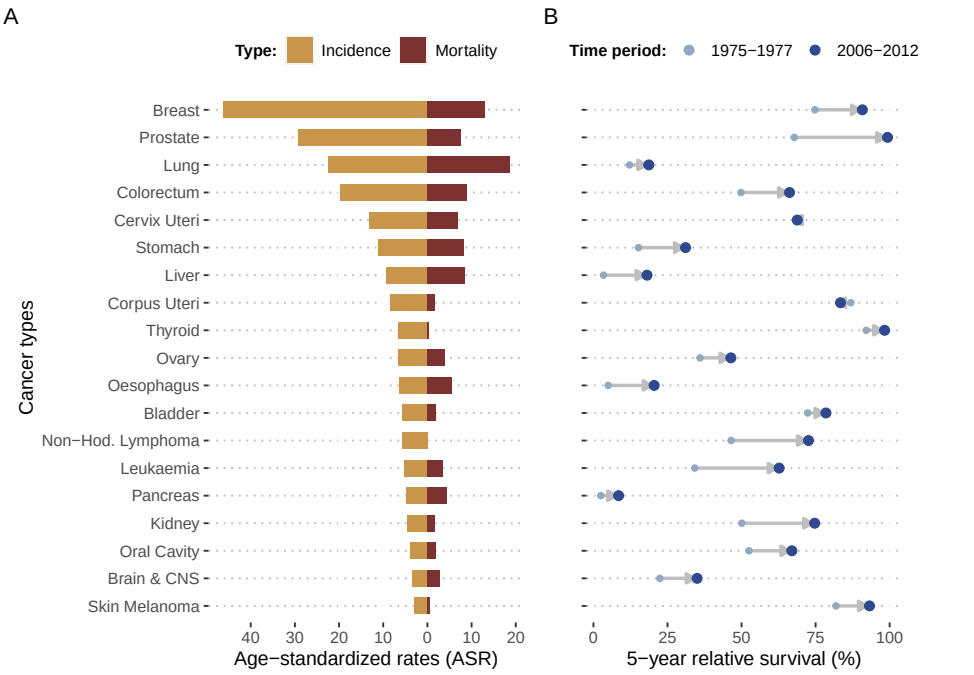



Figure 2.3: **Incidence, mortality and survival per cancer types**. (A) World incidence and mortality for the 19 most frequent cancer types in 2018, expressed with age-standardized rates (adjusted age structure based on world population); data retrieved from Global Cancer Observatory. (B) Evolution of 5-years relative survival for the same cancer types based on US data from SEER registries in 1975-1977 and 2006-2012; data retrieved from Jemal et al. [2017].

the genetic information in the cell. The RNA is then itself translated into new molecules composed of long chains of amino acid residues and called proteins. They are the ones that execute most of the numerous cellular functions: DNA replication, physical structuring of the cell, molecule transport within the cell etc. A rather simplistic but fruitful way to understand this functioning is to consider it as a **progressive transfer of biological information from DNA to proteins**, which has sometimes been summarized as the central dogma of the molecular biology (2.4), first stated Francis Crick [Crick, 1970].

However, many changes would be necessary to clarify this scheme and the uni-directional nature was questioned early on. Above all, a large number of regulations interact with and disrupt this master plan. The genes are not always all transcribed, or at least not at constant intensities, interrupt-

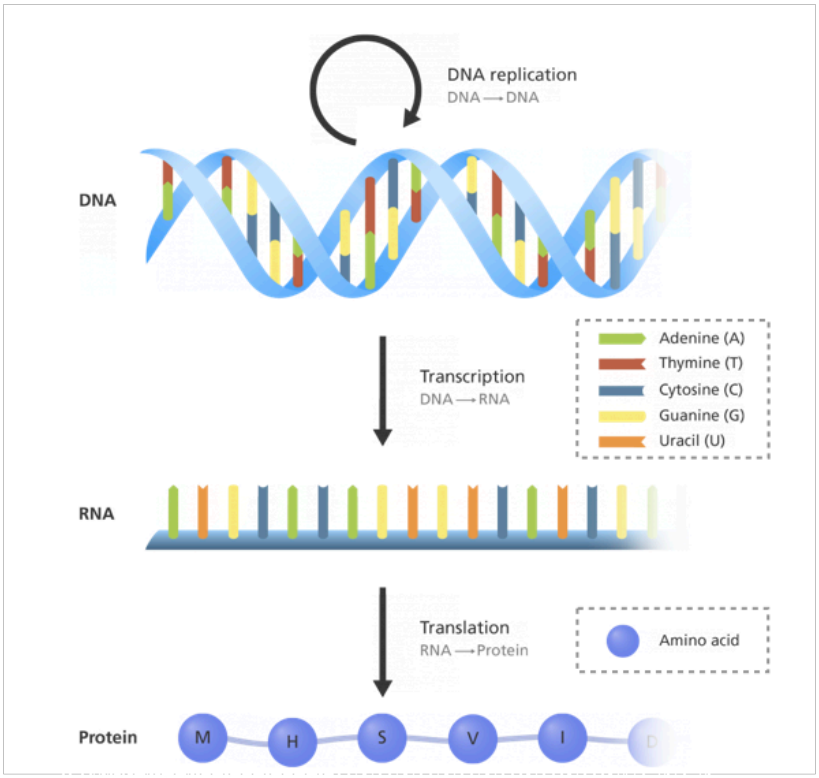



Figure 2.4: **Central dogma of molecular biology.** Schematic representation of the information flow within the cell, from DNA to proteins through RNA, more precisely described in this video (Image credit *Genome Research Limited*).

ing or varying the chain upstream. This modulation in the transcription of genes can be induced by proteins, called transcription factors. After a gene transcription, its expression can still be regulated at various stages. RNAs can also be degraded more or less rapidly. RNAs can be reshaped in their structure during their maturation by a process called splicing, which varies the genetic information they carry. Finally, proteins are subject to all kinds of modifications referred to as post-translational, which can change the chemical nature of certain groups or modify the three-dimensional structure of the whole protein. For instance, some proteins perform their function only if a specific amino acid residue is phosphorylated. In addition, these modifications can be transmitted between proteins, further complicating the flow of information. **All these possibilities of regulation play an absolutely essential role in the life of the cell by allowing it to adapt to different contexts, situations and development stages**.

From the same genetic material, a cell of the eye and a cell of the heart can thus perform different functions. Similarly, the same cell subjected to different stimuli at different times can provide different responses because these molecular stimuli trigger a regulation of its programme. But all these regulatory mechanisms can be corrupted.

### 2.3.2 A rogue machinery

With the above knowledge we can now return to the definition of cancer as an uncontrolled division of cells that can lead to the growth of a tumor that eventually spreads to the surrounding tissues. Therefore, this corresponds to normal processes, like cell division and reproduction, that are no longer regulated as they should be and are out of control. Experiments on different model organisms have gradually identified genetic mutations as a major source of these deregulations [Nowell, 1976, Reddy et al. [1982]] until cancer was clearly considered as a **genetic disease** making Renato Dulbecco, Nobel Laureate in Medicine for his work on oncoviruses, say:

> *If we wish to learn more about cancer, we must now concentrate on the cellular genome.*
> [Dulbecco, 1986].

However, cancer is not a Mendelian disease for which it would be sufficient to identify the one and only gene responsible for deregulation. Indeed, the cell has many protective mechanisms. For example, if a genetic mutation appears in the DNA, it has a very high chance of being repaired by dedicated mechanisms. And if it is not repaired, other mechanisms will take over to trigger the programmed death of the cell, called apoptosis, before it can proliferate wildly. So a cancer cell is probably a cell that has learned to resist this cell death. Similarly, in order to generate excessive growth, a cell will need to be able to replicate itself many times. However, there are pieces of sequences on chromosomes called telomeres that help to limit the number of times each cell can replicate. A cancer cell will therefore have to manage to bypass this protection. Thus we can schematically define the properties that must be acquired by the cancereous cells in order to truly deviate the machinery. In an influential article, these properties were summarized in six hallmarks (Figure 2.5) which are: resisting cell death, enabling replicative immortality, sustaning proliferative signaling, evading growth suppressors, activating invasion and inducing angiogenesis [Hanahan and Weinberg, 2000]. Two new ones were subsequently added in the light of

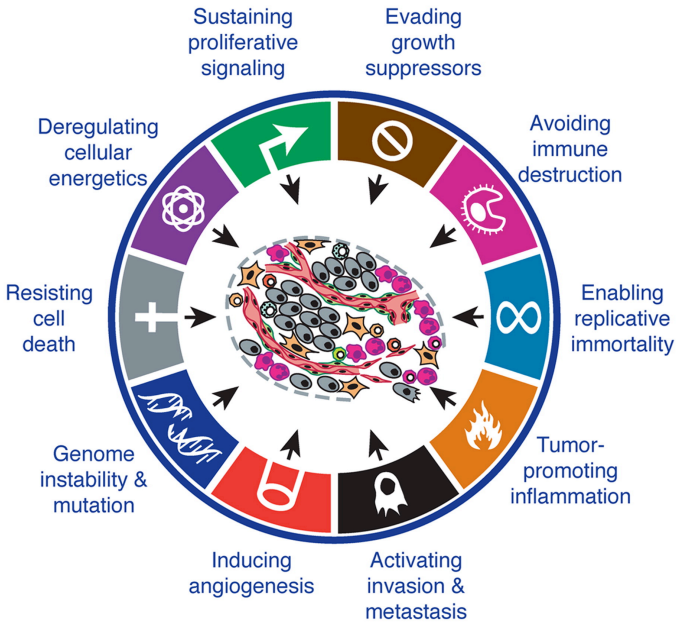



Figure 2.5: **Hallmarks of cancer.** The different biological capabilities acquired by cancer cells. Adapted from Hanahan and Weinberg [2011].

advances in knowledge [Hanahan and Weinberg, 2011]: deregulating cancer energetics and avoiding immne destruction. The acquisition of these capacities generally requires many genetic mutations and is therefore favoured by an underlying genome instability or tumor-promoting inflammation.

Each of these characteristics, or hallmarks, constitutes a research program in its own right. And for each one there are genetic alterations. These are tissue-specific or not, specific to a hallmark or common to several of them [Hanahan and Weinberg, 2000]. In any case, **cancer can only result from the combination of different alterations that invalidate several protective mechanisms** at the same time. This is often part of a multi-step process of hallmark acquisition that has been experimentally documented in some specific cases [Hahn et al., 1999] or more recently inferred from genome-wide data for human patients [Tomasetti et al., 2015]. In summary, it appears that in order to study the functioning of cancer cells it is necessary to look at several mechanisms and to be able to consider them not separately but together, in as many different patients as possible. This ambitious programme has been made possible by a technological revolution.

## 2.4 The new era of genomics

### 2.4.1 From sequencing to multi-omics data

In 2001, the first sequencing of the human genome symbolized the beginning of a new era, that of what will become **high-throughput genomics** [Lander et al., 2001, Venter et al., 2001]. From the end of the 20th century, biological data started to accumulate at an ever-increasing rate [Reuter et al., 2015], feeding and accelerating cancer research in particular [Stratton et al., 2009, Meyerson et al., 2010]. The ability to sequence the human genome as a whole, for an ever-increasing number of individuals, has enabled **less biased and more systematic studies of the causes of cancer** [Lander, 2011]. The number of genes associated with cancer increased drastically and some very important genes such as BRAF or PIK3CA have been identified [Davies et al., 2002, Samuels et al., 2004]. Progress also extended to the gene expression data. Gene-expression arrays have made an important contribution by providing access to transcriptomic data (RNA), i.e., what has been transcribed from DNA and is therefore one step further in terms of biological information. This information has made it possible to further explore the differences between normal and tumor cells [Perou et al., 1999], or even to refine the classification of cancers, which until now has been done mainly according to the tumor site. Breast cancers are thus divided into subtypes with different combinations of molecular markers that facilitate the understanding of clinical behavior [Perou et al., 2000]. One step further, we also note the appearance of prognostic gene signatures such as gene expression patterns correlated with the survival of patients [Van't Veer et al., 2002]. This revolution was then extended to other types of data such as proteins (proteomics), reversible modifications of DNA or DNA-associated proteins (epigenomics), metabolites (metabolomics) and others, each representing a perspective that can complement the others to better understand biological mechanisms, particularly in the case of diseases [Hasin et al., 2017]. We have thus entered the era of multi-omics data [Vucic et al., 2012].

### 2.4.2 State-of-the art of cancer data

With respect to cancer in particular, this wealth of data is particularly represented by a family of **studies conducted by The Cancer Genome Atlas (TCGA) consortium**, started in 2008 [Network et al., 2008]. Cohorts of several hundred patients are thus sequenced over the years for different types of cancer [TCGA et al., 2012], resulting today in a total of 11,000 tumors from 33 of the most prevalent forms of cancer [Ding et al.,

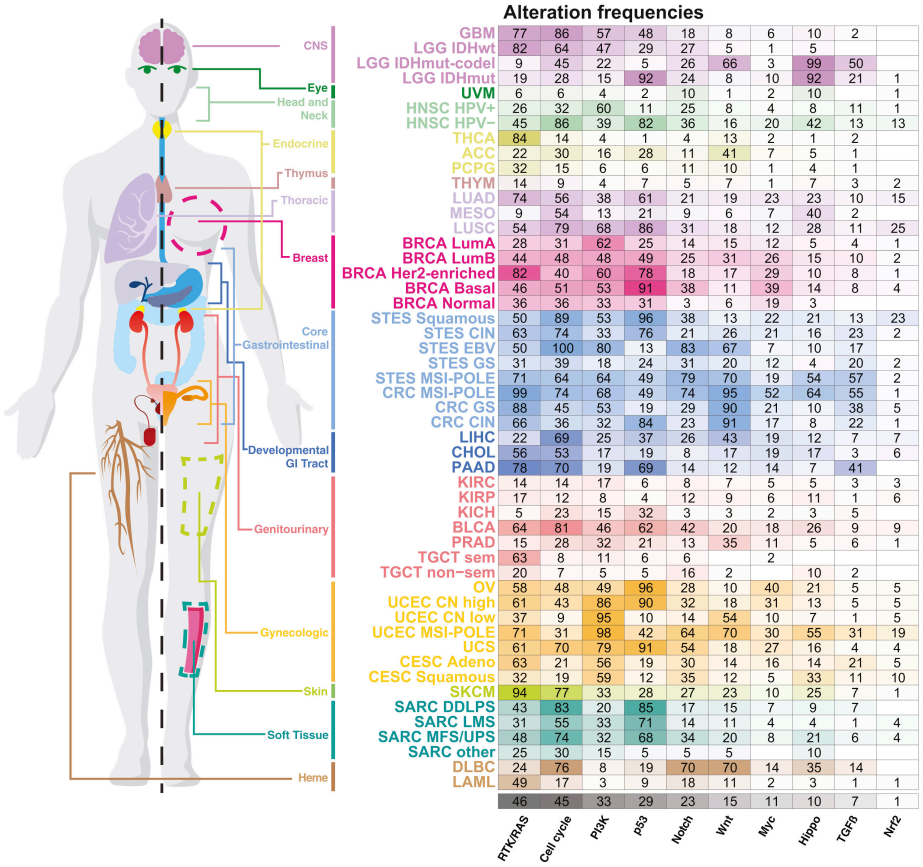



Figure 2.6: **Genetic alterations frequencies for cancer types from TCGA data.** Frequencies of alteration per pahway and tumor types as summaried in Pan-cancer analyses from TCGA data. Reprinted from Sanchez-Vega et al. [2018].

2018]. Figure 2.6 provides a partial but striking overview of the depth of data available under this program. We can see the frequencies of alterations of certain groups of genes for a list of cancer types, making it possible to visualize the disparities already anticipated in section 2.2 based on patient survival. There are indeed important differences between the organs but also between the different subtypes associated with the same organ. And this representation only corresponds to one layer of data, that of genetic alterations. It could be used for transcriptomic, epigenomic or proteomic data, thus giving rise to an incredibly complex photography.

However, the diversity of data available for cancer research extends far

beyond this, both in terms of technology and type of data. This may be data from model organisms such as mice or even tumors of human origin made more suitable for experimentation. In the latter category, it is crucial to mention the **huge amount of data available on cell lines**, extracted from human tumors and transformed to be studied in culture. It is then possible to go beyond descriptive data and vary the experimental conditions in order to study the responses of these cells to perturbations and to enrich our knowledge. This provides an opportunity to know the response to more than 100 drugs of about 700 cell lines [Yang et al., 2012]. The richness of these data, coupled with the omic profiling of each cell line, enables to study the determinants of response to treatment with unprecedented scope [Iorio et al., 2016]. More recently, but following a similar logic, other types of inhibition screenings have been proposed based on a more specific technique called CRISPR-Cas9 [Behan et al., 2019]. The simplicity of the cell lines in relation to the original tumors makes all these studies possible but sometimes hinders the clinical application of the knowledge acquired. For this reason, other types of biological models have been developed, including patient-derived xenografts (PDX) which is an implant of human tumors in mice to ensure the existence of a certain tumor microenvironment [Hidalgo et al., 2014], while maintaining drug screening possibilities [Gao et al., 2015]. These two types of data, cell lines and PDX, have been used in this thesis, in addition to TCGA patient data, thus justifying the limitation of this presentation, which could otherwise be extended to other types of biological models. Similarly, other technologies are becoming increasingly important in the generation of cancer data, such as single-cell sequencing [Navin, 2015], but will not be used in this work.

## 2.5 Data and beyond: from genetic to network disease

All that remains to be done now is to make sense of all these data, to organize them, because **cancer understanding does not flow directly from the abundance of data**, and the ability to produce it may have been outpaced by the ability to analyze it [Stadler et al., 2014]. A striking example is that of the prognostic signatures mentioned above. The many signatures or lists of genes proposed, even for the same cancer type, share relatively few genes, are difficult to interpret and their efficiency is sometimes poorly reproducible [Domany, 2014]. Even more surprisingly, most signatures composed of randomly selected genes were also found to

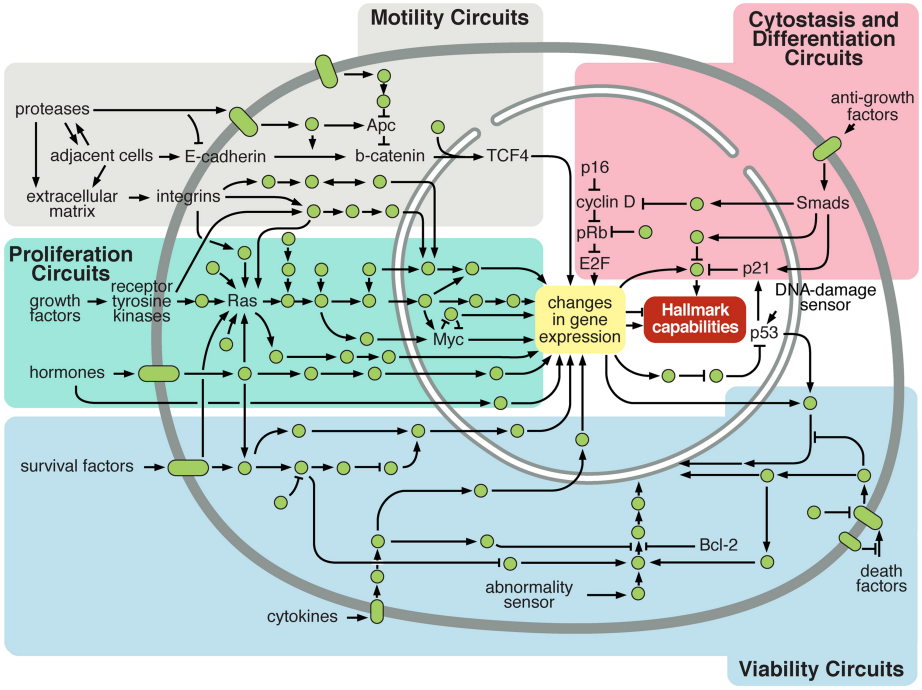



Figure 2.7: **Simplistic representation of cellular circuitry.** Normal cellular circuit sand sub-circuits (identified by colours) can be reprogrammed to regulate hallmark capabilities within cancer cells. Reprinted from Hanahan and Weinberg [2011].

be associated with patient survival [Venet et al., 2011]. One of the main avenues for improving the interpretability of the data is the **integration of the prior knowledge** we have of the phenomena, especially in the case of cancer [Domany, 2014].

This *a priori* knowledge is in fact already present in Figure 2.6 since genetic alterations have been grouped in several categories called pathways. A pathway is a group of biological entities and associated chemical reactions, working together to control a specific cell function like apoptosis or cell division. The interest of these groupings may be understood based on the description of hallmarks. Indeed, if the "aim" of a cancer cell is to inactivate each of the protective functions, then it is more relevant to think not by gene but by function. Inactivating only one of the genes associated with the function may be sufficient and it is no longer necessary to inactivate the others. Numerous alterations in a large number of genes in various patients result often in the same key impaired pathways, like alterations of cell cycle

or angiogenesis for instance [Jones et al., 2008]. It is therefore possible to improve the stability and interpretability of analyses by moving **from the gene scale to the pathway scale** [Drier et al., 2013]. More generally, the integration of biological knowledge often leads to improved performance in various cancer-related prediction tasks, either through the selection of variables or by taking into account the structure of the variables [Bilal et al., 2013, Ferranti et al., 2017]. Increasingly, the biological variables are not interpreted separately but in relation to each other [Barabasi and Oltvai, 2004]. This is reflected in the emergence of more and more resources to summarize and represent signaling pathways and associated networks such as SIGNOR [Perfetto et al., 2016] or the Atlas of Cancer Signaling Network [Kuperstein et al., 2015]. Like other diseases, cancer then goes **from a genetic disease to a network disease** [Del Sol et al., 2010] and one can study how all kinds of genetic alterations affect the wiring of these networks [Pawson and Warner, 2007], and modify the cellular functions leading to the previously described cancer hallmarks as depicted schematically in Figure 2.7. In short, the richness of the data did not make it less necessary to use prior knowledge in order to make the analyses more interpretable and more robust.

The final step, to obtain one of the most complete and integrated visions of cancer biology, is then to integrate omics knowledge with knowledge about the structure of pathways to try to understand in detail how their combinations can lead to so many cancers that are both similar and different. An example of such a representation is given by mapping the TCGA data about genetic alterations, presented in Figure 2.6, on a representation of the different pathways showing not only their internal organization but also their cross-talk [Sanchez-Vega et al., 2018]. This representation is proposed in Figure 2.8 and is one of the most recent and comprehensive view of the kind of tools and data available to the modeler who wants to dissect more deeply the mechanisms involved in cancer.

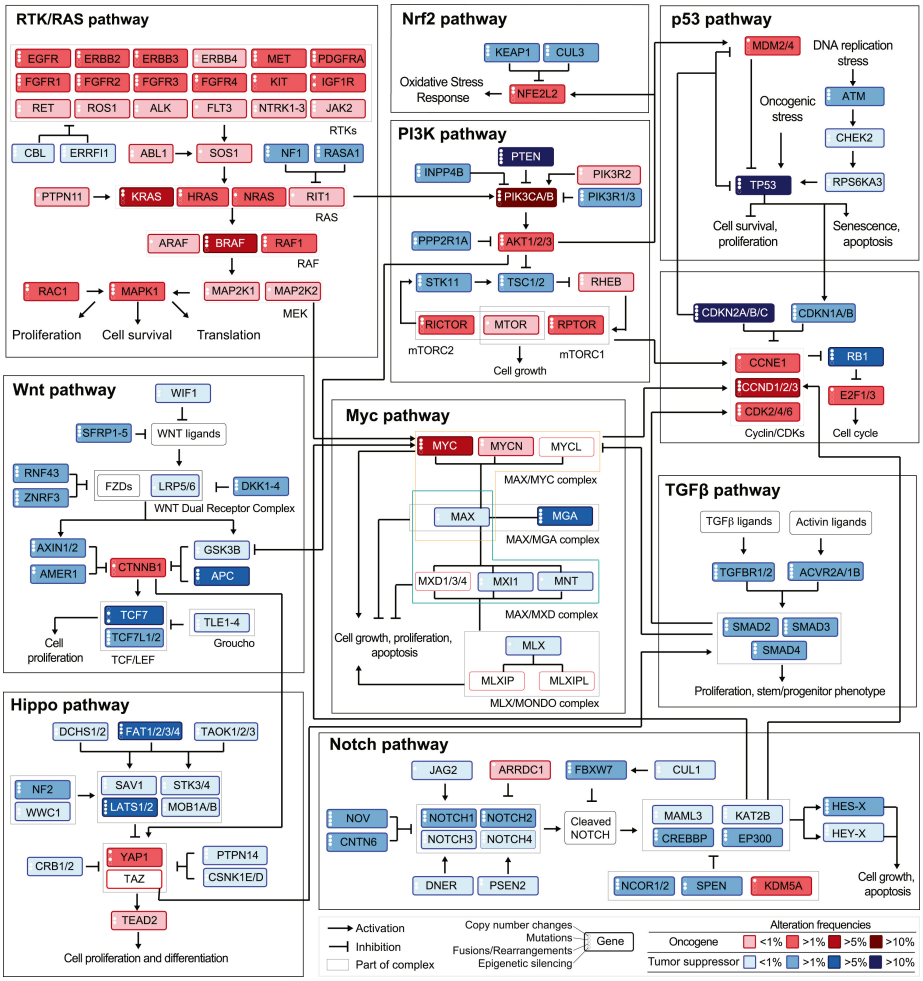



Figure 2.8: **Genetic alterations frequencies from TCGA data mapped on a schematic signaling network.** Frequencies of alteration per pathway and tumor types as summarized in Pan-cancer analyses from TCGA data. Reprinted from Sanchez-Vega et al. [2018].

Chapter 3

# Mechanistic modeling of cancer: from complex disease to systems biology

*"How remarkable is life? The answer is: very. Those of us who deal in networks of chemical reactions know of nothing like it... How could a chemical sludge become a rose, even with billions of years to try."*

*George Whitesides (The improbability of life, 2012)*

The previous chapter identified the need to organize cancer knowledge and data. The integration of biological knowledge, particularly in the form of networks, is a first step in this direction. The deepening of knowledge, however, requires the ability to manipulate objects even more, to experiment, to dissect their behaviour in an infinite number of situations, such as the astronomer with his orrery or physicians with their old anatomical models (Figure 3.1). Is it then possible to create mechanistic models of cancer in the same way?

Figure 3.1: **Dissecting a biological phenomenon using a non-computational model.** Rembrandt, *The Anatomy Lesson of Dr Nicolaes Tulp*, 1634, oil on canvas, Mauritshuis museum, The Hague

## 3.1 Introducing the diversity of mechanistic models of cancer

Modeling cancer is not a new idea. And the diversity of biological phenomena involved in cancer has given rise to an equally important diversity of models and formalisms, which we seek here to give a brief overview in order to better identify the specific models that we will focus on later. One way to order this diversity is to consider the scales of these models (Figure 3.2). Indeed, **cancer can be read at different levels, from the molecular level of DNA and proteins, to the cellular level, to the level of tissues and organisms** [Anderson and Quaranta, 2008]. Models have been proposed at all these scales, using different formalisms [Bellomo et al., 2008] and answering different questions.

Consistent with the evolution of knowledge and data, the early models were at the **macroscopic level**. While methods and terminologies may have changed, there are nevertheless traces of these models as early as the

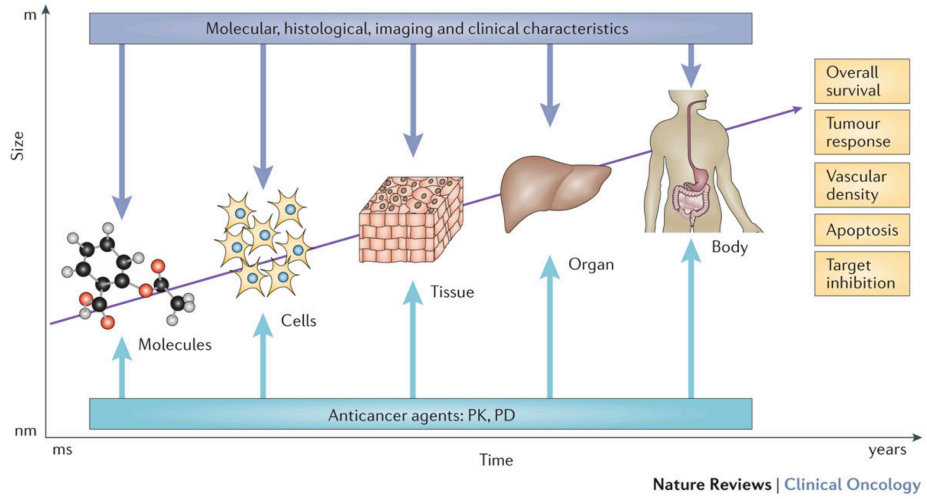



Figure 3.2: **The different scales of cancer modeling.** Cancer can be approached at different scales, from molecules to organs, using different data (dark blue), but often with the direct or indirect objective of contributing to the study of clinically interpretable phenomena (yellow boxes), in particular by studying the influence of anticancer agents (pale blue). Reprinted from Barbolosi et al. [2016].

1950s. We then speak rather of mathematical modeling with a meaning that is intermediate between what we have defined as mechanistic models and statistical models [Byrne, 2010]. First, the initiation of tumorigenesis was theorized with biologically-supported mathematical expressions in order to make sense of cancer incidence statistics [Armitage and Doll, 1954, Knudson, 1971]. These models, however, remained relatively descriptive in that they did not shed any particular light on the biological mechanisms involved and focused on gross characteristics of tumors. The integration of more advanced knowledge as well as the progressive refinement of mathematical formalisms has nevertheless allowed these models to proliferate while gaining in interpretability, with for instance mechanistic models of metastatic relapse [Nicolò et al., 2020]. Always on a macroscopic scale, the study of tumor growth has also been the playground of many mathematicians [Araujo and McElwain, 2004, Byrne, 2010], even predicting invasion or response to surgical treatments using spatial modeling [Swanson et al., 2003]. This line of research is still quite active today and provides a mathematical basis for comparison with tumor experimental growth [Benzekry et al., 2014].

Taking it down a step further, it is also possible to model cancer at the **cellular level**, for example by looking at the clonal evolution of cancer [Altrock et al., 2015]. The aim is then to understand the impact of the processes of mutation, selection, expansion and cohabitation of different populations of cells, at specifc rates. The accumulation of a mutation in a population of cells can thus be studied [Bozic et al., 2010]. Modeling at the cellular level is well suited to the study of interactions between cells, between cancer cells and their environment or with the immune system. Similar to other kinds of studies of population dynamics, formalisms based on differential equations are quite common [Bellomo et al., 2008]; but there are many other methods such partial differential equations or agent-based modeling [Letort et al., 2019].

Finally, at an even smaller scale, it is possible to model the **molecular networks** at work in cells [Le Novere, 2015]. The aim is then to simulate mathematically how the different genes and molecules regulate each other, transmit information and, in the case of cancer, end up being deregulated [Calzone et al., 2010]. These models will be the subject of the thesis and will therefore be defined more precisely and used to detail the concepts and tools of systems biology in the following sections. It can already be noted that while these models can integrate the most fundamental biological mechanisms of living organisms, one of the most burning questions is whether it is possible to link them to the larger scales that are clinically more interesting (tissues, organs etc.). Can these models tell us something about the molecular nature of cancer? About patient survival? Their response to treatment? These questions apply to all of the above models, whatever their scales (Figure 3.2), but are more difficult to answer for models defined at molecular scale that are further from the clinical data of interest. The aim of this thesis is to provide potential answers to these questions. One of the ways of approaching these issues has been to propose multi-scale models, which are nevertheless very complex [Anderson and Quaranta, 2008, Powathil et al., 2015]. We will focus here on the use of models defined almost exclusively at the molecular scale, which is assumed to be prominent, to study what can be inferred on the larger scales.

## 3.2 Cell circuitry and the need for cancer systems biology

Most biological systems, and certainly cells, fall into the category of **complex systems**. These are systems made up of many interacting elements.

While these systems can be found in many different scientific fields, the cell as a complex system is characterized by the diversity and multifunctionality of its constituent elements (genes, proteins, small molecules, enzymes), which nevertheless contribute to organized and *a priori* non-chaotic behaviour [Kitano, 2002]. Thus, the role of a protein such as the p53 tumor suppressor can only be understood by taking into account the interplay between its relationships with transcription factors and biochemical modifications of the molecule itself [Kitano, 2002]. In a cell, as in any complex system, the multiplication of components and interactions can make the response or behaviour of the system unexpected or unpredictable. Non-linear responses, such as abrupt changes in the state of a system, called critical transitions, can be observed in response to a moderate change in the signal [Trefois et al., 2015]. Generally speaking, it is possible to observe **emergent behaviours**, i.e., behaviours of the system as a whole that were not trivially deducible from the individual behaviours of its components. This has been documented, through experiments and simulations, in the study of cell signalling pathways and the resulting biological decisions [Bhalla and Iyengar, 1999, Helikar et al., 2008]. These considerations have thus given rise to **system-level or holistic approaches that aim to integrate data and knowledge into more comprehensive representations, often called systems biology**.

What is true for the cell in general is just as true for cancer in particular. Understanding the intertwining of signaling pathways is necessary to study their contributions to different cancer hallmarks, as shown in Figure 2.7. The concepts described above can thus be transposed to **cancer systems biology** [Hornberg et al., 2006, Kreeger and Lauffenburger, 2010, Barillot et al., 2012]. Indeed, it is often a question of understanding or predicting the impact of perturbations on cellular networks. Understanding how a single genetic mutation disrupts and reprograms networks, or even predicting the responses triggered by a drug on a presumably promising molecular target, makes little sense without integrated approaches. In addition, cancers are characterized by the accumulation of numerous mutations and alterations over time that must be considered concomitantly. These points of view of biologists and modelers reinforce the observation already made in the previous chapter of cancer as a network disease, as a system disease (Figure 2.8).

Finally, to conclude this general presentation, it is important to understand that while small molecular network modeling is not recent, the rise and multiplication of wide range systems biology approaches is very much

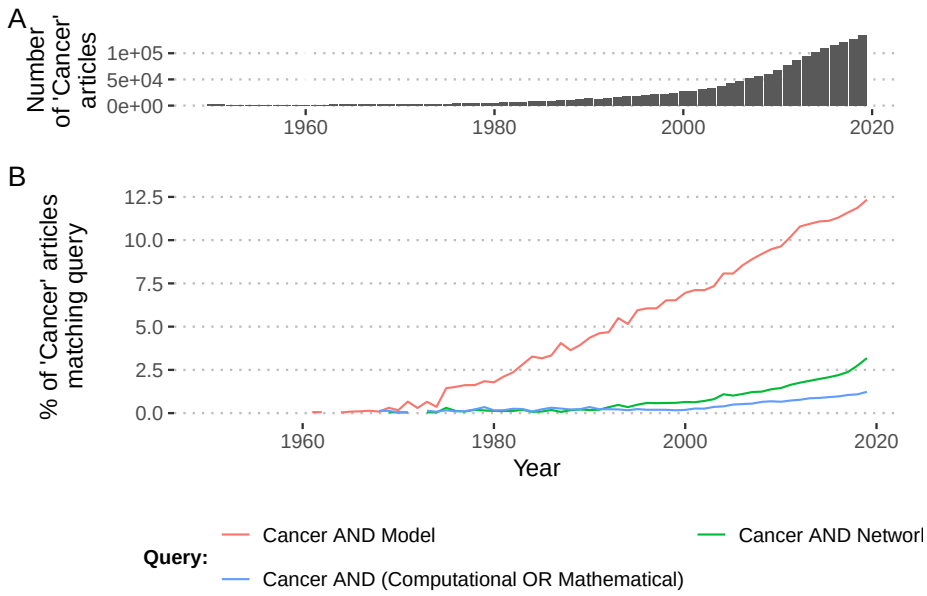



Figure 3.3: **PubMed trends in cancer studies.** (A) PubMed articles with the word *Cancer* in either title or abstract from 1950 to 2019. (B) Proportion of the *Cancer* articles with additional keywords expressed as PubMed logical queries.

related to the production of biological data [De Jong, 2002]. The last few decades have seen the emergence of high-throughput data that has made it possible to identify and link hundreds of genes or proteins involved in cancer. Exploring the interaction and back and forth between these models and the data they use or predict is therefore of utmost importance. In addition, the now **massive amount of data has also imposed mathematical or computational approaches as a central element in the management of this profusion** and more and more modeling approaches are focused on data integration or inference [Fröhlich et al., 2018, Bouhaddou et al., 2018]. More generally, Figure 3.3 shows that while the number of scientific articles devoted to cancer has increased drastically since the 1950s (panel A), the proportion of these same articles mentioning *models*, *networks* or *computational* approaches has also increased (panel B), illustrating a change in paradigms.

## 3.3 Mechanistic models of molecular signaling

Once the context has been defined, both biologically and methodologically, it is possible to begin the exploration of the models that will constitute the core of this thesis: the **mechanistic models of molecular networks** and signaling pathways. Before describing and illustrating some of the existing mathematical formalisms, it is possible to describe the common fundamental elements of this family of approaches.

### 3.3.1 Networks and data

The first step is to identify the relevant biological entities from a question or system of interest (e.g. tumor suppressor genes, signaling cascades of proteins) and then to model their interactions, the regulatory relationships that link them. At this stage the model can generally be represented by a network but this word can cover different realities [Le Novere, 2015]. The simplest network just represents undirected interactions between entities, which therefore only establishes relationships and not causal mechanisms. But modeling requires more precise definitions, in particular concerning the direction of the interaction (is it A that acts on B or the opposite) and its nature (type of chemical reaction, activation/inhibition etc.). This is usually summarized as **activity flows (or influence diagrams) with activation and inhibition arrows** as in Figure 2.7 or Figure 3.5A. These arrows emphasize the transformation of static networks into dynamic objects that can be manipulated and interpreted mechanistically. This work can be taken further by writing bipartite graphs, known as process descriptions, which explicitly show the different states of each variable (first type of nodes), depending on their phosphorylation state for instance, and the reactions that link them (second type of nodes) as in Figure 3.5B. A more precise description of these different representations and their meanings can be found in Le Novere [2015]. **Once the network structure of the model has been defined or inferred by the modeler, it is possible to write the corresponding mathematical formalism** and potentially to refine certain parameters. Finally, the model is often confronted with new data to check its consistency with the biological behaviour studied or possibly make new predictions.

However, all these steps are not linear and sequential, but rather iterative and cyclical. This **modeling cycle**, with back and forth to the data, is

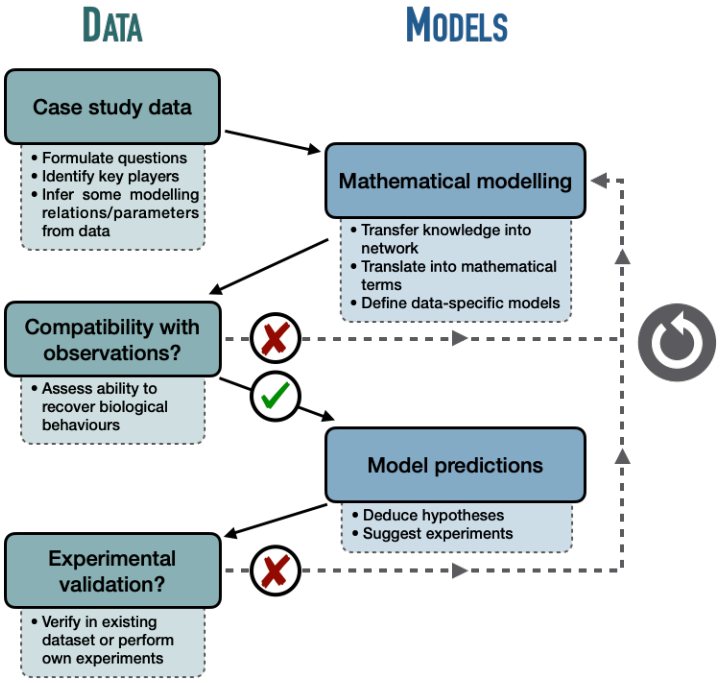



Figure 3.4: **Modeling a biological network: an iterative and cyclical process.** Reprinted from [Béal et al., 2020]. A different and simpler version of this cycle is described in [Le Novere, 2015].

not specific to molecular network models, but it is possible to specify it in this case (Figure 3.4). The names of the key players involved in the question of interest are thus first extracted from adapted data or from the literature. A first mathematical translation of the relationships between the entities is then proposed before verifying the compatibility of this model with the observations, whether qualitative or quantitative. If the compatibility is not good, we come back to the definition or the parameterization of the model. If compatibility is correct, the model can be used to make new predictions or study phenomena that go beyond the initial data set. Ideally, these predictions will be tested afterwards. This cyclic approach with two successive checks is analogous to the use of validation and test data in the evaluation of most learning algorithms. This analogy can sometimes be masked by the qualitative nature of the predictions or by the lack of explicit fitting of the parameters.

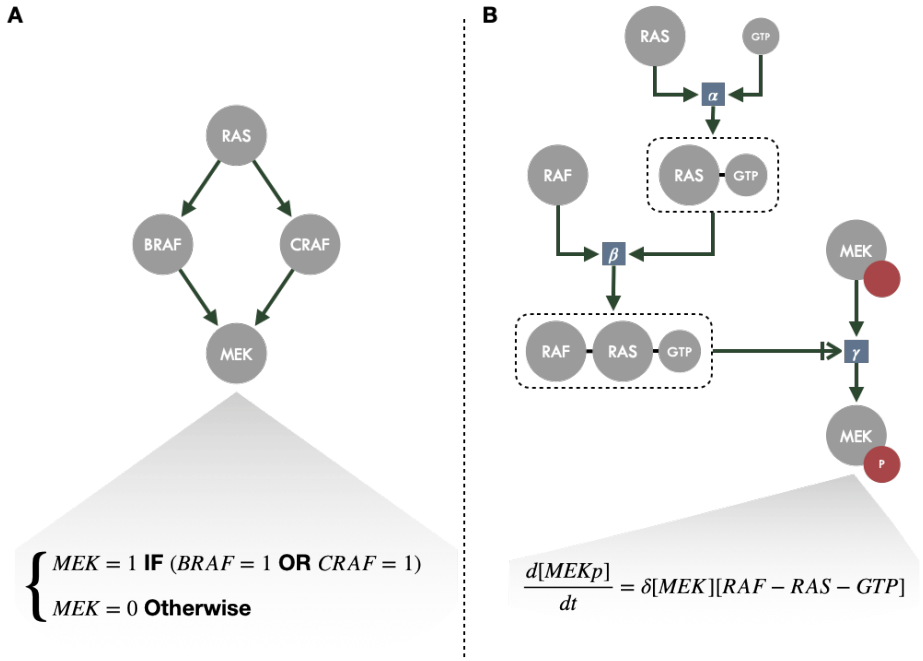



Figure 3.5: **Schematic example of logical and ODE modeling around MAPK signaling.** (A) Activity flow diagram of a small part of MAPK signaling, each node representing a gene or protein, with an example of logical rule for MEK node for the corresponding logical model. (B) Process description of the same diagram with BRAF and CRAF merged in RAF for the sake of simplicity; each square representing a reaction and the corresponding rate; an example differential equation is provided for the phosphorylated (active) form of MEK.

### 3.3.2 Different formalisms for different applications

Beyond these similarities in the construction and representation of models, the precise mathematical formalism that underlies them varies according to the type of question and the data [De Jong, 2002]. For the sake of simplicity, and without exhaustiveness, we propose to divide into quantitative and qualitative formalisms which will be essentially illustrated respectively by **ordinary differential equation (ODE)** models and **logical (or Boolean) models** for which a graphical and schematic comparison is proposed in Figure 3.5.

One of the most frequent approaches is the use of **chemical kinetics** equations to construct ODE systems which are a fairly natural transla-

tion of the process descritption networks described in the previous section [Polynikis et al., 2009]. For instance, each biological interaction can be treated as a reaction governed by the law of mass action and, under certain hypotheses, as a differential equation (Figure 3.5B); the set of reactions in the system then generates a set of differential equations with coupled variables, in an analogous way to the Lotka Volterra system presented in section 1.2.2. Thus the variables generally represent quantities of molecular species, for example concentrations of RNA or proteins, and the stoichiometric coefficients and reaction rates are used to define the system parameters. Approximations are sometimes made to simplify the equations, for example by assuming that they can be written as Michaelis-Menten's enzymatic reactions, which have a simple and well known behaviour. However, the theoretical accuracy of quantitative models has a cost since **each differential equation requires parameters**, such as reaction constants or initial conditions, to which the system is very sensitive [Le Novere, 2015]. The biochemical interpretation of the parameters sometimes allows to find their value in the literature, or in dedicated databeses [Wittig et al., 2012], if the reactions are well characterized, even if possible variations in a given biological or physical context are often unknown. Since knowledge of the values of these parameters is often limited or even non-existent, it may require a very large volume of data (including time series) to fit the many missing parameters which can be difficult if the number of parameters is large [Villaverde and Banga, 2014]. However, recent work has demonstrated the feasibility and scalability of this type of inference with sufficiently rich data [Fröhlich et al., 2018].

At the same time, more qualitative approaches to modeling biological networks have been proposed with discrete variables linked together by rules expressed as logical statements [Abou-Jaoudé et al., 2016]. These models are both more abstract since variables do not have a direct biological interpretation (e.g. concentration of a species) but are more versatile since they can unify different biological realities under the same formalism (e.g. activation of a gene or phosphorlation of a protein). The discrete nature of the variables can then be seen as an asymptotic case of the sigmoidal (e.g. Hill function) relationships often found in biology [Le Novere, 2015]. The step function thus obtained can keep a natural interpretation in the context of biological phenomena: genes activated or not, protein present or absent etc. Similarly, interactions between species are not quantified but are based on qualitative statements (e.g. A will be active if B and C are active), drastically reducing the number of parameters (Figure 3.5A). If the theoretical interest of this formalism to study biological mechanisms

Table 3.1: **Features of quantitative and qualitative modeling applied to biological molecular networks** (adapted from Le Novere [2015])

| | Quantitative modeling | Qualitative modeling |
|---|---|---|
| **Example formalism** | Ordinary differential equation (ODE) models | Logical models |
| **Type of variables** | Direct translation of biological quantities, usually continuous | Abstract representation of activity levels, usually discrete |
| **Objective** | Quantitatively accurate and temporal simulation of an experimental phenomenon | Coarse-grained simulation of qualitative phenotypes |
| **Advantages** | Direct confrontation with experimental data; precise; linear representation of time | Faster design; easy translation of literature-based assertions; simulation of perturbations |
| **Drawbacks** | Difficulty determining or fitting parameters | More difficult to link to data; lower precision |

was proposed quite early [Kauffman, 1969, Thomas, 1973], many concrete applications have also been developed over the years, particularly in cancer research [Saez-Rodriguez et al., 2011a, Remy et al., 2015]. This **logical formalism will constitute the core of the work presented in Part II**, where it will therefore be discussed in greater detail.

These two formalisms, which are among the most frequent for modeling biological networks, share many similarities, in particular the propensity to be built according to bottom-up strategies based on knowledge of the elementary parts of the model, i.e., biological entities and reactions. However, they differ in their implementation and objectives, **one aiming at the most accurate representation possible, the other seeking to capture the essence of the system's dynamics in a parsimonious way** (Table 3.1). The opposition is not irrevocable, as illustrated by the numerous hybrid formalisms that lie within the spectrum delimited by these two extremes such as fuzzy logic or discrete-time differential equations [Aldridge et al., 2009, Le Novere, 2015, Calzone et al., 2018]. To conclude, a comparison between the two approaches applied to the same problem is proposed by Calzone et al. [2018], studying the epithelio-mesenchymal transition (EMT, a biological process involved in cancer), to illustrate in concrete terms their

complementarity.

### 3.3.3 Some examples of complex features

With the help of these models, both qualitative and quantitative, many complex behaviours have been identified. Benefiting from the knowledge accumulated in the study of dynamic systems, a whole zoo of patterns with complex and non-intuitive behaviours such as non-linearities have been highlighted [Tyson et al., 2003]. The MAPK pathway, coarsely described in Figure 3.5, and often simplified as a rather unidirectional cascade, shows switch or bistability behaviors generated by the complexity of its multiple phosphorylation sites [Markevich et al., 2004]. These models have also been put at the service of understanding cancer and the erroneous decision-making by cells resulting from impaired signaling pathways. Thus, Tyson et al. [2011] summarize superbly well the complexity that can be hidden in the dynamics of smallest molecular networks as soon as they contain more than two entites and crossed regulations or feedback loops. Logical models have also made it possible to better dissect some complex phenomena at play in the cell such as emergent behaviours [Helikar et al., 2008] or mechanisms behind mutation patterns in cancer [Remy et al., 2015].

## 3.4 From mechanistic models to clinical impact?

Mechanistic models have therefore undeniably led to a better understanding of the complex molecular machinery of signalling pathways. But beyond the interest that this understanding represents, do these models also have a clinical utility? In other words, **are they of clinical or only scientific value?**

### 3.4.1 A new class of biomarkers

Throughout this thesis, the clinical value of mechanistic models will often be analyzed by analogy to that of biomarkers. Biomarkers are usually defined as measurable indicators of patient status or disease progression, such as prostate-specific antigen (PSA) for prostate cancer screening or BRCA1 mutation for breast cancer risk [Henry and Hayes, 2012]. Biomarkers also encompass multivariate signatures that identify more complex patterns with clinical significance. Taking the logic even further, it was therefore proposed that mechanistic models, which also reveal complex molecular behaviours,

could be considered as biomarkers, capturing perhaps even dynamic information [Fey et al., 2015].

Like oncology biomarkers, the models will be divided into two categories according to their clinical objectives: **prognostic models and predictive models** [Oldenhuis et al., 2008]. Prognostic biomarkers and models are those that provide information on the evolution of cancer independently of treatment. They are therefore generally confronted with survival or relapse data. The protein Ki-67 for example, encoded by the MKI67 gene, is known to be indicative of the level of proliferation and high levels of expression are thus associated with a poorer prognosis in many cancers [Sawyers, 2008]. Predictive biomarkers and models, on the other hand, give an indication of the effect of a therapeutic strategy. The simplest example, but not the only one, concerns biomarkers that are themselves the target of treatment: treatments based on monoclonal antibodies directed against HER2 receptors in breast cancer are only effective if the HER2 receptor has been detected in the patient [Sawyers, 2008]. Without attempting to be exhaustive, some logical and ODE models, with either prognostic or predictive claims, will be described.

### 3.4.2 Prognostic models

One of the first mechanistic models of cell signalling to have been explicitly presented as a prognostic biomarker is the one proposed by Fey et al. [2015] and describing c-Jun N-terminal kinase (JNK) pathway in neuroblastoma cells. A summary of the study is provided in Figure 3.6. The model is an ODE translation of the process description network of Figure 3.6A, further determined and calibrated with molecular biology experimental data obtained using neuroblastoma cell lines. We thus observe the non-linear switch-like dynamics of JNK activation as a function of cellular stress (Figure 3.6B). The precise characteristics of this sigmoidal response can, however, vary from one individual to another as captured by the network output descriptors $A$, $K_{50}$ and $H$. Fey et al. proposed to perform neuroblastoma patient–specific simulations of the model, using patient gene expressions for ZAK, MKK4, MKK7, JNK and AKT genes to specify the initial conditions of the ODE system. Since JNK activation induces cell death through apoptosis, the patient-specific $A$, $K_{50}$ and $H$ derived from patient-specifc models are then analyzed as prognostic biomarkers (Figure 3.6C). Readers are invited to refer to the original article for details on model calibration or binarization of network descriptors [Fey et al., 2015]. The authors also showed that in the absence of positive feedback from $JNK^{**}$ to ${}^{P}MKK7$,

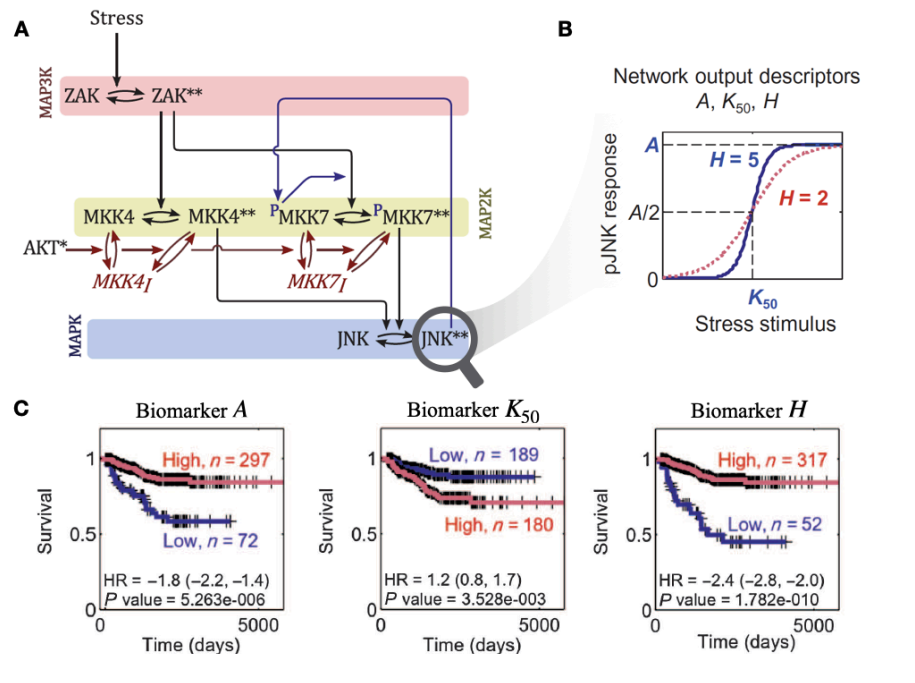



Figure 3.6: **Mechanistic modeling of JNK pathway and survival of neuroblastoma patients, as described by Fey et al. [2015].** (A) Schematic representation, as a process description, for the ODE model of JNK pathway. (B) Response curve (phosphorylated JNK) as a function of the input stimulus (Stress) and characterization of the corresponding sigmoidal function with maximal amplitude $A$, Hill exponent $H$ and activation threshold $K_{50}$. (C) Survival curves for neuroblastoma patients based on binarized $A$, $K_{50}$ and $H$; binarization thresholds having been defined based on optimization screening on calibration cohort.

an important component of non-linearity, the prognostic value is drastically decreased. All in all, this pipeline from ODE model to survival curves, thus provides a **paradigmatic example of the clinical interpretation of mechanistic models of molecular networks** that will be reused in later chapters for illustration purposes. Other ODE models following a similar rationale have been proposed by the same group for colorectal cancer [Hector et al., 2012, Salvucci et al., 2017] or glioblastoma [Murphy et al., 2013, Salvucci et al., 2019b]. Machine learning approaches have also been proposed to ease the clinical implementation of this kind of prognostic models by dealing with the potential lack of patient data needed to personalize them [Salvucci et al., 2019a].

On the logical modeling side, there are also studies including prognostic value validation. Thus, Khan et al. [2017] proposed two logical models of epithelio-mesenchymal transition (EMT) in bladder and breast cancers. These models are inferred from prior mechanisms knowledge and data analysis with particular attention to potential feedback loops. Using these models, it is possible to study the behaviour of them for all combinations of model inputs (growth factors and receptor proteins) and derive subsequent signatures for good or bad prognosis. These signatures are later validated with cohorts of patients. In this case, the mechanistic model does not seek to capture a dynamic behavior but to facilitate and **make understandable the exploration of combinations of input signals that grow exponentially with the number of inputs considered**. Other formalisms, called pathway activity analysis and following the same activity flows principles (Figure 3.5A), have been analysed in the light of their prognostic value. Their greater flexibility enables the direct use of networks of several hundred or thousands of genes, such as those present in the KEGG database [Kanehisa et al., 2012]. The benefit of mechanistic modeling is then to organize high-dimensional data and to facilitate the *a posteriori* analysis of the results.

### 3.4.3 Predictive models

But the explicit representation of biological entities in mechanistic models makes them particularly **suitable for the study of well-defined perturbations such as drug effects**. Indeed, by assuming that the mechanism of action of a drug is at least partially known, it is possible to integrate this mechanism into the model if it contains the target of the drug (Figure 3.7). One can therefore simulate the effect of one drug or even compare several. These strategies have already been implemented in a qualitative way with logical models used to explain resistance to certain treatments of breast cancer [Zañudo et al., 2017] or even highlight the synergy of certain combinations of treatments in gastric cancer [Flobak et al., 2015]. The value of these models, however, is more scientific than clinical in that they focus on a single cell line or a restricted group of cell lines. The possibility to personalize the predictions or recommendations for different molecular profiles of cell lines or patients is therefore not obvious. Still within the context of logical formalism, Knijnenburg et al. [2016] proposed a broader approach: if their model needs to be trained, it can nevertheless provide an analytical framework for several hundred cell lines, while remaining within the scope of the training data to ensure the validity of predictions.

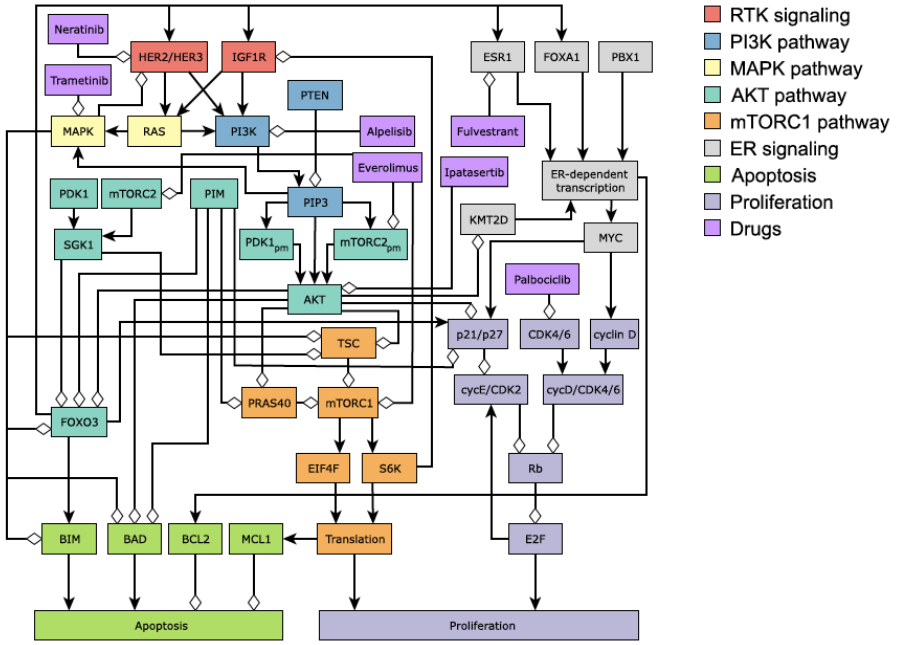



Figure 3.7: **Network model of oncogenic signal transduction in ER+ breast cancer, including some drugs and their targets.** Reprinted from Zañudo et al. [2017].

Conceptually comparable strategies can be found on the side of differential equations where large mechanistic models of cell signalling are also trained to predict the response to different treatments [Bouhaddou et al., 2018, Fröhlich et al., 2018]. A calibrated model can then predict the response to a combination of treatments not tested in the training data, thereby proving the ability of mechanistic models to extend their predictive value beyond the data [Fröhlich et al., 2018]. As with prognostic models, mechanistic approaches other than logical formalisms and ODEs have been proposed and validated [Jastrzebski et al., 2018]. What can be learned from these predictive models is that they require **significant training data to be able to go beyond qualitative predictions and dissect treatment response mechanisms of many cell lines simultaneously**. For obvious practical and ethical reasons, the validation of these models is for the moment limited to preclinical data since they require data for many uncertain therapeutic interventions.

This first bridge between mechanistic models of cell signalling and clinical applications concludes this introductory part. The next part will be

devoted to the definition of new methods to establish this connection based on logical formalism, before the third part proposes a more statistical evaluation of the prognostic and predictive values of the models presented in the previous parts.

# Part II

# Personalized logical models of cancer

Chapter 4

# Logical modeling principles and data integration

*"Je suis l'halluciné de la forêt des Nombres.*
*Ils me fixent, avec leurs yeux de leurs problèmes ;*
*Ils sont, pour éternellement rester : les mêmes.*
*Primordiaux et définis,*
*Ils tiennent le monde entre leurs infinis ;*
*Ils expliquent le fond et l'essence des choses,*
*Puisqu'à travers les temps planent leurs causes."*

*Émile Verhaeren (Les nombres, in Les Flambeaux noirs, 1891)*

Another way of ordering the diversity of mechanistic models presented above is to consider their relationship to biological data. Those that make little use of these data are essentially theoretical scope models that describe the general functioning of signaling pathways and associated systems [Calzone et al., 2010]. Other models propose more quantitative models but require much more data, either from databases or experimental data generated for this purpose in order to fit the parameters. In the latter case, the necessary data is usually perturbation data: how does my system react to this or that inhibition or activation? For a single cell line this

already corresponds to a large amount of data [Razzaq et al., 2018]. And if we want to extend these approaches to many cell lines, the amount of data becomes massive [Fröhlich et al., 2018]. For patient-specific models, access to this perturbation data is even more difficult.

Between theoretical models that are not very demanding in terms of data but not very applicable clinically and models with a clinical focus but very demanding in terms of data, an intermediate alternative is missing. **Can patient-specific mechanistic models be developed that would provide qualitative clinical interpretation with a small amount of data, accessible even in patients?** In this part, composed of three chapters, a middle way will be described to answer positively to this question. This methodology will be based on a historically qualitative mathematical formalism already presented in the previous chapter under the name of logical modeling. Logical modeling in general will be detailed in this chapter before describing an original personalized approach in the next two chapters.

**Scientific content**

This chapter presents the theoretical bases of logical modeling and the tools used thereafter. It does not present any original work but refers to the synthesis and analyses of logical modeling as described in Béal et al. [2019] and Béal et al. [2020].

## 4.1 Logical modeling paradigms for qualitative description

Mathematical models serve as tools to answer a biological question in a formal way, to detect blind spots and thus better understand a system, to organize, into a consensual and compact manner, information dispersed in different articles. In the light of this definition, logical formalism may seem one of the closest to natural language in that it **can translate quite directly the statements present in the literature** such as "protein A activates protein B" or "the expression of gene C requires the joint presence of factors D and E". Indeed, shortly after the first descriptions of control circuits by Jacob and Monod [1961], the interest of logical models to describe biological systems was put forward by Kauffman [1969] and Thomas [1973]. Since then, studies have multiplied [Thomas and d'Ari, 1990], varying the

fields of biological applications and also the mathematical and computational implementations [Naldi et al., 2018b]. The two subsections below summarize the characteristics common to most of the logical formalisms, before detailing the implementation chosen in this thesis in section 4.2. A review of the use of data in logic models will finally be proposed in section 4.3.

### 4.1.1 Regulatory graph and logical rules

A logical model is based on a network called **regulatory graph** (Figure 4.1), where each **node** represents a component (e.g. genes, proteins, complexes, phenotypes or processes), and is associated with discrete levels of activity (0, 1, or more when justified). The use of a discrete formalism in molecular network modeling relies on the highly non-linear nature of regulation, and thus on the existence of a regulatory threshold. Assuming that each variable represents a level of expression, it will take the value 0 if the level of expression of the entity is below the regulation threshold, i.e., insufficient to carry out the regulation; and the value 1 if it is above the threshold and regulation is possible. In other words, the control threshold discretizes the state space, here the expression levels. It is therefore possible to distinguish several thresholds for the same variable, corresponding to distinct controls that do not take place at the same expression levels. The variable is then multivalued. This extension greatly enriches the formalism, because it allows to distinguish situations that are qualitatively different and that would be confused with Boolean variables. In the continuation of this thesis, we will consider by default that the activity levels are binary, 0 corresponding to an inactive entity and 1 to an active entity. The **edges** of this regulatory graph correspond to influences, either positive or negative, which illustrate the possible interactions between two entities (Figure 4.1). Positive edges can represent the formation of active complexes, mediation of synthesis, catalysis, etc. and they will be later depicted as green arrows ($\leftarrow$). Negative edges on the other hand can represent inhibition of synthesis, degradation, inhibiting (de)phosphorylation, etc. and they will be depicted as red turnstiles ($\vdash$).

Then, each node of the regulatory graph has a corresponding Boolean variable associated to it. The variables can take two values: 0 for absent or inactive (OFF), and 1 for present or active (ON). These variables change their value according to a logical rule assigned to them. The state of a variable will thus depend on its **logical rule**, which is based on logical statements, i.e., on a function of the node regulators linked with logical

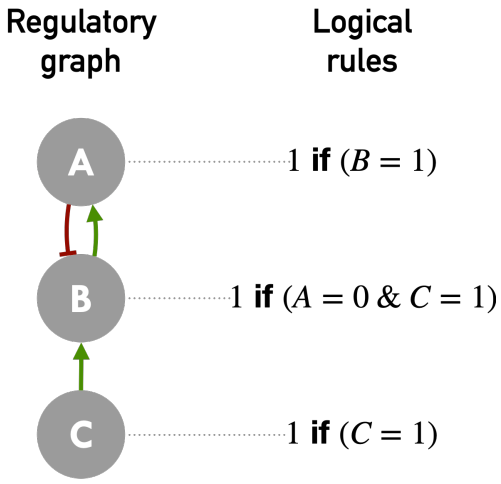



Figure 4.1: **A simple example of a logical model.** Regulatory graph on the left with positive (green) and negative regulations (red); a set of possible corresponding logical rules on the right.

connectors AND (&), OR (|) and NOT (!). These operators can account for what is known about the biology behind these edges. If two input nodes are needed for the activation of the target node, they will be linked by an AND gate; to list different means of activation of a node, an OR gate will be used; and negative influences will rely on NOT gates. The rules corresponding to the toy model in Figure 4.1 could be interpreted literally like this: A is activated to 1 if B is active; B is updated to 1 in the absence of A and the presence/activity of C; C is an input of the model and therefore not regulated. It can be noted that the logical rules cannot be deduced only from the regulatory graph, which is less precise and ambiguous. One could thus imagine that B is activated if C is, OR if A is not, thus changing the behavior of the model.

### 4.1.2 State transition graph and updates

In a Boolean framework, the variables associated to each node can take two values, either 0 or 1. We define a model state as a vector of all node states. All the possible transitions from any model state to another are dependent on the set of logical rules that define the model. These transitions can be viewed into a graph called a **state transition graph** (STG), where nodes are model states and edges are the transitions from one model state to another. STG nodes will be later depicted with rounded squares instead of circles in order to emphasize the difference with regulatory graphs. That

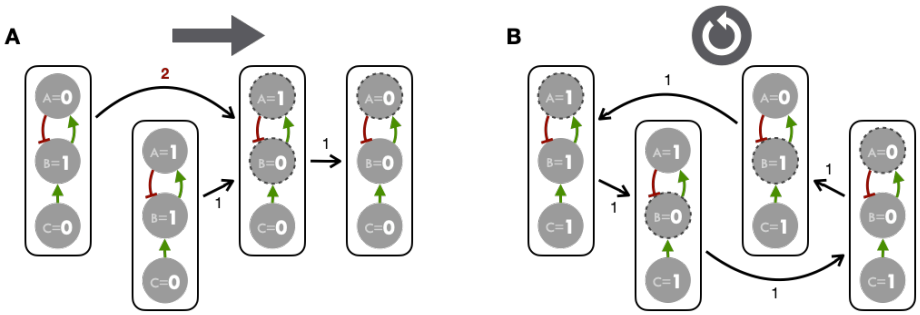



Figure 4.2: **State transition graph and synchronous updates.** Stable state (A) and limit cycle (B) attractors obtained for the example logical model with synchronous updates (all possible updates simultaneously). Figures above/below STG edges correspond to the number of nodes updated in each transition.

way, trajectories from an initial condition to all the final states can be determined. In a model with $n$ nodes, the STG can contain up to $2^n$ model state nodes; thus, if $n$ is too big, the construction and the visualization of the graph become difficult.

Based the simple logical model of Figure 4.1 it is nevertheless possible to represent the STG comprehensively. The idea for this is to start from a state of the system and track the successive states defined by the logical rules and the corresponding updates. The first strategy to construct this STG is to change simultaneously at each time step all the variables that can be changed (Figure 4.2). This method is referred to as a **synchronous updating strategy**. In the second method, referred to as a **asynchronous updating strategy**, variables are changed one at a time (Figure 4.3) and therefore each state has as many successors as there are components whose state must be changed according to logical rules (Figure 4.3A). Performing only one transition at a time implies having to choose this transition between the different possible options, which can be done according to pre-established rules or stochastically as explained in section 4.2. The latter asynchronous method will be used exclusively in the work presented thereafter.

We then define attractors of the model as long-term asymptotic behaviors of the system. Two types of attractors are identified: stable states, when the system has reached a model state whose successor in the transition graph is the model state itself; and cyclic attractors, when trajectories in the transition graph lead to a group of model states that are cycling.

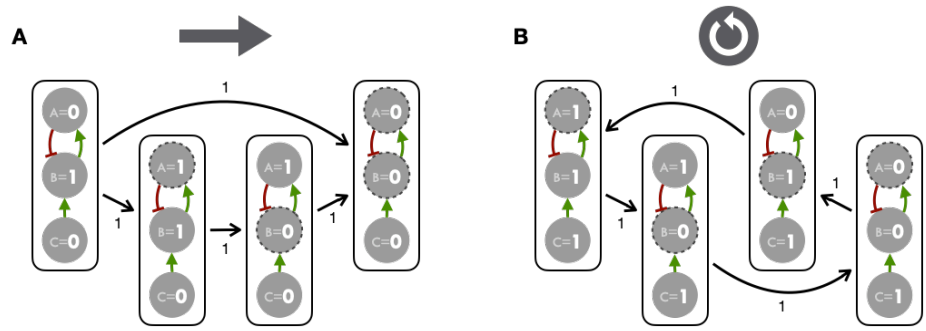



Figure 4.3: **State transition graph and asynchronous updates.** Stable state (A) and limit cycle (B) attractors obtained for the example logical model with asynchronous updates (one update at a time). Figures above/below STG edges correspond to the number of nodes updated in each transition.

For both synchronous and asynchronous updating strategies, the toy model shows the existence of **two types of attractors: a stable steady state and a limit cycle**, depending on the initial value of $C$. There are two disconnected components of the STG for this example that correspond to the two possible values for the input $C$. If $C$ is initially equal to 0 (inactive), then there exists only one stable state: $A = B = C = 0$. All the trajectories in the state transition graph lead to a single final model state. If $C$ is initially equal to 1, then the attractor is a limit cycle. The path in the STG cycles for any initial model state of this connected component. Note that for the asynchronous and synchronous graphs, the precise paths or limit cycles may differ. To conclude, it is important to emphasize and illustrate the characteristics of asynchronous updates in this toy example. In Figure 4.3A, the transition from the initial state ($A = C = 0; B = 1$) suggests two distinct possibilities, so it is necessary to **define additional rules or heuristics to choose between possible transitions**. We will come back to this by specifying the logical modeling implementation chosen in this thesis in section 4.2, in which a stochastic exploration of these alternatives is proposed.

### 4.1.3 Tools for logical modeling

Numerous tools have been developed to build logical models and study the dynamics of the systems under investigation, each with its own specificity. They allow, for example, to represent regulation networks; to edit, modify or infer logical rules; to identify stable states; to reduce models; to visualize

graphs of synchronous or asynchronous transitions. Some also allow to integrate temporal data; to discretize expression data; to simulate the model stochastically or to integrate delays; to identify existing models, etc. Among them, we can cite GINsim [Naldi et al., 2018a], BoolNet [Müssel et al., 2010], pyBoolNet [Klarner et al., 2016], BooleanNet [Albert et al., 2008], CellCollective [Helikar et al., 2012], bioLQM [Naldi, 2018], MaBoSS [Stoll et al., 2012, 2017], PINT [Paulevé, 2017], CaspoTS [Ostrowski et al., 2016], or CellNOptR [Terfve et al., 2012]. The interaction between all these tools, their interoperability and complementarity are highlighted in the form of a notebook Jupyter [Naldi et al., 2018b], and some of them are described in more details in section 4.3.

## 4.2 The MaBoSS framework for logical modeling

In the present study, all simulations have been performed with MaBoSS, a **Ma**rkovian **Bo**olean **S**tochastic **S**imulator whose design is summarized in Figure 4.4 and precisely described by Stoll et al. [2012] and Stoll et al. [2017]. This framework is based on an asynchronous update scheme combined with a continuous time feature obtained with Gillespie algorithm [Gillespie, 1976], allowing simulations to be continuous in time despite the discrete nature of logical modeling.

### 4.2.1 Continuous-time Markov processes

The implementation of asynchronous updates proposed by the MaBoSS software is based on **continuous time Markov processes applied on a Boolean state space**. The precise relations of this software with the Markovian formalism, and the associated demonstrations, are available in the original publication by Stoll et al. [2012] and detailed in particular in the related supplementary document[1]. The essence of the formalism, as set out in this article, is outlined below.

Before returning to the simplified example of Figure 4.4, it is possible to describe the general scheme of Markov processes and Gillespie algorithm for simulating logical models. We restrict ourselves here to a Boolean model (binary variables), composed of $n$ variables. The network state of the system is a vector $S$ of binary values such as $S_i \in \{0, 1\}$, with $i = 1, ..., n$ and $S_i$

[1] Additional information on Markov processes and MaBoSS in the following document

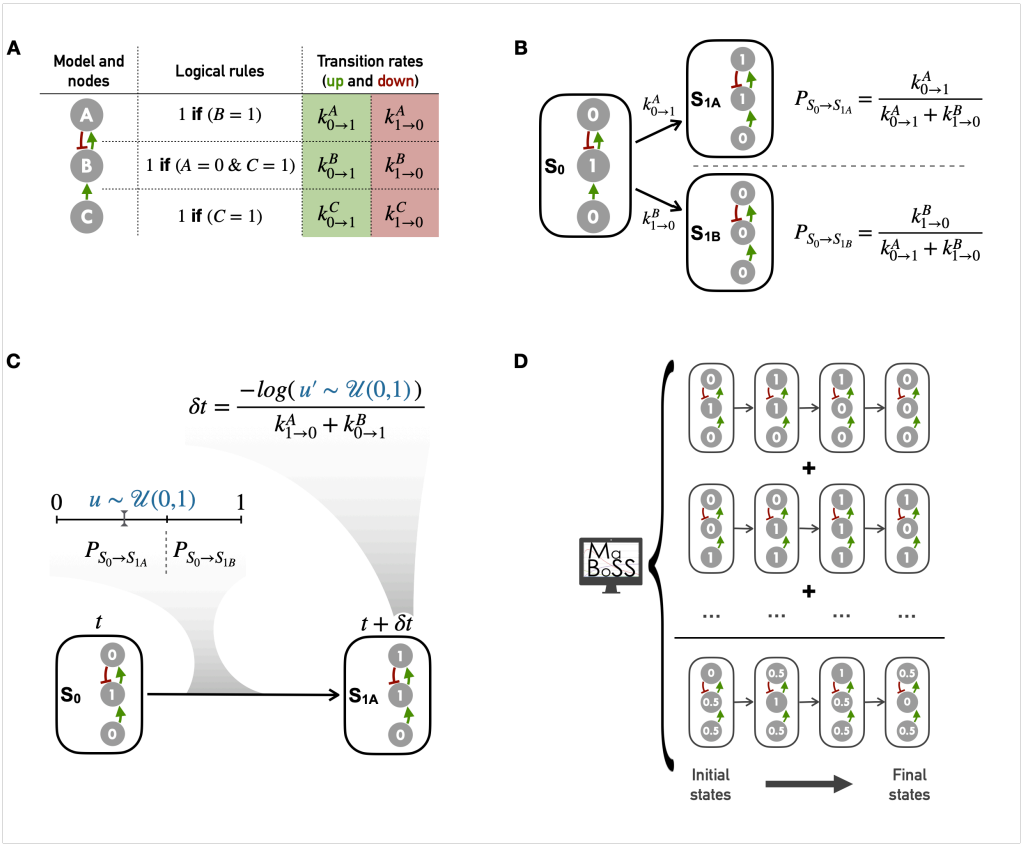



Figure 4.4: **Main principles of MaBoSS simulation framework and Gillespie algorithm.** (A) A logical model with regulatory graph, logical rules and transition rates. (B) A subset of the corresponding state transition graph with two possible transitions in asynchronous update for a given initial state; the probability probability associated with each transition comes from Gillespie algorithm. (C) Schematic diagram of the Gillespie algorithm applied to the asynchronous transition depicted in (B); random selection of a specific transition and time by the algorithm from two uniform random variables. (D) Schematic representation of a logical model simulation with MaBoSS: average trajectory obtained from the mean of many individual stochastic trajectories, each resulting from the succession of transitions as represented in (C).

representing the state of the node $i$. The network state space of all possible network states is called $\Omega$. The evolution of the state of the system can be represented by a stochastic process $s : t \rightarrow s(t)$ defined on $t \in I \subset \mathbb{R}$ applied on the network state space. For each time $t \in I \subset \mathbb{R}$, $s(t)$ represents a random variable applied on the network state space. Thus, $P[s(t) = S] \in [0, 1], \forall S \in \Omega$ and $\sum_{S \in \Omega} P[s(t) = S] = 1$. This **stochastic process verifies the Markov property which stipulates the absence of memory**, *i.e.*, the conditional probability distribution of future states of the process depends only on the present state, not on the sequence of past states. The resulting stochastic process, called Markov process is defined by an initial condition $P[s(0) = S], \forall S \in \Omega$ and the conditional probabilities $P[s(t) = S|s(t') = S'], \forall S, S' \in \Omega, \forall t, t' \in I, t' < t$. In this work we will focus on continuous time Markov processes where it has been shown that all conditional probabilities are functions of transition rates $\rho_{(S' \rightarrow S)}(t) \in [0, \infty[$ [Van Kampen, 2004]. Only the case of **time independent transition rates** (and Markov processes) will be explored below.

It is then possible to re-state the description of the logical models in this formalism. One of the first representations of the model is the regulatory network with a set of directed arrows linking the $n$ nodes (Figure 4.4A, left column). For each node $i$, a logical rule $L_i(S)$ is defined based only on the nodes $j$ for which there exists an arrow from node $j$ to $i$, as represented in Figure 4.4A middle column, *e.g.*, $L_B =$ (NOT $A$) AND ($C$) with $L_B$ the logical rule of node $B$. The notion of **asynchronous transition** can then be defined as a pair of network states $(S, S') \in \Omega$, written $(S \rightarrow S')$ such that:

$$S'_j = L_B(S) \text{ for a given } j$$
$$S'_i = S_i \text{ for } i \neq j$$

This means that during an asynchronous transition, only one node changes state, among those whose logical rule allows it. In Figure 4.4B, two possible asynchronous transitions from the same initial state are represented. Then, transition rates $\rho_{(S \rightarrow S')}$ are non-zero only if $S$ and $S'$ differ by only one node. In that case, each Boolean logic $L_i(S)$ is replaced by two functions $k^i_{0 \rightarrow 1/1 \rightarrow 0}(S) \in [0, \infty[$. The transition rates are defined as follows: if $i$ is the node that differs from $S$ and $S'$, then:

$$\rho_{(S \rightarrow S')} = k^i_{0 \rightarrow 1}(S) \text{ if } S_i = 0$$

$$\rho_{(S \to S')} = k^i_{1 \to 0}(S) \text{ if } S_i = 1$$

Therefore, the **continuous Markov process is completely defined by all these $k^i$ and an initial condition**. The state transition graph can also be re-defined as a graph in $\Omega$, with an edge between $S$ and $S'$ if and only if $\rho_{S \to S'} > 0$.

### 4.2.2 Gillespie algorithm

Although the Markov process is completely defined, the cost of its theoretical resolution increases exponentially with the number of nodes since the transition matrix of the system is of size $2^n \times 2^n$. One solution consists in **sampling the probability space by simulating time trajectories in the state transition graph**, which is what MaBoSS performs through Gillespie algorithm, also called kinetic Monte-Carlo. The principle of the algorithm is to compute a finite set of individual stochastic trajectories of the Markov process and to use them to infer probabilities for the given Markov process (Figure 4.4D). For the sake of simplicity the remainder of this section will describe the simulation of a single individual stochastic trajectory, first with a focus on the formal description of the algorithm and then with a more qualitative explanation.

A stochastic trajectory $\hat{S}(t)$ is a function from a predefined time interval $[0, t_{max}]$ to $\Omega$. The iterative computation of the trajectory is obtained as follows:

1. From a state $S$ at $t_0$, sum the rates of all possible transitions for leaving this state: $\rho_{total} = \sum_{S'} \rho_{(S \to S')}$
2. Draw two uniform random numbers $u, u' \in [0, 1]$
3. Compute the transition time $\delta t = -log(u)/\rho_{total}$
4. Order the potential target states $S'^{(j)}$ and their corresponding transition rates $\rho^{(j)} = \rho_{(S \to S'^j)}$
5. Choose the new state $S' = S'^{(k)}$ such that $\sum_{j=0}^{k-1} \rho^{(j)} < u' \rho_{total} < \sum_{j=0}^{k} \rho^{(j)}$, with $\rho^{(0)} = 0$
6. Define the trajectory $\hat{S}(t)$ such as $\hat{S}(t) = S$ for $t \in [t_0, t_0 + \delta t[$ and $\hat{S}(t_0 + \delta t) = S'$
7. Repeat iteratively from $S'$ until $t_{max}$ is reached

Thus, Gillespie algorithm provides a **stochastic way to choose a specific transition among several possible ones** and to infer a corre-

sponding time for this transition. To achieve this, transition rates seen as qualitative activation or inactivation rates, must be specified for each node (Figure 4.4A). They can be set either all to the same value by default, in the absence of any indication, or in various levels reflecting different orders of magnitude: post-translational modifications are quicker than transcriptions for instance. These transition rates are translated as **transition probabilities** in order to determine the actual transition (Figure 4.4B). Indeed, the probability for each possible transition to be chosen for the next update is the ratio of its transition rate to the sum of rates of all possible transitions. Higher rates correspond to transitions that will take place with greater probability, or in other words more quickly. At each update, the Gillespie algorithm performs the procedure described above and depicted schematically in Figure 4.4C. Two uniform random variables $u$ and $u'$ are drawn and used respectively to select the transition among the different possibilities (with $u$) and to infer the corresponding time (with $u'$). Based on the described formula, time $\delta t$ follows an exponential law whose average is equal to the inverse of the sum of all possible transition rates (Figure 4.4C). In the present work, except otherwise stated, all transition states will be initially assigned to 1.

### 4.2.3 A stochastic exploration of model behaviours

Since MaBoSS computes stochastic trajectories, it is relevant to compute several trajectories in order to get an insight of the average behavior by generating a **population of stochastic trajectories** over the asynchronous state transition graph (Figure 4.4D). The aggregation of stochastic trajectories can also be interpreted as a description of a heterogeneous population. In fact, in all the examples in next chapters, all simulations have consisted of thousands of computed trajectories. The larger the model, the larger the space of possibilities and the more trajectories are required to explore it. Since several trajectories are simulated, initial values of each node can be defined with a continuous value between 0 and 1 representing the probability for the node to be defined to 1 for each new trajectory. For instance, a node with a 0.6 initial condition will be set to 1 in 60% of simulated trajectories and to 0 in 40% of the cases.

In the present work, we will focus on the *asymptotic* state of these simulations instead of transient dynamics and we will call **node scores** the asymptotic agregated score obtained by averaging all trajectories at a given final time point. Indeed, asymptotic states are more closely related to logical model attractors than transient dynamics and are therefore less dependent

on updating stochasticity and more biologically meaningful [Huang et al., 2009]. Note that the simulation time should be chosen carefully to ensure that the asymptotic state is achieved, and the term *final state* may be considered as safer. All in all this modeling framework is at the intersection of logical modeling and continuous dynamic modeling. If the definition of time remains rather abstract and difficult to interpret experimentally, the stochastic exploration of trajectories makes it possible to refine the purely binary interpretation of the variables.

### 4.2.4 From theoretical models to data models?

To sum up, logical formalism makes it possible to design qualitative models that reflect *a priori* knowledge of the phenomena being studied. Thus, they allow answering questions for which there is little quantitative information on the precise mechanisms involved in a disease, for instance a lack of data related to the expression of genes, the quantity of key proteins or the speed of certain processes. Logical models can confirm that a network is a good illustration of the underlying biological question. However, the construction of the model from the literature often results in a generic consensus network that cannot explain the differences observed between certain patients with different molecular profiles. In order to propose a patient-specific mechanistic approach, it seems crucial to **use the biological data** available. How is this possible in a formalism that is by definition quite abstract?

## 4.3 Data integration and semi-quantitative logical modeling

The higher level of abstraction of the logical formalism sometimes makes the necessary back and forth between theoretical modeling and experimental or clinical data less easy. However, many theoretical approaches have been developed over the years to enable this dialogue at all stages, from the construction to the validation of a logical model, as summarized in Figure 4.5. This section summarizes some of these approaches to show **how the use of biological data enriches logical models and brings them closer to clinical applications** in precision medicine. The purpose of this presentation is also to better contextualize the original approach presented in the following chapter. It should be noted that the methods presented below are all applicable to logical models, and illustrated with such examples where possible. However, some methods are not specific to this formalism and can be applied to other modeling frameworks.

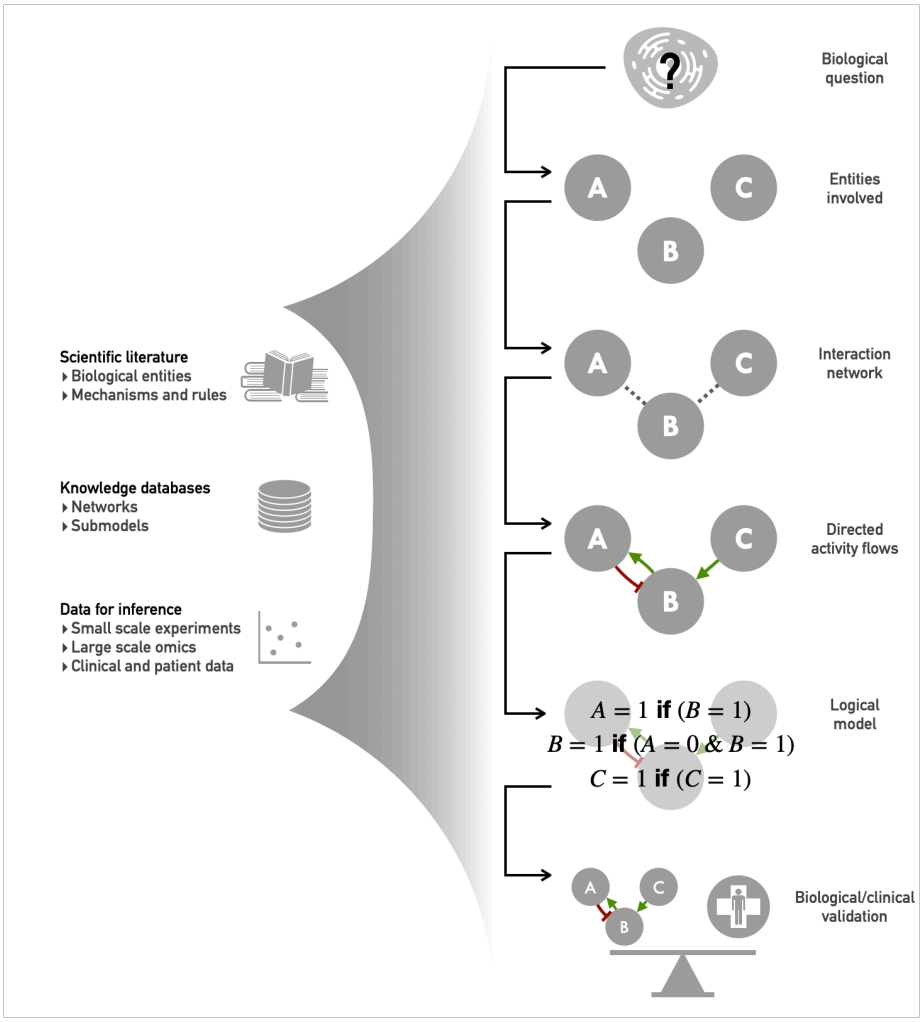



Figure 4.5: **Data integration in logical modeling.** The main types of data used are shown on the left; the essential steps of the logical modeling are shown linearly on the right.

### 4.3.1 Build the regulatory graph

Faced with a biological question (Figure 4.5, first step), it is crucial to identify the main actors in the process in order to **define the outline of the model** (Figure 4.5, second step). A first approach relies on the existing scientific literature on the topic: which biological species and which interactions have been identified as relevant to my problem? In a more automatic way, it is possible to extract information from different databases in order to establish an initial list of biological entities and interactions associated with a biological phenomenon or even a gene of interest [Kanehisa et al., 2012, Perfetto et al., 2016]. As an example, starting from the study of E2F1 gene as the hub of many regulatory mechanisms, Khan et al. [2017] have reconstructed a dense network of interactions in the vicinity of E2F1, which will be used for the construction of their subsequent model. The main difficulty here is to choose and select the relevant biological information adapted to the context of the model to be created, depending for example on the type of cancer studied or the desired level of precision.

But if the literature can be considered as processed data, it is also possible to use directly experimental data related to the problem under study. Key actors of biological processes identified by statistical analysis, such as differentially expressed genes or the most frequently mutated genes in a patient cohort, are selected and used as a starting point for the construction of the model [Remy et al., 2015]. More comprehensive approaches can use differential analysis tools on signaling pathways, rather than individual genes, to choose the relevant processes to include by contrasting different groups of patients based on their grades, metstatic status, resistance to treatments etc. [Martignetti et al., 2016, Montagud et al., 2017]. Similarly, the study of regulatory networks involving transcription factors may justify the use of ChIP-seq data to identify possible new transcriptional regulations not previously listed [Collombet et al., 2017].

Once the main actors have been identified, it is necessary to **infer the links** between them (Figure 4.5, third and fourth step). However, starting from a list of genes and proteins of interest, how can we ensure that the regulatory relationships are complete and relevant? While a careful reading of the literature can provide locally interesting information, the use of omics data is also a resource that can be broken down into different levels of precision. The major interest of these methods, assuming that the data are adequate and sufficiently massive, is to be able to extract information as large as the dataset, potentially on hundreds of entities, and above

all specific to the object of study: a cancer subtype or a particular cell line can thus generate their own interaction network [Lefebvre et al., 2010]. Inference methods extract biological knowledge hidden in large databases, summarize it and represent it via networks. Many methods construct co-expression networks, which are non-oriented graphs, with different metrics and methods [Margolin et al., 2006, Vert et al., 2007]. Other approaches seek to infer causal relations between components, allowing the reconstruction of directed graphs where the links between entities are oriented, and sometimes even signed as activating (positive) or inhibiting (negative) regulations. These methods often make use of time series [Hill et al., 2016] or perturbation data [Meinshausen et al., 2016], but also more recently from observational data [Verny et al., 2017]. The information extracted from the data is then directly readable in the form of activity flows as described in the SBGN standards [Novère et al., 2009], thus providing a representation adapted to the construction of qualitative models and *a fortiori* of logical models [Le Novere, 2015]. Closer to the objective of defining logical models, certain methods allow the study and inference of co-regulation expressed with logical operators [Elati et al., 2007], thus facilitating the passage from the definition of an interaction network to the construction of a true logical model.

### 4.3.2 Define the logical rules

Precision must then be taken further by defining the logical rules that complete the network (Figure 4.5, fifth step). The first source of aggregated data to define logical rules is the scientific literature. The modeler looks for the state of knowledge on a given regulatory mechanism and translates it into a **local logical rule**, according to the desired level of precision. For example, it has been observed that the protein kinase AKT can stabilize the oncogene MDM2 by phosphorylation, which leads to the degradation of p53 by forming a complex with it: this example can be translated by a simple inhibition relationship of AKT on p53 if this level of precision is considered sufficient or else intermediate species such as MDM2 can be used [Cohen et al., 2015]. Then, the effect of inhibition must be defined: can MDM2 alone inhibit p53 or does the presence of other activators outweigh this effect? This kind of considerations allows to define the logical combinations between the different inputs of a network node. In some cases, experimental data can be used to answer such questions: is a single activator sufficient or is the presence of all activators necessary? Which of the activator or inhibitor prevails in the case of simultaneous presence? While this information is often found in the literature, one should generate one's

own experimental data to ensure an answer tailored to the study context, using a variety of experimental molecular biology techniques. For example, in order to elucidate the relationship between Foxo1 and Cebpa in a model of differentiation of myeloid and lymphoid cells, Collombet et al. [2017] first established the physical relationship between these species by ChIP-seq before determining the nature of this relationship using an ectopic expression experiment of Foxo1 in macrophage cells.

Other, more global approaches have been developed in recent years, driven by the influx of data from high-throughput sequencing techniques. Based on this rich and complex data, it has become possible to **infer entire logical models, with precisely defined rules and interactions** [Ostrowski et al., 2016]. The algorithms CellNOpt [Terfve et al., 2012] and caspo [Videla et al., 2017] provide two examples of these approaches, and more recently the SCNS tool described a graphical interface to infer logical models from single cell data [Woodhouse et al., 2018]. This model-inference goes beyond simpler structure-inference by defining the logical rules, but it is generally based on a predefined topological structure to which time series or perturbation data are added. These data provide access to the response dynamics of a system. By questioning the way the system reacts, these data are therefore richer than a snapshot and thus facilitate the transition from correlation to causality, and thus the inference of logical rules. In practice, the use of proteomic or phospho-proteomic data is often recommended because these data account for the activity of the protein and are in fact the closest to the cellular response [Ostrowski et al., 2016, Terfve et al., 2012, 2015]. In spite of the richness of this type of data, model inference is sometimes still an under-determined problem that can lead to a large number of models with different logical rules equally compatible with the data. In such situations, it is then a matter of choosing the model on the basis of biological relevance criteria or of accepting to use families of models, or ensemble models, instead of limiting oneself to a single model [Videla et al., 2017]. In all cases, constructing logical rules directly from data specific to the problem can make it possible to obtain logical rules that are also specific to the context or the system under study [Saez-Rodriguez et al., 2011b]. For example, the inference of logical models specific to one or some cancer cell lines is a powerful tool to study their particularities [Razzaq et al., 2018].

### 4.3.3 Validate the model

Finally, the data can be used to validate the **biological or clinical relevance of the models** (Figure 4.5, sixth step). Compared to a system of differential equations, logical modeling has the particularity of being more abstract and therefore less directly reliable to an experimental reality for its validation. A system of differential equations can be compared to the chemical kinetics of the biological system under study. Compared to continuous formalisms, the dynamics of logic model simulation is more difficult to take into account but it is possible to verify it qualitatively, for example by validating the cyclic nature of activation trajectories for a model simulating the cell cycle [Fauré et al., 2006] or cellular decisions as a function of the activation signal [Calzone et al., 2010]. A second, more frequent approach consists in looking at the model's steady states and associating them with physiological conditions [Weinstein et al., 2017, Cohen et al., 2015]. A third strategy focuses on the asymptotic state reached during the stochastic simulation of the model(s), a state representing a mixture of the different steady states according to the probability that the model has of reaching them. It is also possible, in some model checking frameworks, to study the ability of models to reach certains states, interpreted as cellular types, from given initial conditions [Abou-Jaoudé et al., 2015].

In many models, to facilitate the analysis, **nodes representing phenotypes have been added as "read-out" of the activity of certain entities**. Thus, if a model includes a node named *Proliferation*, it will then be simpler to draw interpretations from the simulations performed with the model that will be linked to experimental observations of tumor growth or cell proliferation [Grieco et al., 2013, Steinway et al., 2015]. To validate these models, the activity of phenotypes when forcing some node activity to 0 or 1 is compared with the results of gene mutations reported in experiments carried out on mice or cell lines [Fauré et al., 2006, Cohen et al., 2015]. Another similar method for validating the relevance of a logical model is based on the analysis of the effects of different therapeutic molecules. The mechanistic nature of logical modeling makes it possible to simulate the effect of these molecules, at least for targeted treatments with known mechanisms of action. It is then possible to simulate the effect of an inhibitory molecule by forcing the activity of its target to 0 and to compare with data [Zañudo et al., 2017, Iorio et al., 2016, Knijnenburg et al., 2016].

Beyond validation, some studies have predicted **new therapeutic targets** based on logical models, for instance by pointing out weaknesses in

the topology of a regulatory system [Sahin et al., 2009]. Taking advantage of the versatility of the formalism to study combinations of therapeutic molecules, logical modeling has also proved fruitful in predicting the best therapeutic combinations and their synergies, in the context of gastric cancers for example [Flobak et al., 2015]. Experimental confirmation of the predictions resulting from the modeling is then the ultimate stage in the validation of a logical model, completing the fruitful round trip between models and data.

CHAPTER 5

# Personalization of logical models: method and prognostic validation

*"All happy families are alike; each unhappy family is unhappy in its own way."*

*Leo Tolstoy (Anna Karenina, 1877)*

Now that logical modeling has been introduced, it is possible to come back to the question that structures this part and to refine it. **Is it possible to use routine omics data to obtain logical models that provide qualitative clinical interpretation?** We thus propose a sequential approach, separating the model construction process from the integration of biological data. A generic logical model is first built, based on the literature knowledge, and the data are then used to specify the model. Indeed, the model as defined from the literature is often generic in the sense that it summarizes the state of knowledge on a probably heterogeneous pathology or population. Assuming that this general regulatory scheme provides a relevant framework for the system, it may then be relevant to use more precise omics data to impose biologically sourced constraints on the model: inactivation of a gene in a patient, activation of a protein or a signalling pathway by overexpression or phosphorylation, etc. This ap-

proach, called **PROFILE** (PeRsonalization OF logIcaL ModEls), allows the integration of both discrete (mutations) and continuous data (RNA expression levels, proteins) based on the MaBoSS software, and leads to specific models of a cell line or a patient.

**Scientific content**

This chapter presents the method developed during the thesis to personalize logical models, i.e., generate patient-specific models from a single generic one. The principles of the method and some analyses on patient data have been comprehensively described in Béal et al. [2019] and briefly summarized in Béal et al. [2020]. Analyses on cell lines are unpublished.

## 5.1 From one generic model to data-specific models with PROFILE method

The PROFILE method is summarized in Figure 5.1 and the different steps are successively described in the following subsections.

### 5.1.1 Gathering knowledge and data

The first steps are therefore to build a logical model adapted to the biological question (Figure 5.1, upper left) and to collect omics data that will be used to personalize the model (Figure 5.1, upper right). The construction of the model can be based on literature or data (see previous chapter). In the latter case, the data used to build the model will preferably be distinct from those used to personalize the model.

#### 5.1.1.1 A generic logical model of cancer pathways

In this chapter, which is essentially methodological in nature, we will use a **published logical model of cancer pathways** to illustrate our PROFILE methodology. It is based on a regulatory network summarizing several key players and pathways involved in cancer mechanisms: RTKs, PI3K/AKT, WNT/$\beta$-catenin, TGF-$\beta$/Smads, Rb, HIF-1, p53 and ATM/ATR [Fumia and Martins, 2013]. The later analyses will be mainly focused on two read-out nodes, *Proliferation* and *Apoptosis*. Based on the model's logical rules *Proliferation* node is activated by any of the cyclins

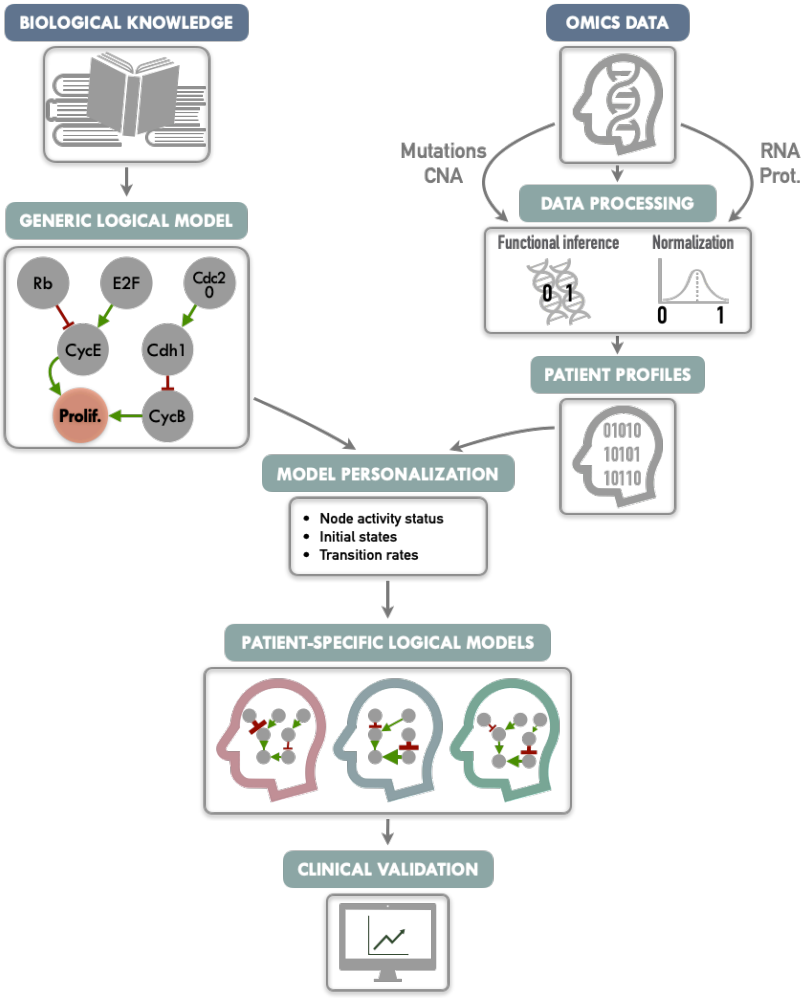



Figure 5.1: **Graphical abstract of PROFILE method to personalize logical models with omics data.** On the one hand (upper left), a generic logical model, in a MaBoSS format is derived from literature knowledge to serve as the starting-point. On the other hand (upper right), omics data are gathered (e.g., genome and transcriptome) as data frames, and processed through functional inference methods (for already discrete genome data) or binarization/normalization (for continuous expression data). The resulting patient profiles are used to perform model personalization, i.e., adapt the generic model with patient data. The merging of the generic model with the patient profiles creates a personalized MaBoSS model per patient. Then, biological or clinical relevance of these patient-specific models can be assessed.

(CyclinA, CyclinB, CyclinD, and CyclinE) and is, thus, an indicator of cyclin activity as an abstraction and simplification of the cell cycle behavior. *Apoptosis* node is regulated by Caspase8 and Caspase9. This generic model contains 98 nodes and 254 edges. Further details and visual representation are provided in section B.1 and Figure B.1. Model files are available in MaBoSS format in a dedicated GitHub repository.

#### 5.1.1.2 Cancer data to feed the models

In order to showcase the method, **breast-cancer patient data** are gathered from METABRIC studies [Curtis et al., 2012, Pereira et al., 2016]. 1904 patients have data for both mutations, copy number alterations, RNA expression and clinical status (e.g. survival). This number rises to 2504 patients if we only look at the mutations. Additional analyses were also performed based on the smaller and clinically less complete TCGA breast cancer data [TCGA et al., 2012]. These are detailed in Béal et al. [2019] but not included in this thesis. A more comprehensive description of these two databases can be found in section A.3.

In addition to these examples proposed in the original article, an application to **cell line data** is proposed in section 5.2.1 to link to the next chapters. A cohort of 663 cell lines from different types of cancer will be used. The data are from Cell Models Passports [van der Meer et al., 2019] and are described in more detail in the appendix A.1. In all cases, samples and cell lines will sometimes be referred to as patients for the sake of simplicity.

### 5.1.2 Adapting patient profiles to a logical model

Before describing precisely the methodologies for using the data to generate patient-specific models, it is important to understand that these data will need to be transformed. This is the transformation of raw omics data into processed profiles that can be used directly in logical modeling.

#### 5.1.2.1 Functional inference of discrete data

Since the logical formalism is itself discrete, the integration of discrete data is more straightforward, at least at the first glance. The most natural idea, used in many previous works, is to **interpret the functional effect of these alterations** and to encode it directly in the model. For instance, a deleterious mutation is integrated into the model by setting the corresponding node to 0 and ignoring the logical rule associated to it. For activating

mutation, the node is set to 1. The main obstacle is therefore to estimate the functional impact of the alterations in order to translate them as well as possible in the model.

For mutations, based on the variant classification provided by the data, inactivating mutations (nonsense, frame-shift insertions or deletions and mutation in splice or translation start sites) are assumed to correspond to loss of function mutations and therefore the corresponding nodes of the model are forced to 0. Then, missense mutations are matched with OncoKB database [Chakravarty et al., 2017]: for each mutation present in the database, an effect is assessed (gain or loss of function assigned to 1 and 0, respectively) with a corresponding confidence based on expert and literature knowledge. Then, mutations targeting oncogenes (resp. tumor-suppressor genes), as defined in the 2020+ driver gene prediction method [Tokheim et al., 2016], are assumed to be gain of function mutations (resp. loss of function) and therefore assigned to 1 (resp. 0). To rule potential passenger mutations out, each automatic assignment of a oncogene/tumor-suppressor gene muations requires that the effect of the mutation has been identified as significant by predictive software based on protein structure such as SIFT [Kumar et al., 2009] or PolyPhen [Adzhubei et al., 2010].

For integration of copy number alterations, we use the discrete estimation of gain and loss of copies from GISTIC algorithm processing [Mermel et al., 2011]. The loss of both alleles of a gene (labelled -2) can thus be interpreted as a 0. Conversely, a significant gain of copies (labelled +2) denotes a gene that tends to be more highly expressed although the interpretation is more uncertain.

#### 5.1.2.2 Normalization of continuous data

The integration of continuous data, such as RNA expression levels, in logical modeling is more difficult. The stochastic framework of MaBoSS provides however some possibilities. The main continuous mechanistic parameters of MaBoSS are the initial conditions of each node (its initial probability of being activated among the set of simulated stochastic trajectories) and the transition rates associated with the nodes (its probability to have its transition performed in an asynchronous update). In order to facilitate the use of continuous data through one of these two possibilities, we propose to transform them so that the **values are continuous between 0 and 1**, what we will refer to hereafter as normalized data. **It is assumed that these continuous data can be good proxies of biological activity**, 0

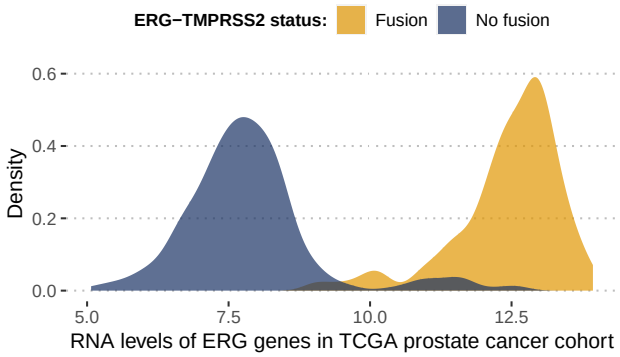



Figure 5.2: **Bimodal distribution of ERG gene in TCGA prostate cancer cohort.** This bimodality is largely explained by the fusion status of ERG gene. Patients for whom the gene has fused with TMPRSS2 have a much higher level of RNA expression for ERG.

corresponding to a very low level of activity of the biological entity and 1 to a very high level. This assumption will have to be explained and justified each time: high level of expression of an RNA or significant phosphorylation of a protein interpreted as continuous markers of an important biological activity for example.

One of the assumptions of our analysis is that the interpretation of continuous data can only be relative and not absolute. It is indeed difficult to define an absolute threshold of RNA level at which a gene will be considered as activated. This may depend on contexts, technologies or even the way in which the data have been processed. On the other hand, it is possible to estimate that a gene is over-expressed for a patient compared to a cohort of interest. In contrast, the effect of a mutation can be estimated more independently. Thus, the **continuous data will be normalized for the whole cohort studied**, for each gene individually. In order to retain biological information as much as possible, distribution patterns are identified and normalized in different ways (Figure 5.4). We will illustrate the process by taking the example of the expression data expressed with continuous RNA levels. Beforehand, genes with no variation in expression level or too many missing values are discarded from the analysis. Then, we seek to identify first the genes that have a **bimodal** distribution. Indeed, these naturally fit into a binary formalism and this bimodality often has an underlying biological explanation. As an example, in the TCGA prostate cancer cohort (used in section 6.3), a gene called ERG has a bimodal distribution when looking at RNA levels in all patients. This distribution is

almost entirely explained by an underlying genetic alteration that is the fusion of the ERG gene with the TMPRSS2 gene promoter (Figure 5.2), which is very common in this cancer [Tomlins et al., 2005]. In the data we identify bimodal patterns based on three distinct criteria: **Hartigan's dip test of unimodality, Bimodality Index (BI) and kurtosis**. The dip test measures multi-modality in a sample using the maximum difference between empirical distribution and the best unimodal distribution, i.e., the one that minimizes this maximum difference [Hartigan and Hartigan, 1985]. Values below 0.05 indicate a significant multi-modality. In PROFILE, this dip statistic is computed using the R package *diptest*. The Bimodality Index (BI) evaluates the ability to fit two distinct Gaussian components with equal variance [Wang et al., 2009]. Once the best 2-Gaussian fit is determined, along with the respective means $\mu_0$ and $\mu_1$ and common variance $\sigma$, the standardized distance $\delta$ between the two populations is given by

$$\delta = \frac{|\mu_0 - \mu_1|}{\sigma}$$

with $\mu_i$ the mean of Gaussian component $i$, and the BI is defined by

$$BI = [p(1-p)]^{1/2}\delta$$

where $p$ is the proportion of observations in the first component. In PROFILE, BI is computed using the R package *mclust*. Finally, the kurtosis method corresponds to a descriptor of the shape of the distribution, of its tailedness, or non-Gaussianity. A negative kurtosis distribution, especially, defines platykurtic (flattened) distributions, and potentially bimodal distributions. It has been proposed as a tool to identify small outliers subgroups or major subdivisions [Teschendorff et al., 2006]. In our case, we focus on negative kurtosis distributions to rule out non-relevant bimodal distributions composed of a major mode and a very small outliers' group or a single outlier. Although Dip test, BI and negative kurtosis criteria emerge as similar tools in the sense that they select genes whose values can be clustered in two distinct groups of comparable size, we choose to combine them in order to correct their respective limits and increase the robustness of our method. For that, we consider that **all three conditions (Dip test, Bimodality Index and kurtosis) must be fulfilled in order for a gene to be considered as bimodal**. The thresholds of each test are inspired by those advocated in the papers presenting the tools individually. Dip test

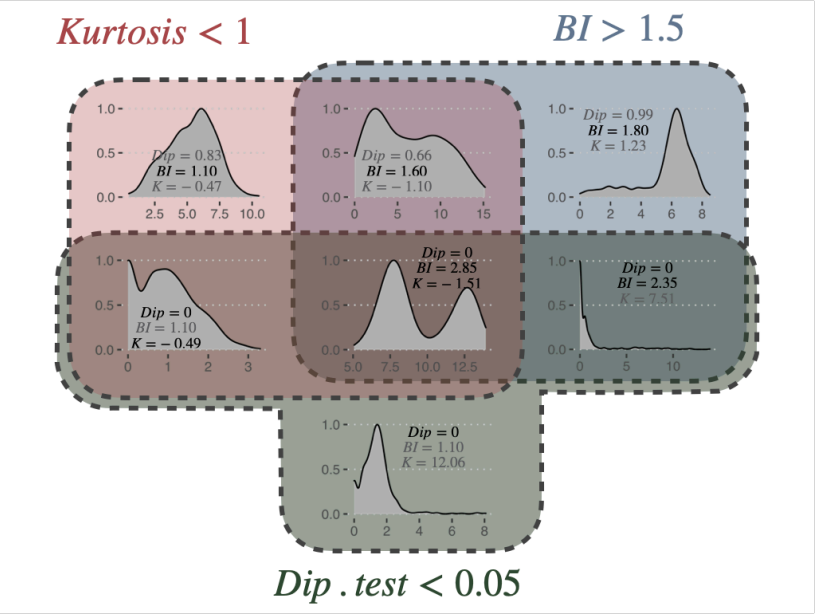



Figure 5.3: **Bimodality criteria and their combinations.** Examples of gene expression distributions for the different combinations of bimodality criteria: Dip test, Bimodality Index (BI) and kurtosis (K). Plots are organized in a Venn diagram.

is a statistical test to which the classical 0.05 threshold has been chosen. In the article describing BI, authors explored a cut-off range between 1.1 and 1.5 and we chose 1.5 for the present work. Regarding kurtosis, the usual cut-off is 0, but since this criterion does not directly target bimodality, this criterion has been relaxed to $K < 1$. Several examples of the relative differences and complementarities between these criteria can be seen in Figure 5.3.

Non-bimodal genes are further classified as unimodal or zero-inflated distributions, looking at the position of the distribution density peak (Figure 5.4A). Then, based on this three category classification of genes, a **pattern-preserving normalization** can be performed, as summarized in Figure 5.4B. For a bimodal gene $i$, a 2-component Gaussian mixture model is fitted using *mclust* R package resulting in a *lower* component $C_{i,0}$ (with mean $\mu$) and an *upper* component $C_{i,1}$. Denoting $X_{i,j}$ the expression value for gene $i$ and sample $j$, $X_{i,j}$ has a probability to belong to $C_{i,0}$ or $C_{i,1}$ such as $P[X_{i,j} \in C_{i,0}] + P[X_{i,j} \in C_{i,1}] = 1$. These probabilities result from posterior inference using Bayes' rule. For bimodal genes, the normalization

processing is therefore defined as:

$$X_{i,j}^{norm} = P[X_{i,j} \in C_{i,1}]$$

For unimodal distributions, we transform data through a sigmoid function in order to maintain the most common pattern which is unimodal and nearly-symmetric:

$$X_{i,j}^{norm} = \frac{1}{1 + e^{-\lambda(X_{i,j} - median(X_i))}}$$

Since the slope of the function depends on $\lambda$, we adapt it to the dispersion of initial data in order to maintain a significant dispersion in $[0, 1]$ interval: more dispersed unimodal distributions are mapped with a gentle slope, peaked distributions with a steep one. We map the median absolute deviation $MAD(X_i) = median(|X_i - median(X_i)|)$ on both sides of the median respectively to 0.25 and 0.75 to ensure a minimal dispersion of the mapping. Thus, the proposed mapping results in:

$$\lambda = \frac{log(3)}{MAD(X_i)}$$

Last, zero-inflated distributions are transformed by linear normalization of the initial distribution:

$$X_{i,j}^{norm} = \frac{X_{i,j} - min(X_i)}{max(X_i - min(X_i))}$$

The transformation is applied to data between 1st and 99th quantiles to be more robust to outliers. Values outside this range are respectively assigned to 0 and 1. All the categoriation of distributions and the subsequent normalizations are summarized in Figure 5.4. With the help of the categories described here, it is also possible to binarize the continuous data quite simply. This binarization is required for some methods of network inference or logical modeling but will not be used in the examples presented below. Readers may refer to Béal et al. [2019] for more details.

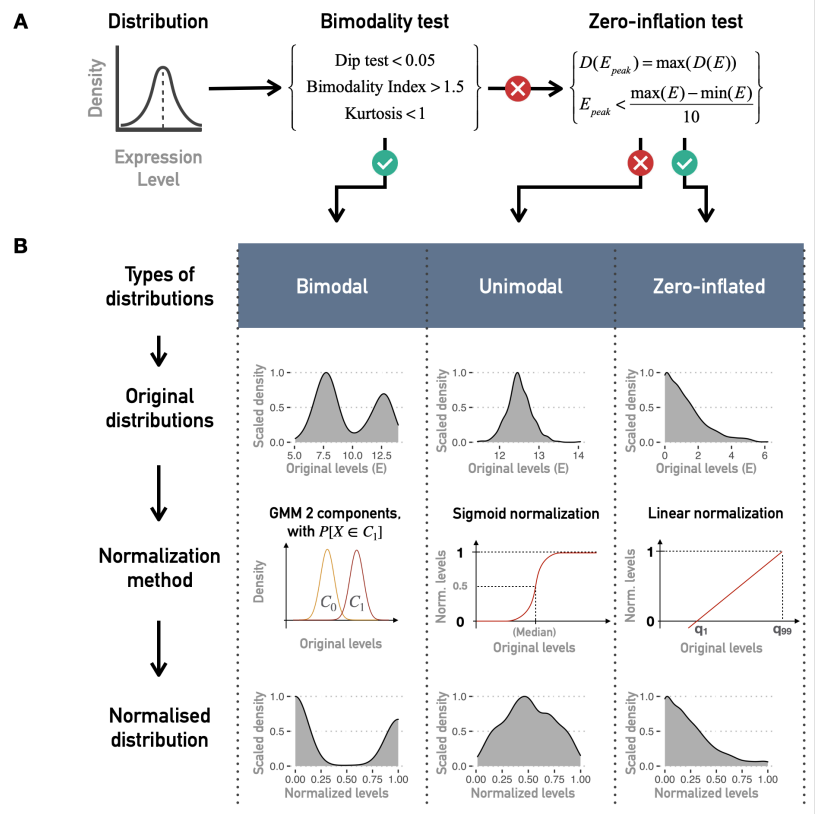



Figure 5.4: **Normalization of continuous data for logical modeling.** (A) Combinations of tests and criteria to classify distributions of continuous data (such as gene expression for one gene and all patients) as bimodal, unimodal or zero-inflated. (B) Normalization methods for each kind of distribution.

### 5.1.3 Personalizing logical models with patient data

It is now possible to redefine more precisely the ways of integrating data into a logical model defined with MaBoSS, as sketched at the beginning of the previous section. **Personalization is defined here as the specification of a logical model with data from a given patient**: each patient has a personalized model tailored to his/her data, so that all personalized models are different specifications of the same logical model, using data from different patients (Figure 5.1). Based on MaBoSS formalism and the processed patient data, there are several possibilities to personalize a generic logical model with patient data. One possibility to have patient-specific models is to force the value of the variables corresponding to the altered

genes in a given patient, i.e., constraining some model nodes to an inactive (0) or active (1) state (Figure 5.5A). In order to constrain a node to 0 (resp. 1), the initial value of the node is set to 0 (resp. 1) and $k_{0\to1}$ (resp. $k_{1\to0}$) to 0 to force the node to maintain its initially defined state. For instance, the effect of a TP53 inactivating mutation can be modeled by setting the node *p53* in the model and its initial condition to 0 and ignoring the logical rule of p53 variable. Because of the type of data used, this personalization method is referred to as **discrete personalization**. It has also been called *strict node variants* in Béal et al. [2019] because this data integration overwrites the logical rules.

Another possible strategy is to modify the initial conditions of the variables of the altered genes according to the results of the normalization (i.e., the probability of initial activation for one node among the thousands of stochastic trajectories). These initial conditions can capture different environmental and genetic conditions. Nevertheless, in the course of the simulation, these variables will be prone to be updated depending on their logical rules. Finally, as MaBoSS uses Gillespie algorithm to explore the STG, data can be mapped to the transition rates of this algorithm. In the simplest case, all transition rates of the model are set to 1, meaning that all possible transitions are equally probable. Alternatively, it is possible to separate the speed of processes by setting the transition rates to different values to account for what is known about the reactions: more probable reactions will have a larger transition rate than less probable reactions [Stoll et al., 2012]. For this, different orders of magnitude for these values can be used. They are set according to the activation status of the node (derived from normalized values) and an amplification factor $F$, designed to generate a higher relative difference in the transition rates, and are therefore defined for each node $i$ and sample $j$:

$$k_{i,j}^{0\to1} = F^{2(X_{i,j}^{norm}-0.5)}$$

$$k_{i,j}^{1\to0} = \frac{1}{k_{i,j}^{0\to1}}$$

Thus, if a gene has a value of 1 based on its RNA profile, $k_{0\to1}$ (resp. $k_{1\to0}$) will be $10^2$ (resp. $10^{-2}$) with an amplification factor of 100. This amplification factor is therefore a hyper-parameter of the method. Very low values of $F$ will have no impact while higher values will make some transitions almost impossible and the method will then approach the discrete personalization described above. Some quantitative illustrations of the influence of $F$ are

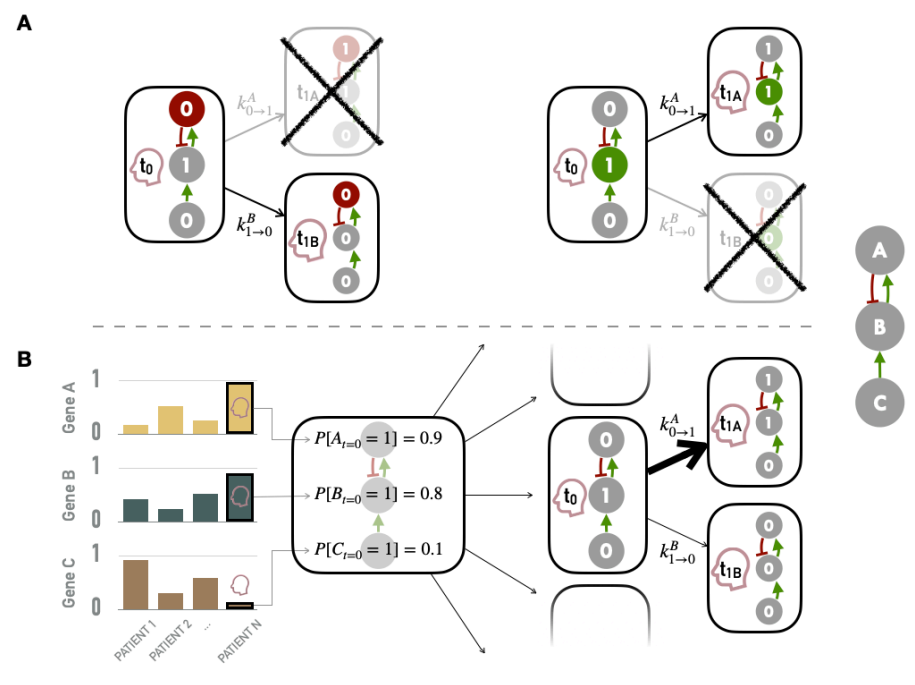



Figure 5.5: **Methods for personalization of logical models.** (A) Personalization with discrete data, such as mutations, with some nodes forced to 0 based on loss of function alteration (left) or 1 based on gain of function/constitutive activation (right). (B) Personalization with continuous data used to define the initial conditions of nodes and to influence the transitions rates and the subsequent probabilities of transition in asynchronous update; the graph on the left represents the normalized values of genes A, B and C for patients 1, 2 and N; the right side represents the personalization of logical model using values from patient N (red profile), first defining the initial probabilities of node activation (middle) and then influencing the probabilities of transitions from one state to another (right).

provided in Béal et al. [2019]. The integration of continuous data through the initial conditions of the nodes and the transition rates are combined to form a second personalization method called **continuous personalization** and described in Figure 5.5B. This method has also been called *soft node variants* to emphasize its difference with discrete/strict personalization: it may influence the trajectories in the solution state space leading to a change in probabilities of the resulting stable state but it does not overwrite the logical rules. To illustrate a little more explicitly the impact of continuous personalization, if a given node has a normalized value of 0.8 after data processing (based on proteins levels for instance), it will be initialized as 1 in 80% of the stochastic trajectories, its transition rate $k_{0\to1}$ will be increased (favoring its activation) and its transition rate $k_{1\to0}$ will be decreased (hampering its inactivation). These changes increase the probability that this node will remain in an activated state close to the one inferred from the patient's data, while maintaining the validity of its logical rule. Thus, continuous personalization appears as a smoother way to shape logical models' simulations based on patient data.

In summary, different types of data can be used, with different integration methods. Note that it is quite natural to use genetic alterations (mutations, CNA) to specify definitive changes in models (such as those of discrete personalization) since this corresponds to biological reality: a mutation cannot be undone or reversed. Conversely, continuous alterations in expression or phosphorylation are subject to modification and regulation, thus justifying their interpretation in a less strong and definitive way (such as continuous personalization). Finally, it follows from these definitions that there are different strategies for personalizing a logical model since discrete and continuous personalizations can each use different types of data; and moreover, these two strategies can be combined. **Except otherwise stated, mutations (resp. RNA or protein) will always be integrated using discrete (resp. continuous) personalization and the joint integration of both types of data will therefore combine both methods.** The relative merits of the different personalization strategies will be discussed below.

## 5.2 An integration tool for high-dimensional data?

Once the method has been defined, it is imperative to study its validity and possible limitations. This comes down to answering the question: **do**

**personalized models capture a biological reality**, and in our case do they discriminate between different types of cancer?

### 5.2.1 Biological relevance in cell lines

These questions can be addressed using cell line data. Using the logical model of cancer pathways from Fumia and Martins [2013], it is possible to study the 663 cell lines from different types of tumors by integrating their processed omics profiles to the generic logical model to obtain as many personalized models. If we focus on the read-out of *Proliferation*, one of the easiest to interpret, there are several ways to study its relevance. For each cell line and each personalization strategy (and corresponding data type) we can define a personalized model and derive the asymptotic value the *Proliferation* node, called *Proliferation* score. This score is therefore *a priori* different for all cell lines that present a different molecular profile. For the whole population of cell lines, this score can be confronted with other markers of proliferation such as the levels of Ki67 [Miller et al., 2018], here replaced as an example by the RNA levels of the corresponding MKI67 gene. It can then be observed that the simulated *Proliferation* indicator, derived from the personalized models, correlates positively with the biomarker, but only when RNA has been used in the personalization (Figure 5.6A). The **correlation makes qualitative sense, but the heterogeneity appears to be very large and most of the variability is not captured by the models**. This heterogeneity is also visible by focusing on some types of cancer (Figure 5.6B). Thus this kind of comparison only validates the models' ability to retrieve a RNA biomarker (not used in personalization) when they themselves integrate other RNA data. It is also consistent that scores from models personalized with mutations only have less uniform distributions due to the discrete nature of the data and the many identical profiles: many cell lines are not distinguishable by mutations only.

It is possible to go one step further by comparing these personalized *Proliferation* scores with the doubling time of the cell lines, i.e., the time it takes for the cell line population to double. A cell line described as proliferative (high *Proliferation* score) should thus have a low doubling time. This can be observed qualitatively by using a subgroup of cell lines for which this information is available (Figure 5.6C). These correlations are not significant and once again summarize a large heterogeneity. Predicting doubling times is, however, a rather difficult task with poor accuracies, even with the help of more flexible machine learning low [Kurilov et al., 2020].

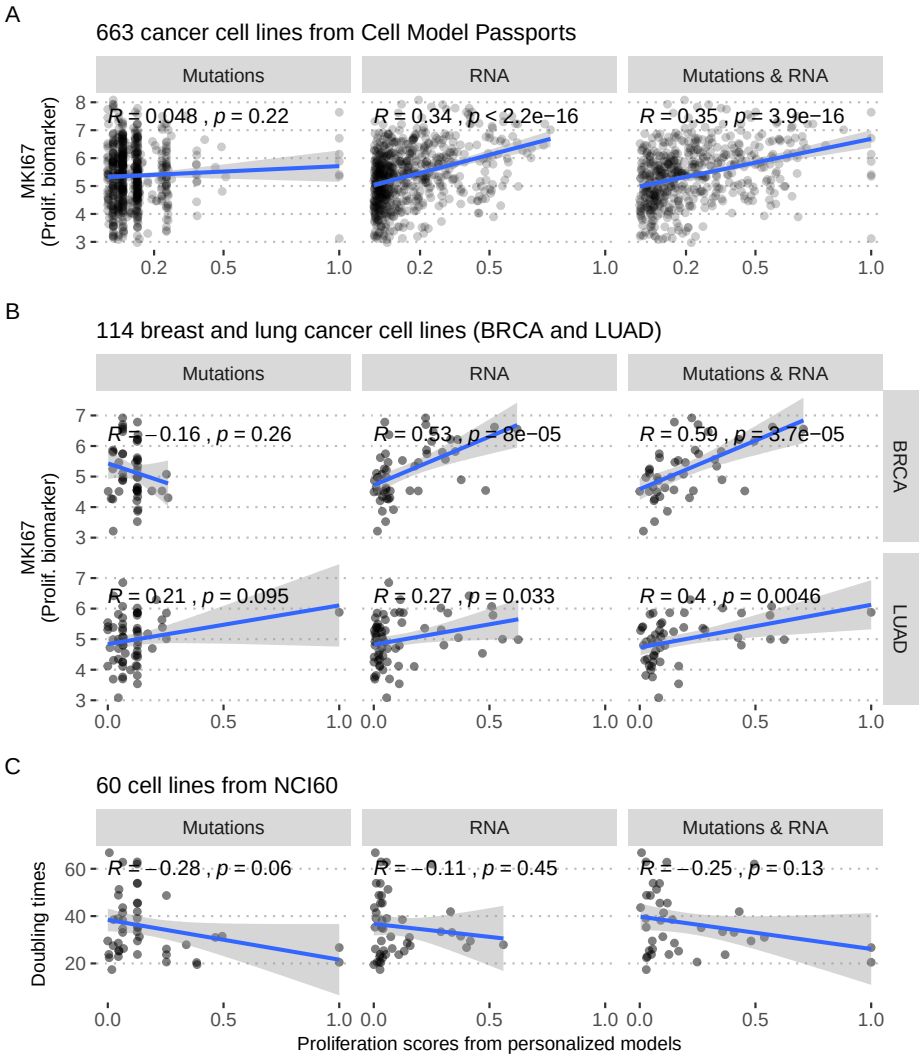



Figure 5.6: **Validation of personalized *Proliferation* scores in cell lines.** (A) Comparison with MKI67 proliferation biomarker for all cancer cell lines. (B) Same with breast (BRCA) and lung (LUAD) cancer only. (C) Comparison with doubling times in a subset of 60 cell lines.

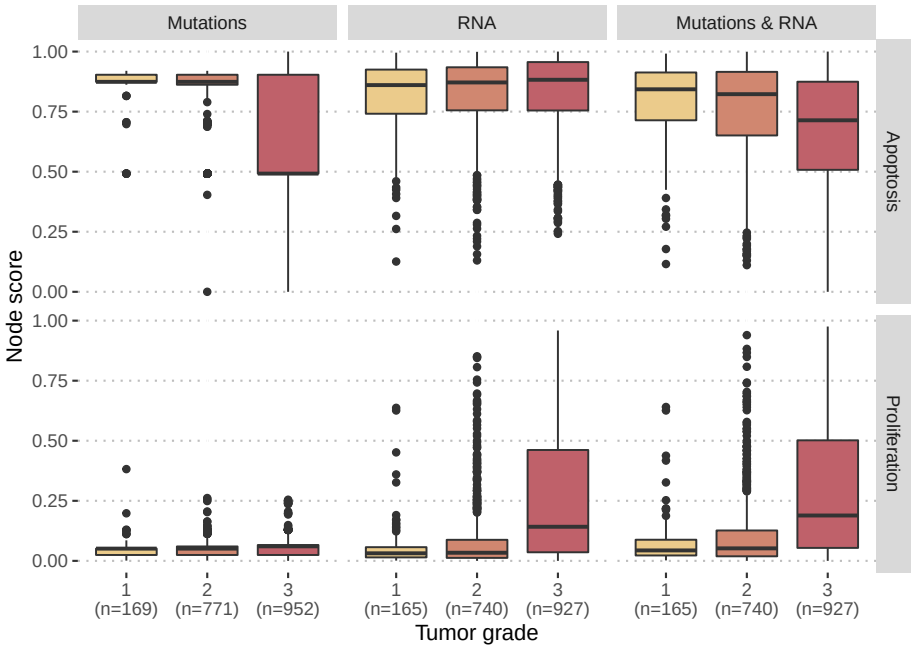



Figure 5.7: **Comparaison of personalized scores with tumor grades for breast cancer patients in METABRIC cohort.** Comparisons are provided for different personalization strategies (with mutations and/or RNA) and two different model nodes (*Proliferation* and *Apoptosis*).

### 5.2.2 Validation with patient data

Patient data can as well be used to reproduce analyses of the same type as those previously performed with the MKI67 biomarker, as was done in Béal et al. [2019], but we focus here on the more clinical applications of the personalized mechanistic models. By analogy with the validations proposed for other mechanistic models [Fey et al., 2015], it is also possible to evaluate the **prognostic value of personalized logical models on patient data**. For example, when studying breast cancer patients in the METABRIC cohort, *Proliferation* and *Apoptosis* scores differ according to tumor grade. The more advanced tumors (grade 3) are associated with higher *Proliferation* scores and lower *Apoptosis* scores (Figure 5.7). This is in line with the natural interpretation that could be given since proliferation is by definition a sign of cancer progression while apoptosis, a programmed death of defective cells, is on the contrary a protective mechanism. While these trends are monotonous and clearly significant for the third strategy using both mutations and RNA ($p < 10^{-12}$ with Jonckheere–Terpstra test for ordered differences among classes, for both nodes), this is not the case

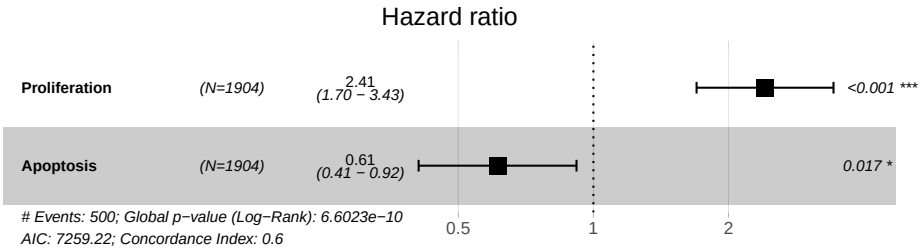



Figure 5.8: **Hazard ratios for *Proliferation* and *Apoptosis* in a survival Cox model in METABRIC cohort.** Higher *Proliferation* (resp. *Apoptosis*) scores correspond to higher (resp. lower) probabilities of death.

when the two types of data are used separately: mutations (resp. RNA) are not sufficient to personalize *Proliferation* (resp. *Apoptosis*) scores in a meaningful way. The personalisation method therefore seems to be able to combine discrete and continuous data in such a way that some of the biological information is preserved.

This comparison to clinical data can be extended to **patient survival data** in the same cohort. If we focus on the strategy integrating both mutations and RNA, we observe that in a Cox model of survival, *Proliferation* is significantly associated with a higher risk of event while *Apoptosis* is associated with a lower risk, which is again consistent (Figure 5.8). In a more schematic and visual way, it is possible to transform these continuous *Proliferation* and *Apoptosis* scores into binary indicators (using medians) and observe their impact on survival, as it has been done in previously mentioned studies [Fey et al., 2015, Salvucci et al., 2019b]. The shortcomings of such approaches will be discussed from a statistical point of view in Part III. We then observe the same behaviour for the two personalized scores (Figure 5.9A and B). Interestingly, if we combine the indicators to create groups that are expected to be of very bad prognosis (high *Proliferation*, low *Apoptosis*) or of very good prognosis (low *Proliferation*, high *Apoptosis*), we further discriminate patients and confirm the qualitatively meaningful interpretation of the personalized scores. It should be noted that the clinical validations presented here remain voluntarily simple and quite close to those proposed in similar articles. Discussions and statistical developments will be proposed in Part III.

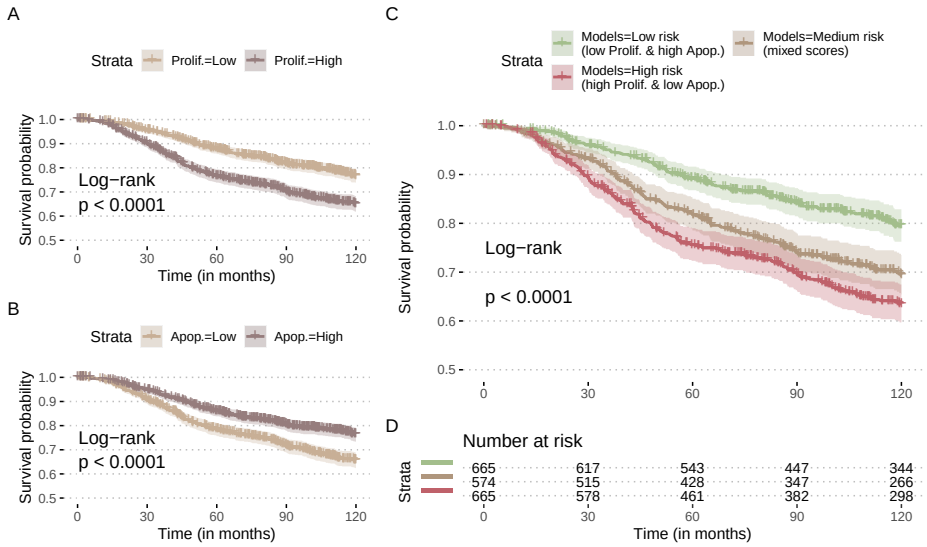



Figure 5.9: **Prognostic value of *Proliferation* scores for breast cancer patients in METABRIC cohort.** (A) Survival curve for overall survival stratified with *Proliferation* scores from personalized models integrating mutations and RNA; scores have been binarized based on median and survival censored at 120 months. (B) Same with *Apoptosis* scores. (C) Survival curve stratified with combinations of *Proliferation* and *Apoptosis* scores, based on the same thresholds, and the corresponding number of patients at risk (D).

### 5.2.3 Perspectives

In summary, this kind of application of personalized models allows the **integration of quite heterogeneous and moderately dimensional biological data in a constrained framework** that orders the relationships between variables and guides interpretations. Comparison with external biological or clinical data then makes it possible to verify the absence of major contradictions in the definition of the model. However, the interest of these mechanistic approaches in this type of task appears as quite moderate compared to statistical models. The qualitative aspect is not necessarily compensated here by the integration of knowledge into the structure of the model, especially in examples that use an extremely broad logical model, which has not been specifically designed for the problems to which it is applied. It is then necessary to study the application of these personalized models to more suitable problems, in which the explicitly mechanistic nature of the models can be exploited.

Chapter 6

# Personalized logical models to study an interpret drug response

*"Il serait excellent que tout médecin ait la possibilité d'expérimenter un grand nombre de médicaments sur lui-même. Sa compréhension de leurs effets en serait tout autre."*

*Mikhail Bulgakov (Morphine, 1927)*

Historically, all mechanistic models of molecular networks, and logical models in particular, have been widely used to study response to treatments [Flobak et al., 2015, Jastrzebski et al., 2018]. Indeed, biological entities, many of which are prospective therapeutic targets, are explicitly represented in the model, making it possible to simulate their inhibition. This is what will be presented in this chapter using the personalized logical models described above. Can they be used to study the response of biological systems to perturbations, in this case the response of cell lines to gene or protein inhibitions? Compared to the numerous statistical models designed to predict the sensitivity of cell lines to treatements, what information do these personalized mechanistic models provide?

**Scientific content**

This chapter extends the method presented in the previous chapter to investigate drug response with personalized logical models. The first application to cell lines of all cancer types was presented orally at ISMB2020 in Basel but is not published.
The example about BRAF in melanomas and colorectal cancers is under review and the corresponding pre-print is available as Béal et al. [2020]. In this joint work, only the construction of the generic logical model and the model-checking procedure were mostly carried out by collaborators and especially by an intern under my joint supervision. These two steps will therefore be described more succinctly.
Finally, the work on prostate cancer presented in a third section will be submitted soon. It is also a joint work, in which my participation focused on the application of the PROFILE method.

## 6.1 One step further with drugs

One of the main clinical consequences of the underlying molecular complexity of cancers is the divergent response to treatment, even for *a priori* similar tumors. In the light of high-throughput sequencing data, the mechanisms governing these responses are somewhat better understood, for patients and especially for model organisms such as cell lines [Heiser et al., 2012, Garnett et al., 2012]. But beyond a few simple cases, the diversity of response biomarkers once again calls for **holistic approaches** to unravel the underlying mechanisms.

### 6.1.1 Modeling response to cancer treatments

To study these observed differences in drugs response in various cancers, some approaches based on mathematical modeling were developed to explore the complexity of differential drug sensitivities[1]. A number of **machine learning-based methods for predicting sensitivities** have been proposed [Costello et al., 2014], either without particular constraints or with varying degrees of prior knowledge; but they do not necessarily pro-

[1]Sensitivity is understood throughout this chapter in the biological sense, *i.e.*, the response of a biological system (here cell lines) to an external disturbance (*e.g.*, a drug). This definition is extended to personalized logical models whose response to the same perturbations is studied.

vide a mechanistic understanding of the response. Some other approaches focused on the description of the processes that might influence the response by integrating knowledge of the signaling pathways and their mechanisms and translated it into a mathematical model [Eduati et al., 2017, Jastrzebski et al., 2018, Fröhlich et al., 2018]. The first step of this approach implies the construction of a network recapitulating knowledge of the interactions between selected biological entities (driver genes but also key genes of signaling pathways), extracted from the literature or from public pathway databases, or directly inferred from data [Verny et al., 2017]. This static representation of the mechanisms is then translated into a dynamical mathematical model with the goal to not only understand the observed differences [Jastrzebski et al., 2018] but also to predict means to revert unexpected behaviours.

One way to address issues related to patient response to treatments is to **fit these mechanistic models to the available data**, and to train them on high-throughput cell-line specific perturbation data[2] [Eduati et al., 2017, Jastrzebski et al., 2018, Klinger et al., 2013]. These mechanistic models are then easier to interpret with regard to the main drivers of drug response. They also enable the *in silico* simulations of new designs such as combinations of drugs not present in the initial training data [Fröhlich et al., 2018]. However, these mechanistic models contain many parameters that need to be fitted or extracted from the literature. Some parsimonious mathematical formalisms have been developed to make up for the absence of either rich perturbation data to train the models or fully quantified kinetic or molecular data to derive the parameters directly from literature. One of these approaches is the logical modeling, which uses discrete variables governed by logical rules. Its explicit syntax facilitates the interpretation of mechanisms and drug response [Zañudo et al., 2017, Iorio et al., 2016] and despite its simplicity, semi-quantitative analyses have already been performed on complex systems including drug response studies [Knijnenburg et al., 2016, Eduati et al., 2020].

### 6.1.2 An application of personalized logical models

But logical formalism has also shown its relevance regarding drug response in cases where the model is not automatically trained on data but simply constructed from literature or pathway databases and where biological ex-

[2]In this thesis, this term refers to data from biological systems (e.g. cell lines) that have been disrupted according to different technologies or molecules: drugs, CRISPR/Cas9 etc. The dynamic response of the studied system to these perturbations is thus accessed instead of being restricted to a static knowledge of the system.

periments focus on a particular cell line [Flobak et al., 2015]. The study is then restricted to one cell line only from which some data and parameters have been experimentally inferred. Using the PROFILE method, it is possible to generate personalized logical models associated with different cell lines and then use them to study the response to treatment. **Since the models are not trained with perturbation data but simply specified/constrained by interpreting the molecular profiles, it is possible to personalize the logical models with a rather limited amount of data**.

The principles are summarized in Figure 6.1. A generic model (Figure 6.1A) is first transformed into as many personalized models as there are cell lines with an omics profile. These persnalized models are then simulated by adding the effect of a given treatment (Figure 6.1B). The treatments that can be studied are generally targeted inhibitors. Generally speaking, one must be able to **translate the mechanism of action of the treatment into the logical model**. The impact of more systemic treatments such as chemotherapy or radiotherapy is more difficult to study with these methods, in any case with most of the logical models published to date, even if in theory, precise modeling of the pathways associated with these treatments (such as DNA repair) could contribute to this.

It is then possible to analyze the personalized scores for each cell line (asymptotic values of the phenotypic read-outs of the model) with or without the effect of treatment. If the model includes more than two phenotypes of interest, such as the one in Figure 6.1A, one can visualize these behaviors in the PCA space of the personalized scores, as shown schematically in Figure 6.1C. In this case the directions of the original phenotype features (*Proliferation*, *Apoptosis*, *Quiescence*) have been added in the PCA-transformed space in order to facilitate the interpretation of positions and drug-induced displacements. In this mock example, based on personalized models, treatment would promote a shift from proliferative to more apoptotic or quiescent behaviors, in particular in the red and green cell lines, which are *a priori* more sensitive to the treatment.

### 6.1.3 A pan-cancer attempt

This versatile analysis framework was first applied during this thesis to a **large pan-drug and pan-cancer analysis**. On the basis of generic logical models such as those previously presented (see appendix B.1 and B.2), and in view of the abundance of available data (across cancer tissues

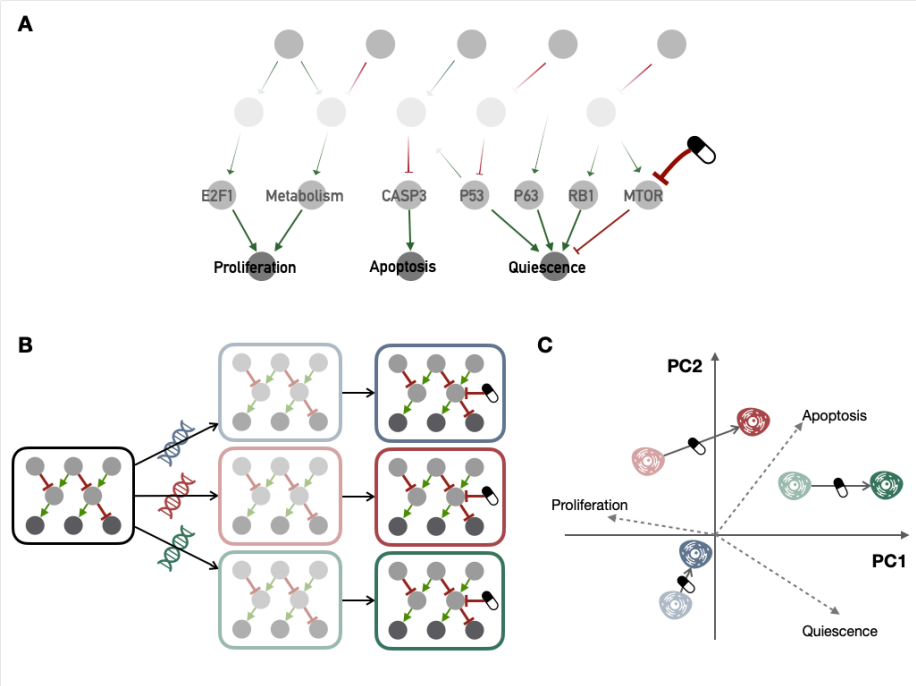



Figure 6.1: **Schematic extension of PROFILE-personalized logical models to drug investigation.** (A) Schematic representation of a logical model of cancer molecular networks, in particular the one described in appendix B.2 and used in the next subsection. (B) Sequential pipeline for drug response investigation with PROFILE, starting from a generic logical model, then transformed into several personalized models with different molecular profiles (correspondong to several cell lines); these models are finally simulated with a defined drug inhibition. (C) A possible analysis of the predictions of personalized models obtained from the generic model described in (A); a PCA is computed based on the final phenotype scores from personalized model, it allows to interpret biologically how the models represent cell lines (*e.g.*, more or less proliferative) and especially what impact of treatment they predict (*e.g.*, decrease *Proliferation* or increase *Apoptosis*).

and drugs such as in GDSC cell line dataset, see appendix B.1), there were no theoretical obstacle to such an analysis. Although the simulations were carried out without any problems, the analysis nevertheless proved extremely difficult to interpret. We will highlight the various problems encountered, propose an illustration and some perspectives that led to the work presented in the following section.

Based on the PROFILE methodology and GDSC data, hundreds of personalized models can be obtained, each corresponding to a cell line. For each of these personalized models, several dozen of potential drugs have a mechanism of action that can be mechanistically translated into the logical model. We thus obtain **tens of thousands of "personalized model/drug" pairs that correspond to experimentally evaluated drug sensitivies** (cf appendix A.1.2 for details). Firstly, the comparison of simulated and experimental data is not straightforward. As the models are qualitative, it is necessary to carry out the validation in this spirit. The idea is not to predict sensitivity quantitatively, rather to verify their relative relevance. In the first place, do we recover the cell lines that are most sensitive to a given drug? With several hundred cell lines, it is difficult to make this reflection graphically as in Figure 6.1C. More quantitative approaches, such as correlation, would require the **definition of a precise sensitivity proxy in personalized models**. Should we choose the personalised *Proliferation* score obtained with drug? Or the drug-induced displacement in the mechanistic model (the drug arrows in Figure 6.1C)? Or is a combination of phenotypes used, if so which one? As for experimental metrics, which ones to choose, and what interpretations do they allow? Whatever the choice, dose-response AUC or IC50 (see details in the appendix A.1.2), a problem arises: can the sensitivities of a cell line to different drugs be compared? Such a comparison would allow the most clinically interesting questions of precision medicine to be asked: for a given molecular profile, can the model predict the best treatment to administer? However, AUCs are comparable for different drugs only if the concentration ranges tested are similar; and IC50s are extrapolated, sometimes well beyond the concentrations tested. Qualitative comparisons for a given drug therefore seem the most meaningful, as long as a relevant proxy in personalized models can be justified.

Aware of these difficulties, if one decides to do a correlation analysis, for each drug, of the personalized correlation scores with experimental sensitivities, one realizes that **some experimental responses correlate well with the behaviours of personalized models while others do not.**

But it is difficult to decide between two different interpretations: does this mean that correctly predicted drugs are well modeled and others are not? Or does it mean that some correlations appear to be better by chance because so many drugs have been modeled? A case study can be illustrated more precisely with the example shown in Figure 6.2. In order to simplify the analysis presented schematically in Figure 6.1C, the 663 cell lines were averaged by cancer type (according to TCGA denominations) and the drug-induced shifts are all represented from the origin in the PCA space. There is evidence that the effect of the drug on personalized models (using only mutations) tends to make them less proliferative and more apoptotic/quiescent (Figure 6.2A). This shift is strongest for those types of cancer that are actually most sensitive to this inhibitor experimentally (*i.e.*, low AUC), such as skin cutaneous melanomas (SKCM) in particular, and colorectal (COAD/READ) or pancreatic (PAAD) cancers to a lesser extent. The ability of personalized models to explain this difference can be understood by a known underlying biological reality: the prevalence of BRAF or RAS mutations in these cancers. The three aforementioned cancers are thus very frequently mutated for one of the two genes (Figure 6.2B). Then, the model translates the fact that these two genes are located just upstream of MAP2K1. It is therefore natural that an inhibition just downstream of these important mutations is particularly effective (Figure 6.2C). In a case such as this, the relevance of the model can be explained and justified *a posteriori*. This analysis is much more difficult in the vast majority of cases, whether the correlations are apparently good or not.

This example highlights a problem of scope. **The fact that the method enables to study hundreds of cell lines and dozens of drugs does not mean that it is relevant in each case**. The description of pathways in the model is more or less accurate. For example, a node at the model boundaries probably has many regulators missing. Is it then relevant to investigate the response of personalized models to its inhibition? It is therefore necessary to restrict the drugs studied. Similarly, even if the logical model summarizes many important pathways, it is probably unsuitable for certain cell lines or certain types of cancer with different etiologies. However, it is difficult to restrict the scope of the analysis in an unbiased way without having designed a model *de novo* for a specific purpose.

For all these reasons, it was decided to leave aside this naive, broad-spectrum approach in favour of starting from a more specific biological question and constructing the appropriate logical model.

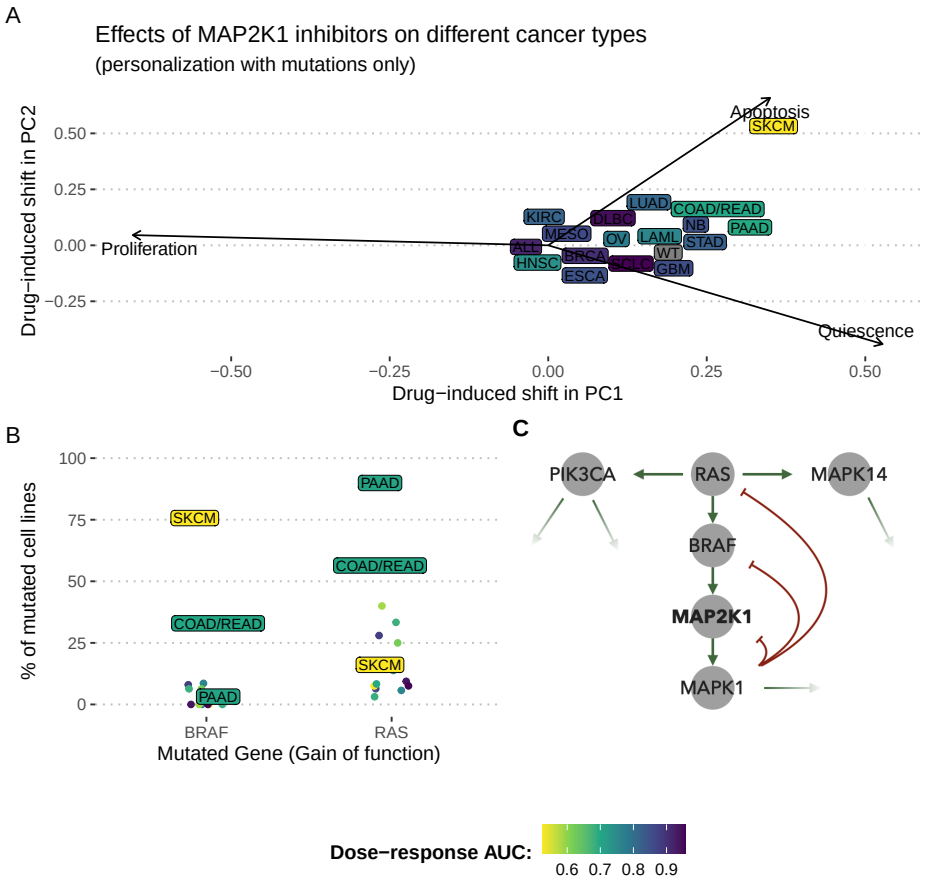



Figure 6.2: **PROFILE-generated models and sensitivites to MAP2K1 inhibitors averaged per cancer type.** (A) Effects of MAP2K1 inhibitors on personalized logical models averaged per cancer types and represented in a normalized PCA space with super-imposed original phenotypes. (B) Proportion of BRAF- and RAS-mutated cell lines in some cancer types. (C) Zoom on the MAPK pathway of the logical model used.

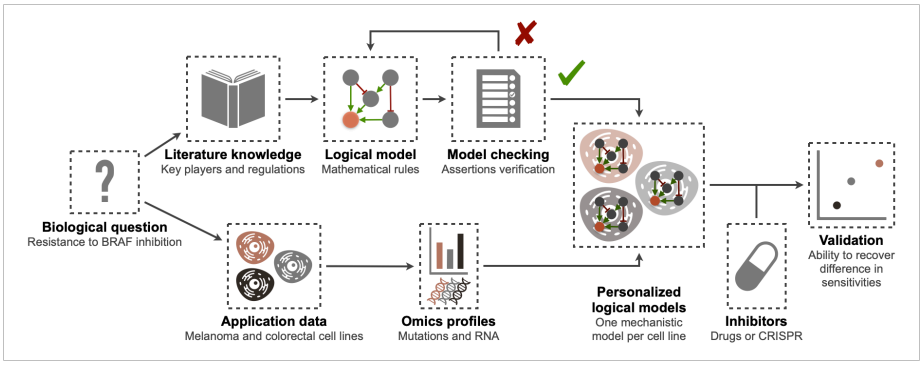



Figure 6.3: **BRAF modeling flowchart: from a biological question to validated personalized logical models.** Logical models are written with MaBoSS, and the checking model procedure is therefore provided in the same formalism. Cell line data are taken from Cell Model Passports [van der Meer et al., 2019].

## 6.2 Case study on BRAF in melanoma and colorectal cancers

In order to address the limitations outlined in this exploratory analysis, we propose here a pipeline based on logical modeling and able to go from the formulation of a specific biological question to the validation of a mathematical model on pre-clinical data, in this case a set of cell lines, and the subsequent interpretation of potential resistance mechanisms[3] (Figure 6.3). As before, **one of the main points of differentiation with existing mechanistic approaches, is that this framework does not rely on any training of parameters but only on the automatic integration and interpretation of molecular features**.

### 6.2.1 Biological and clinical context

The construction of a mathematical model must be based first and foremost on a precise and specific biological question, at the origin of the design of the model. Here, we choose to explore the different responses to treatments in tumors from diverse cancers that bear the same mutation. A well-studied example of these variations is the BRAF mutation

[3]For the sake of completeness, all the steps will be described below or in appendix; the *"Logical model"* and *"Model checking"* steps that I supervised jointly without implementing them directly will be described more succinctly.

and especially its V600E substitution. BRAF is mutated in 40 to 70% of melanoma tumors and in 10% of colorectal tumors, each time composed almost entirely of V600E mutations [Cantwell-Dorris et al., 2011]. In spite of these similarities, **BRAF inhibition treatments have experienced opposite results with improved survival in patients with melanoma [Chapman et al., 2011] and significant resistance in colorectal cancers [Kopetz et al., 2010]**, suggesting drastic mechanistic differences. Some subsequent studies have proposed context-based molecular explanations, often highlighting surrounding genes or signalling pathways, such as a feedback activation of EGFR [Prahallad et al., 2012] or other mechanisms [Poulikakos et al., 2011, Sun et al., 2014]. These various findings support the need for an integrative mechanistic model able to formalize and explain more systematically the differences in drug sensitivities depending on the molecular context. The purpose of the study we propose here is not to provide a comprehensive molecular description of the response but to verify that the existence and functionality of the suggested feedback loops around the signalling pathway in which BRAF is involved [Prahallad et al., 2012] may be a first hint towards these differences. For a more thorough study of these cancers, we refer to other works [Eduati et al., 2017, Baur et al., 2020, Cho et al., 2016].

### 6.2.2 A logical model centred on BRAF

A logical model summarizing the main molecular interactions at work in colorectal cancers and melanomas is thus built from the literature and completed with databases. As previously mentioned, the objective is to understand whether it is possible to model and explain differences in responses to BRAF inhibition in melanoma and colorectal cancer patients using the same regulatory network. **The fact that the two cancers share the same network but differ from the alterations and expression of their genes constitutes our prior hypothesis**. The focus of this model is put on two important signaling pathways involved in the mechanisms of resistance to BRAF inhibition which are the ERK1/2 MAPK and PI3K/AKT pathways [Ursem et al., 2018, Rossi et al., 2019]. The generic network presented in Figure 6.4 recapitulates the known interactions between the biological entities of the network that was first built from the literature, and then verified and completed with potential missing connections using SIGNOR database [Perfetto et al., 2016]. More details and references about the model can be found in appendix B.3. All in all, the logical model formalizes the knowledge compiled from different sources and highlights the role of SOX10, FOXD3, CRAF, PI3K, PTEN and of EGFR in resistance

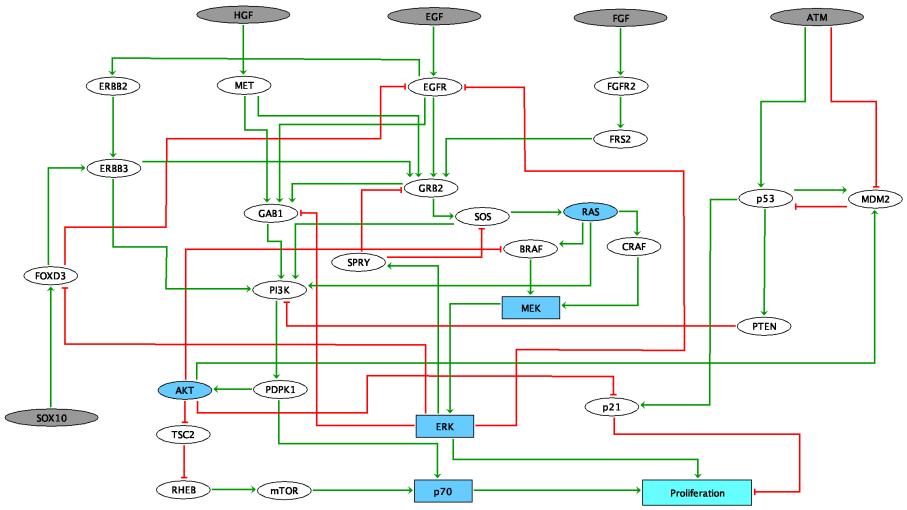



Figure 6.4: **Logical model of signaling pathways around BRAF in colorectal andmelanoma cancers.** Grey nodes represent input nodes, which may correspond to the environmental conditions. Square nodes represent multi-valued variable (MEK, ERK, p70 and Proliferation). Dark blue nodes accounts for families (several genes/entities for one node). Light blue node represents the phenotypic read-out of the model, *i.e.*, *Proliferation*.

to anti-BRAF treatments.

Once the structure of the model was defined, and before moving on to its personalization, its consistency with the literature was checked using a **model-checking procedure**. Indeed, due to the complexity of the system, properly taking into account the interactions between entities does not automatically guarantee that the model will reproduce certain dynamic behaviours. It is therefore a question of verifying whether the model reproduces certain biological assertions found in the scientific literature. An example of a biological assertion may be the reactivation of the MAPK (mitogen-activated protein kinase) pathway through EGFR signal after BRAF inhibition in colorectal cancer [Prahallad et al., 2012]: it is possible to check whether a simulation of this situation with the model gives the same result or not. Because there are many such assertions and because it is useful to verify them as the model is built, automatic model-checking tools have been defined, based on the MaBoSS syntax and inspired by the Python *unittest* library. More details are provided in Béal et al. [2020] and in a corresponding GitHub repository. The list of biological assertions used to validate the model is detailed in the appendix B.3.

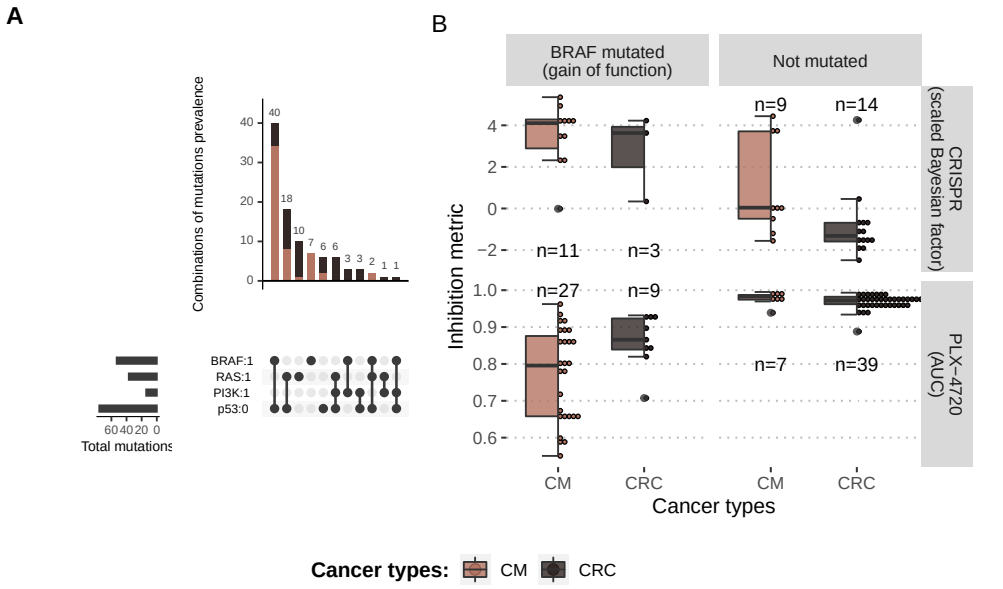



Figure 6.5: **Descriptive analysis of cell lines for melanomas and colorectal cancers.** (A) Number of cell lines for the four most frequently mutated genes and their combinations (plot from UpSetR package [Conway et al., 2017]). (B) Differential sensitivities to BRAF inhibition by the drug PLX-4720 (lower panel) or by CRISPR inhibition (upper panel), depending on BRAF mutational status and cancer type. Numbers of cell lines in each category are indicated. Note that high sensitivities correspond to low AUC and high scaled Bayesian factors.

### 6.2.3 Cell lines data

The omics profiles of colorectal and melanoma cell lines are downloaded from Cell Model Passports portal [van der Meer et al., 2019]. 64 colorectal cancer (CRC) cell lines and 65 cutaneaous melanoma (CM) cell lines are listed in the database, with at least mutation or RNA-seq data (59 CM and 53 CRC with both mutations and RNA-seq data). These omics profiles are used to generate cell-line-specific logical models as described in PROFILE method (Figure 5.1). The prevalence of mutations and their combination for the two types of cancer can be seen in Figure 6.5A and is consistent with the clinical situation described above with melanomas more frequently BRAF-mutated and colorectal cancers more frequently RAS-mutated.

In order to validate the relevance of personalized models to explain differential sensitivities to drugs, some experimental screening datasets are used. **Drug screening data** are downloaded from the Genomics of Drug

Sensitivity in Cancer (GDSC) dataset [Yang et al., 2012] which includes two BRAF inhibitors: PLX-4720 and Dabrafenib. The cell lines are treated with increasing concentration of drugs and the viability of the cell line relative to untreated control is measured. The dose-response relative viability curve is fitted and then used to compute the **area under the dose-response curve (AUC)** [Vis et al., 2016]. AUC is a value between 0 and 1: values close to 1 mean that the relative viability has not been decreased, and lower values correspond to increased experimental sensitivity to inhibitions (details in appendix A.1.2). The results obtained with the two drugs are very strongly correlated (Pearson correlation of 0.91) and the analyses presented here will therefore focus on only one of them, PLX-4720.

In a complementary way, some results of **CRISPR/Cas9 screening** are also downloaded from Cell Model Passports. This technology, which is described in more detail in the appendix A.1.3, allows targeted inhibitions of certain genes. Two different datasets from Sanger Institute [Behan et al., 2019] and Broad Institute [Meyers et al., 2017] are available. We use **scaled Bayesian factors** to assess the effect of CRISPR targeting of genes. These scores are computed based on the fold change distribution of single guide RNAs [Hart and Moffat, 2016]. The highest values indicate that the targeted gene is essential to the cell fitness. The agreement between the two databases is good [Dempster et al., 2019] but we choose to focus on the Broad database, which is more balanced in terms of the relative proportions of melanomas and colorectal cancers.

Figure 6.5B illustrates both the relative quantities of cell lines for which drug or CRISPR screening data are available (depending on their BRAF status) as well as differences in experimental sensitivity to BRAF inhibition. The greater sensitivity of BRAF-mutated melanomas compared to BRAF-mutated colorectal cancers is well observed for PLX-4720. However, the overlap in the distributions requires a deeper look into the data and a search for more precise explanations of the differences in sensitivity, including within each type of cancer. The finding appears to be similar for CRISPR despite a sample size that is too small; the higher average sensitivity of melanomas even extends to non-mutated BRAF.

### 6.2.4 Validation of personalized models using CRISPR/Cas9 and drug screening

The validation of personalized logical models using these screening data is done with the following rationale. First, the models are personalized

using omics data from the cell lines. Then, two separate simulations are performed for each personalized model: one without the inhibition, the other by creating and activating a BRAF inhibitor to mimic the drug or CRISPR inhibition. The ratio of the *Proliferation* phenotype obtained with inhibition and without inhibition is the proxy used to be compared with the different screening metrics each of which is also standardized (AUC calculated on relative viability for drugs and Bayes factor computed from fold-changes and then scaled).

#### 6.2.4.1 Differential sensitivities to BRAF targeting explained by personalized logical models

Once the logical model consistency has been validated, personalized models are generated for each cell line by integrating their interpreted genomic features directly as model constraints or parameters. **Sensitivities to BRAF inhibition inferred from models are then compared to experimentally observed sensitivities** (Figure 6.6). In all the following analyses, we focus on three different personalization strategies using: only mutations as discrete personalization (Figure 6.6A, upper row), only RNA as continuous personalization (Figure 6.6A, middle row) or mutations combined with RNA (Figure 6.6A, lower row). These choices reflect first of all the following *a priori*: mutations are much more drastic and permanent changes than RNA, whose expression levels are more subject to fluctuation and regulation. The objective is also to answer the following questions: What type of data is most likely to explain the differences in responses? Is it relevant to combine them? Figure 6.6 shows an example of the type of analyses possible with personalized models, zooming in more and more on the details from panel A to panel C.

The first approach consists in using only mutations as discrete personalization (Figure 6.6, A, upper row): the mutations identified in the dataset and that are present in the regulatory network are set to 1 for activating mutations and set to 0 for inactivating mutations. In this case, the *Proliferation* scores from personalized models significantly correlate with both BRAF drug inhibitors (PLX-4720 and Dabrafenib) and both CRISPR datasets (using Pearson correlations). Note that the opposite directions of the correlations for the drug and CRISPR datasets are due to the fact that cell lines sensitive to BRAF inhibition result in low AUCs, and high scaled Bayesian factors, respectively, and, if the models are relevant, to low standardized *Proliferation* scores. Looking more closely at the corresponding scatter plot for PLX-4720 (Figure 6.6B, upper left), it can be seen

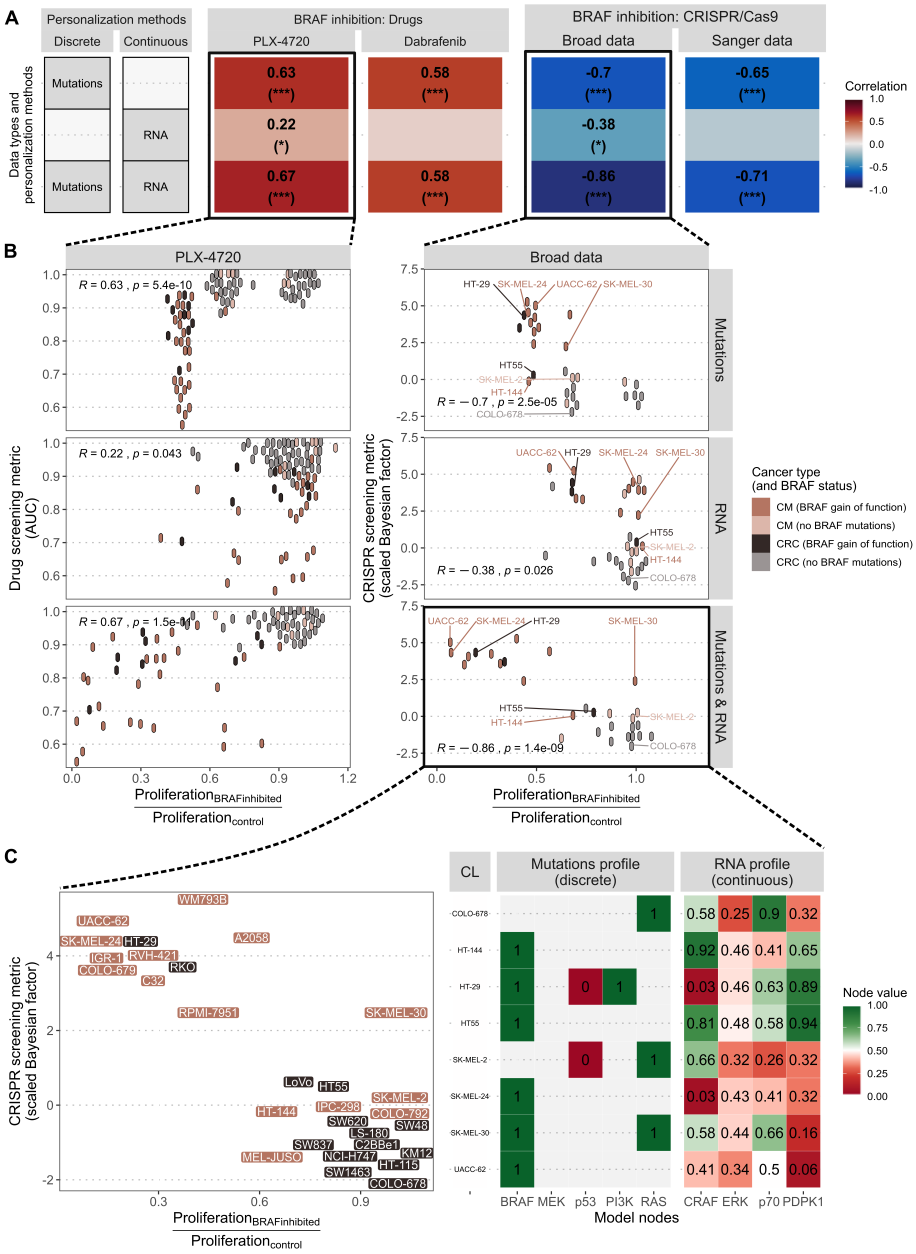



Figure 6.6: **Validation of personalized models of BRAF inhibition with cell lines data.** (A) Pearson correlations between normalized *Proliferation* scores from models and experimental sensitivities to BRAF inhibition (drug or CRISPR); only significant correlations are displayed. (B) Scatter plots with non-overlapping points corresponding to correlations of panel A for one drug (PLX-4720) and one CRISPR dataset (Broad) only. (C) Enlargement of one scatter plot in B (left) with the table describing the omics profiles used for each cell line to explore the response mechanisms (right); interactive version in Figure 6.7 or GitHub files.

that this correlation results from the **model's ability to use mutations' information to recover the highest experimental sensitivities of the BRAF-mutated cell lines** that form an undifferentiated cluster on the left side. These cell lines are indeed relatively more sensitive than non-mutated BRAF cell lines. However, the integration of mutations alone does not explain the significant differences within this subgroup (AUC between 0.55 and 0.9). A very similar behaviour can be observed when comparing model simulations with CRISPR data (Figure 6.6B, upper right).

Using only RNA data as continuous personalization (Figure 6.6A and B, middle rows) is both less informative and more difficult to interpret. For continuous data such as RNA-sequencing data, we normalize the expression values and set both the initial conditions and the transition rates of the model variables to the corresponding values. Correlations with experimental BRAF inhibitions appear weaker and more uncertain. The key point, however, is that the **combination of mutations and RNA, as depicted in Figure 6.6 A and B lower rows, seems to be more relevant**. This is partially true in quantitative terms but it is even easier to interpret in the corresponding scatter plots (Figure 6.7). Comparing first the Broad CRISPR scatter plots using mutations only (Figure 6.6B, upper right) and using both mutations and RNA (Figure 6.6B, lower right), we can observe that non-responsive cell lines (scaled Bayesian factor below 0), grouped in the lower right corner and correctly predicted using only mutations stayed in the same area: these strong mutational phenotypes have not been displaced by the addition of RNA data. Other cell lines previously considered to be of intermediate sensitivity by the model (*e.g.*, COLO-678 or SK-MEL-2) were shifted to the right, consistent with the lack of sensitivity observed experimentally. Finally, BRAF-mutated cell lines, previously clustered in one single column on the left using only mutations (with normalized *Proliferation* scores around 0.5), have been moved in different directions. Many of the most sensitive cell lines (scaled Bayesian factor above 2) have been pushed to the left in accordance with the high sensitivities observed experimentally (*e.g.*, HT-29 or SK-MEL-24). It is even observed that the model corrected the position of the two BRAF mutated cell lines, but whose sensitivity is experimentally low (melanoma cell line HT-144 and colorectal cell line HT-55). Only one cell line (SK-MEL-30) has seen its positioning evolve counter-intuitively as a result of the addition of RNA in the personalization strategy: relatively sensitive to the inhibition of BRAF, it has, however, seen its standardised *Proliferation* score approach 1. All in all, this contribution of RNA data results in significant correlations even when restricted to BRAF-mutated cell lines only ($R = 0.69$, $p.value = 0.006$).

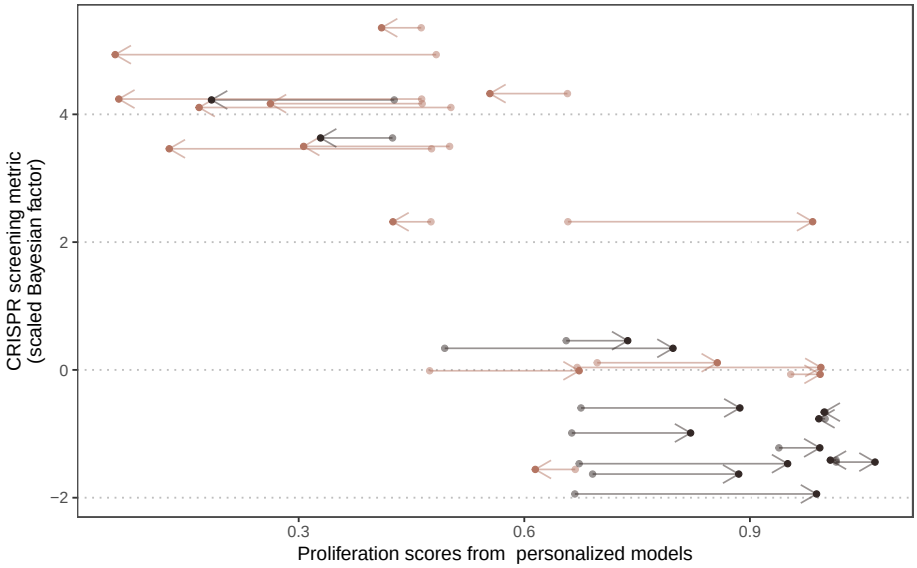



Figure 6.7: **Multi-omics integration and enhanced value of RNA in addition to mutations.** For each cell line, an arrow shows the impact of adding RNA in the customization strategy. This graph is present in an interactive format in the online version of the thesis in order to give easy access to the omic profile corresponding to each point.

A similar analysis can be made of the impact of adding RNA data to personalization when comparing with the experimental response to PLX-4720 (Figure 6.6B, upper and lower left). Most of the non-sensitive cell lines (upper right corner) have not seen the behaviour of the personalized models change with RNA addition. However, the numerous BRAF-mutated cell lines previously grouped around standardized *Proliferation* scores of 0.5, are now better differentiated and their sensitivity predicted by personalized models has generally been revised towards lower scores (*i.e.*, higher sensitivity). Similar to the CRISPR data analysis, three sensitive cell lines have been shifted to the right and are misinterpreted by the model. As a result, the correlation restricted to BRAF-mutated cell lines is no longer significant (R=0.26, p.value=0.1).

#### 6.2.4.2 An investigative tool

These **personalized models are not primarily intended to be predictive tools but rather used to reason and explore the possible mechanisms and specificities of each cell line, for example by studying the molecular alterations at the origin of the observed behaviour**

(Figure 6.7). To continue on the previous examples, the two melanoma cell lines, HT-144 and SK-MEL-24, share the same mutational profiles but have very different sensitivities to BRAF targeting (Figure 6.6C). This inconsistency is partially corrected by the addition of the RNA data, which allows the model to take into account the difference in CRAF expression between the two cell lines. In fact, CRAF is a crucial node for the network since it is necessary for the reactivation of the MAPK pathway after BRAF inhibition. Therefore, the high sensitivity of SK-MEL-24 may be explained by its low CRAF expression level, which makes the reactivation of the MAPK pathway more difficult for this cell line. Conversely, in HT-144, the high level of CRAF expression allows the signal to flow properly through this pathway even after BRAF inhibition, thus making this cell line more resistant. The importance of CRAF expression is also evident in HT-29, a CRC BRAF mutated cell line with other important mutations (PI3K activation and p53 deletion). However, it remains sensitive to treatment, due to its very low level of CRAF expression.

Another interesting contribution of RNA appears in the melanoma cell line UACC-62, which is particularly sensitive to treatment. The model is able to correctly predict its response once RNA levels are integrated. In this case, the reason for sensitivity seems to be due to the low level of PDPK1, which makes it difficult to activate p70 and thus to trigger the resistance linked to PI3K/AKT pathway activation. Similarly, the CRC resistant cell line, HT55, which carries only the BRAF mutation, expresses high levels of PDPK1, in addition to high levels of CRAF, supporting the idea that the presence of both MAPK and PI3K/AKT pathways may confer resistance to BRAF inhibition treatments. We can also mention a cluster of RAS mutated cell lines, usually NRAS mutated for melanomas (*e.g.*, SK-MEL-2) and KRAS for colorectal cancers (*e.g.*, COLO-678), which are classified by the model as resistant. Interestingly, in these cell lines, a low level of CRAF is not enough to block the signal of the MAPK pathway, which is stronger in the model because of the simulation of the RAS mutation (RAS is set to 1). Only SK-MEL-30 appears to be incorrectly classified and is observed to be more sensitive than the other cell lines with a similar mutation profile. This could be due to the fact that our network is incomplete and not able to account for some alterations responsible for this cell line sensitivity. The problem may also come from the fact that this cell line contains a frameshift mutation of RPS6KB2 (p70 node) not referenced in OncoKB and therefore not included in the simulation.

The versatility of the logical formalism makes it possible to test other

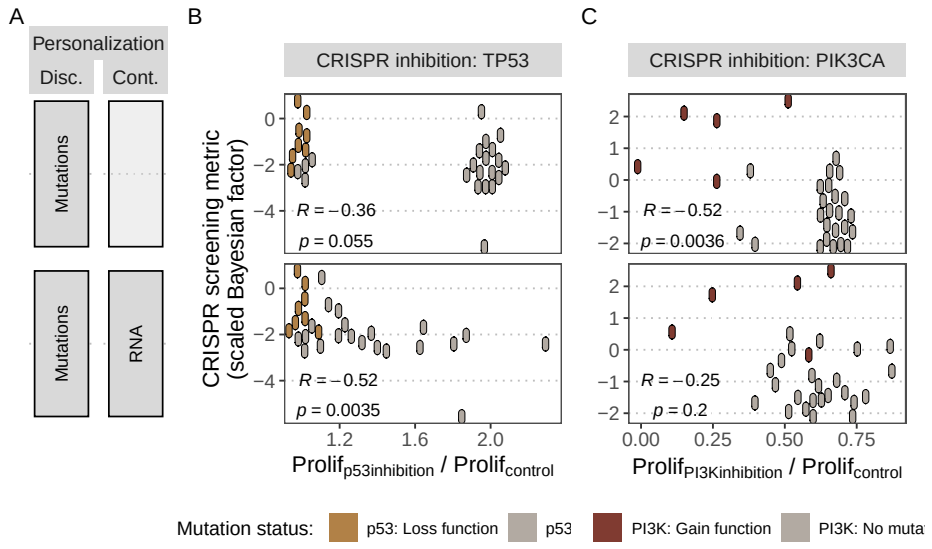



Figure 6.8: **Application of personalized models to other CRISPR targets.** (A) Personalization strategies using either mutations only (as discrete data) or combined with RNA (as continuous data) with their corresponding scatter plots in panels B and C. (B) Scatter plot comparing normalized *Proliferation* scores of p53 inhibition in the models with experiment sensitivity of cell lines to TP53 CRISPR inhibition, indicating p53 mutational status as interpreted in the model. Pearson correlations and the corresponding p-values are shown. (C) Similar analysis as in panel B with PI3K model node and PIK3CA CRISPR inhibition.

node inhibitions as in Figure 6.8, but remains limited by the scope of the model. Since the present model has been designed around BRAF, its regulators have been carefully selected and implemented, which is not necessarily the case for other nodes of the model. Therefore, these personalized models can be used to study how comprehensive the descriptions of the regulation of other nodes or parts of the model are. Thus, model simulations show that response trends to TP53 inhibition are consistently recovered by the model (Figure 6.8B) but the simple regulation of p53 in the model results in coarse-grained patterns, although slightly improved by addition of RNA data. Similar analyses regarding the targeting of PIK3CA (in CRISPR data) simulated, in the model, by the inhibition of PI3K node, can be performed (Figure 6.8C). **Low correlations are an indication highlighting the insufficient regulation of the node, probably confirming the scope issues raised in the pan-cancer-preliminary analysis**.

### 6.2.5 Comparison of the mechanistic approach with machine learning methods

In order to provide comparison elements unbiased by prior knowledge or by the construction of the model, we performed some simple machine learning algorithms. Random forests are used as an example of a machine learning approach to compare with mechanistic models and are implemented with *randomForestSRC* R package [Breiman, 2001a]. Random forests can be seen as an aggregation of decision trees, each trained on a different training set formed by uniform sampling with replacement of the original cohort. Prediction performances are computed using out-of-the bag estimates for each individual (i.e, average estimate from trees that did not contain the individual in their bootstrap training sample) and summarized as percentage of variance explained by the random forest. In this case, random forests have been fitted with inputs (mutations and/or RNA data) and outputs (sensitivities to drug or CRISPR BRAF inhibition) similar to those of logical models and the corresponding predictive accuracies are reported in Figure 6.9A. The first insight concerns data processing. The percentages of variance explained by the models are similar (around 70% of explained variance for drug sensitivity prediction) in the following three cases: unprocessed original data (thousands of genes), unprocessed original data for model-related genes only (tens of genes), and processed profiles of cell lines (tens of genes). This supports the choice of a model with a small number of relevant genes, which appear to contain most of the information needed for prediction. Second, the absolute level of performance appears much lower for CRISPR (between 30 and 50%) probably suffering from the lower number of samples, especially in cases where the number of variables is the highest. This tends to **reinforce the interest of mechanistic approaches that do not use any training on the data for smaller datasets, less suitable for learning**. Finally, while mutations and RNA data seem to provide the same predictive power (especially for drugs), using the two together does not necessarily result in a better performance in this case.

It is also possible to compute the variable importance that assesses the contribution of variables to the overall performance. The solution adopted in this paper to measure it, and called VIMP in the package, consists in introducing random permutations between individuals for the values of a variable and quantifying the variation in performance resulting from this addition of noise. In the case of key variables for prediction, this perturbation will decrease the performance and will result in a high variable importance

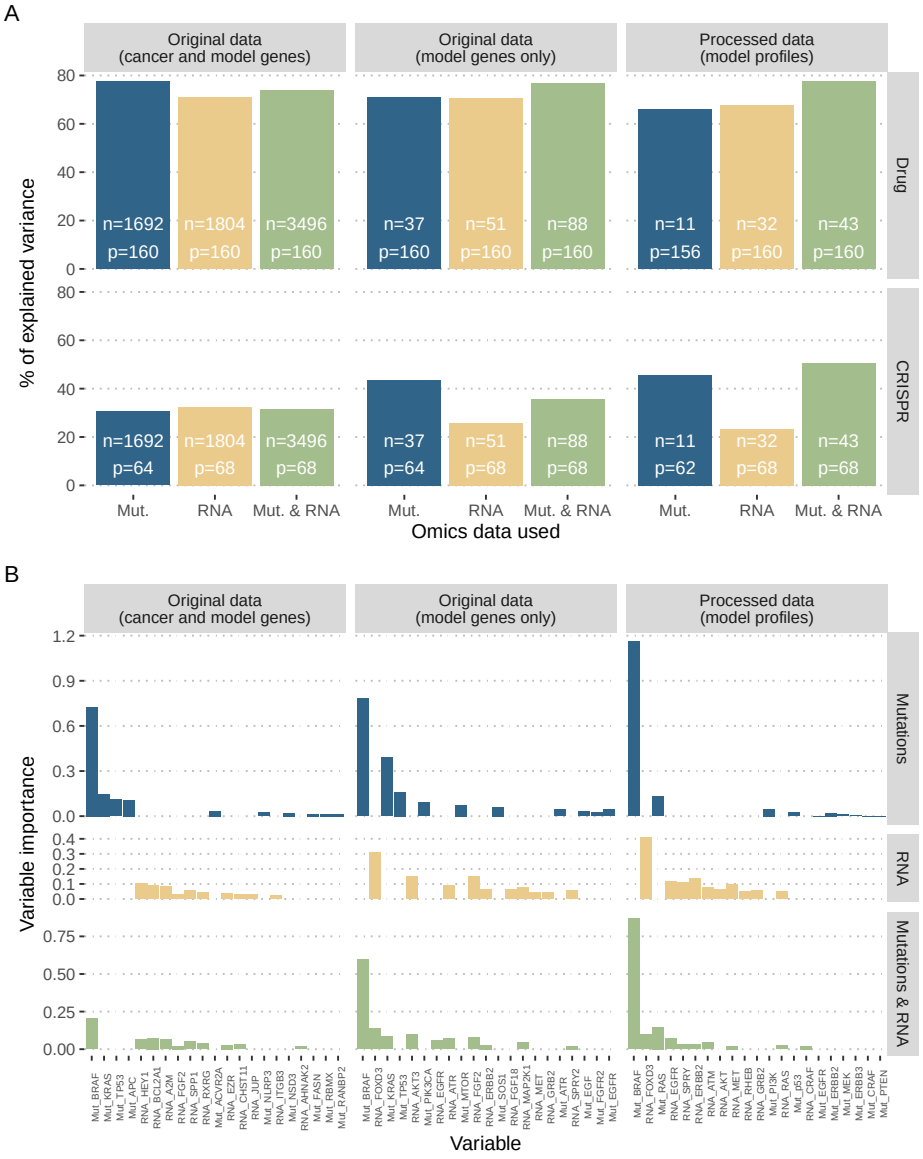



Figure 6.9: **Random forests to predict and explain sensitivity to BRAF inhibition.** (A) Performances of random forests for BRAF sensitivity prediction measured with percentage of explained variance; different learning task with unprocessed original data (thousands of genes), unprocessed original data for model-related genes only (tens of genes), and processed profiles of cell lines (tens of genes); $n$ samples and $p$ variables per learning task. (B) Variable importance for drug prediction only, with the 10 best variables with positive importance for each case.

[Ishwaran, 2007]. **Variable importance in these different random forests are reported in Figure 6.9B and are consistent with the analysis of mechanistic models**. The mutational status of BRAF is definitely the most important variable followed by mutations in RAS or TP53. Concerning RNA levels, the most explanatory variables seem to be FOXD3 or PTEN, in line with the definition of the logical model.

## 6.3 Application on prostate cancer study and challenges

Before summarizing the potential and limitations of the PROFILE approaches described in this and the previous chapter, a final example may be mentioned. Indeed, another application of the PROFILE method, quite similar to the examples presented in the previous and this chapter, has been carried out on prostate cancer. Chronologically, this project was one of the first applications of the method. However, as this project was more collaborative than personal, the previous chapters have been illustrated by more exclusively personal work when they were equivalent. We will therefore only briefly mention here the differences and insights specific to this study.

First, a logical model specific to prostate cancer was developed by some collaborators (Pauline Traynard and Arnau Montagud) over a long period of time, resulting in a large and comprehensive model of 146 nodes, which is described in more detail in the appendix B.4 and Figure B.3. Using the TCGA prostate cancer dataset (A.3.3) prognostic validation of the model was first carried out, similarly to Figure 5.7, by comparing individualized scores of some phenotypes in the model (*i.e.*, *Proliferation*) with clinical markers, in this case Gleason score, a grading system specific to prostate cancer. The qualitative evolution of the personalized *Proliferation* scores is also qualitatively validated (predicted proliferating tumors are on average of higher grade) but, despite the specificity and magnitude of the model, much of the variability is not explained.

The use of cell line data was also explored using Cell Model Passports data, restricted to the 7 prostate cell lines. The size of the model then allows qualitative predictions to be made on the proliferative, apoptotic and metastatic qualities of the different lines. Except for proliferation, however, experimental validation of the relevance of these predictions is difficult using public data or the literature. But again, after these preliminary validations, the focus of the study was on treatment response with a slightly different

rationale than in the previous example. Focusing on a particular cell line (LnCaP) and its corresponding personalized logical model, the idea is to **simulate with the models all possible inhibitions or combinations of inhibitions in order to identify possible vulnerabilities or relevant treatment synergies**. Experimental validation on the cell line was then carried out for certain genes that could be targeted depending on the existence of the treatments. The **efficacy of certain inhibitions highlighted by the simulations, such as that of HSP90, was confirmed experimentally** on this particular cell line. Despite the limitations of the approach in this application to prostate cancer, the study demonstrates the feasibility of the method for investigating the complexity of therapeutic responses and guiding experimental validation.

## 6.4 Limitations and perspectives

The emergence of high-throughput data has made it easier to generate models for prognostic or diagnostic predictions in the field of cancer. The numerous lists of genes, biomarkers and predictors proposed have, however, often been difficult to interpret because of their sometimes uncertain clinical impact and little overlap between competing approaches [Domany, 2014]. Methods that can be interpreted by design, which integrate *a priori* biological knowledge, therefore appear to be an interesting complement able to reconcile the omics data generated and the knowledge accumulated in the literature.

These benefits come at the cost of having **accurate expert description of the problem** to provide a relevant basis to the mechanistic models. This is particularly true in this work since the personalized models all derive from the same structure (the initial generic logical model) of which they are partially constrained versions. It is therefore necessary to have a generic model that is both sufficiently accurate and broad enough so that the data integration allows the expression of the singularities of each cell line. If this is not the case, the learning of logical rules or the use of ensemble modeling could be favoured, usually including perturbation time-series data [Razzaq et al., 2018]. It should also be noted that, in the logical models presented here, the translation of biological knowledge into a logical rule is not necessarily deterministic and unambiguous. The choices here have been made based on the interpretation of the literature only. And the presence of certain outliers, *i.e.*, cell lines whose behaviour is not explained by the models, may indeed result from the limitations of the model, either in its

scope (important genes not integrated), or in its definition (incorrect logical rules). More global or data-driven approaches to define the model would be possible but would require different training/validation steps and different sets of data.

The **second key point is the omics data used**. For practical reasons, we have focused on mutation and RNA data. The legitimacy of the former is not in doubt, but their interpretation is, on the other hand, a crucial point whose relevance must be systematically verified. The omission or over-interpretation of certain mutations can severely affect the behaviour of personalized models. Validation using sensitivity data provides a good indicator in this respect. However, the question is broader for RNA data: are they relevant data to be used to personalize models, *i.e.*, can they be considered as **good proxies for node activity?** The protein nature of many nodes in the model would encourage the use of protein level data instead, or even phosphorylation levels if they were available for these data. One perspective could even be to push personalization to the point of defining different types of data or even different personalization strategies for each node according to the knowledge of the mechanisms at work in the corresponding biological entity. A balance should then be found to allow a certain degree of automation in the code and to avoid overfitting.

Despite these limitations, the results described above support the **importance of combining the integration of different types of data to better explain differences in drug sensitivities**. There was no doubt about this position of principle in general [Azuaje, 2017], and in particular in machine learning methods [Costello et al., 2014, Aben et al., 2016]. The technical implementation of these multi-omic integrations is nevertheless more difficult in mechanistic models where the relationships between the different types of data need to be more explicitly formulated [Klinger et al., 2013]. The present work therefore reinforces the possibility and value of integrating different types of data in a mechanistic framework to improve relevance and interpretation and illustrates this by highlighting the value of RNA data in addition to mutation data in predicting the response of cell lines to BRAF inhibition. In addition, one piece of data that could be further exploited is that of the specific behaviour of the drugs or inhibitors studied, since for instance some BRAF inhibitors have affinities that vary according to mutations in the BRAF gene itself. The integration of truly precise data on the nature of the drug is nevertheless limited by logical formalism and is more often found in more flexible approaches, *e.g.*, in deep learning [Manica et al., 2019].

The application presented in this chapter, focused on BRAF inhibitors, made it possible to verify the good performance of the models through different types of data (drug or CRISPR/Cas9). However, the molecular profiles used to personalize the models were all derived from cell lines, reported in the same database [van der Meer et al., 2019]. It would be possible to use different types of data such as organoids, patient-derived xenografts (PDX), etc. The critical clinical question will then be: **do the mechanisms highlighted for cell lines transfer easily to tumours in vivo?** The ability to identify common reasons explaining the response to treatments has been studied by different statistical approaches with the aim of promoting translational medicine [Mourragui et al., 2019, Kim et al., 2019]. The ability of personalized mechanistic models to follow this path remains to be explored.

To conclude, we provide a comprehensive pipeline from clinical question to a validated mechanistic model which uses different types of omics data and adapts to dozens of different cell lines. This work, which is **based only on the interpretation of data and not on the training of the model**, continues some previous work that has already demonstrated the value of mechanistic approaches to answer questions about response to treatment, especially using dynamic data [Saez-Rodriguez and Blüthgen, 2020], and sometimes about similar pathways [Klinger et al., 2013]. In this context, our approach proves the interest of logical formalism to make use of scarce and static data facilitating application to a wide range of issues and datasets in a way that is sometimes complementary to learning-based approaches.

# Part III

# Statistical quantification of the clinical impact of models

Chapter 7

# Information flows in mechanistic models of cancer

*"Et l'effet qui s'en va nous decouvre les causes."*
*Alfred de Musset (Poésies nouvelles, 1843)*

The mechanistic models of cancer presented in the previous section have allowed us to integrate the omics data, to "make them speak" in order to better understand the clinical characteristics of cell lines or patients. But beyond their undeniable intellectual and scientific interest, do they have a direct clinical utility? Given the abundance and complexity of patient data available to physicians, the use of computer tools and mathematical models is inevitable and increasingly frequent. Because of their explicit representation of phenomena, mechanistic models can provide a more easily understood alternative for physicians or patients. Is it therefore desirable and relevant to use these models in support of medical decision making? And how can their clinical validity and impact be rigorously measured?

First of all, the purpose of this chapter is to outline some of the limitations of the previously presented evaluations of mechanistic models, together

with some recommended statistical tools. These evaluations answered the question: do the models have any clinical utility? We will show that an additional question could be: **do mechanistic models have an incremental clinical utility**, in comparison to the direct use of the data used to construct or specify them? This chapter is intended as a statistical introduction for systems biologists to some of the problems encountered in model evaluation.

**Scientific content**

This chapter is relying on literature for the first section and unpublished content for the second. The exploratory analyses presented below have helped to clarify considerations expressed qualitatively in previous chapters and formed the starting point for subsequent chapters on the clinical impact of cancer models.

## 7.1 Evaluation of models as biomarkers

### 7.1.1 Evaluation framework and general principles

First of all, mechanistic models of cancer should be considered as biomarkers among others, and therefore evaluated as such. This means focusing on the clinical information provided by the model outputs. In the previous examples, these outputs would be for example the $H/K_{50}/A$ biomarkers from Fey's model (described in section 3.4.2) or the personalized *Proliferation* scores from the mechanistic models in the examples in sections 5.2.2 or 6.2.4.1. The prognostic or predictive value of model outputs can then be evaluated according to the methods and recommendations present in the literature on prognostic or predictive biomarkers. Without going into too much detail, guidelines in this area are quite numerous and detailed, both for prognostic biomarkers [McShane et al., 2005, Sauerbrei et al., 2018] and predictive biomarkers [Janes et al., 2014]. Most of the points mentioned in these articles should apply identically for the particular type of biomarker that are the outputs of mechanistic models of cancer. The purpose of this thesis is not to exhaustively list these recommendations for the evaluation of biomarkers, so we will simply highlight the **most salient issues identified in the systems biology literature**.

### 7.1.2 Some frequent problems and recommended statistical tools

Concerning continuous and prognostic model outputs/biomarkers, they are sometimes confronted with naturally binary data (*e.g.*, event or not). In this case, many methods exist, among which the **Area Under the receiver operating Curve (ROC), usually denoted as AUC** [Søreide, 2009]. With a continuous biomarker $X$ and a binary outcome to predict $D$, the ROC curve plots sensitivity, $P(X > c|D = 1)$, against $(1 -$ specificity$) = (1 - P(X \leq c|D = 0))$, for all possible values $c$. the AUC is then simply computed as the area under this curve. The resulting AUC is computed as the area under this curve and measures the ability of the biomarker to discriminate between the two classes of interest and is a common tool for the evaluation of biomarkers.

However, prognostic validation often requires time-to-envent data such as survival data. Very schematically, if we study patients suffering from cancer and we are interested in the *death* event from a time $t_0$, which we define to be common to all patients, different cases are possible. Some patients have died and we therefore know their status and the time of their death $t_1$. Others are still alive at the time $t_{max}$ when the study stops (administrative censorship) or have withdrawn from the study at time $t_2$ so that their fate is then unknown. These patients are said to be right-censored. These data are extremely frequent and require specific methods such as Cox's proportional risk model [Cox, 1972] or accelerated failure time models. The **reasoned use of these dedicated survival models, and the associated assumptions, should be preferred to the forced binarization of survival data**, sometimes encountered in an apparent concern for simplification. In general, the validation of prognostic biomarkers using survival data therefore requires specific metrics such as **time dependent AUC for censored survival data** [Heagerty et al., 2000]. This measure, however, requires a more complex definition of sensitivity and specificity to accommodate censored data [Heagerty and Zheng, 2005] and to be applicable to real biomarker validation data [Buyse et al., 2006]. Despite fairly frequent use [Ching et al., 2018], the use of another metric called c-index is not recommended for assessing a model's ability to predict risk over a given time horizon [Blanche et al., 2019]. However, the clinical interpretation of AUC values is not straightforward and the complementarity of other approaches that focus on **the usefulness of risk models at the population level** rather than the ability to discriminate has been highlighted [Pepe et al., 2008]. The costs and benefits of prognostic models with various

AUC values have been studied in an applied context by Gail and Pfeiffer [2018].

Another frequent issue, already encountered in the examples from previous chapters, is the discretization of continuous markers. This is often done in order to classify patients into high and low risk groups for example for prognostic biomarkers. In the first place, although discretization may be required clinically, it is not necessary to evaluate the clinical value of the biomarker beforehand. Secondly, the choice of thresholds is crucial. In particular, in the case of biomarkers derived from mechanistic models, the artificial nature of the markers often makes difficult a binarization based on an *a priori* interpretation of the values. Choosing the cut-off point in order to maximise the significance or separation of the survival curves, as proposed in Fey et al. [2015] and presented in Figure 3.6, is however not recommended [Altman et al., 1994], among other things because it can be interpreted as uncorrected multiple testing. Such practices may thus contribute to the low clinical reproducibility of the contribution of certain biomarkers [Hilsenbeck et al., 1992]. For this problem in particular, tools have been proposed in the literature on clinical biomarkers, such as the **predictivenesse curve** [Mboup et al., 2020]. Similarly but in a more general framework, Janes et al. [2014] propose the use of **risk curves** to better evaluate predictive biomarkers beyond the crude computation of statistical interaction between the biomarker and the treatment in randomized clinical trial.

Another potential issue, particularly important in subsequent analyses (see section 7.2.1), is the **incremental value of biomarkers**. For instance, in the context of prognostic biomarkers, it is of course necessary to present univariable analyses showing the relationship between the marker and the outcome, which is almost systematically done, but also to question the value of this biomarker compared to other prognostic factors already known: does it add information or is it redundant? It is theoretically possible to consider the potential increase in AUC resulting from the addition of the new biomarker to the model. The increase in AUC, however, requires the addition of very strong markers [Gail and Pfeiffer, 2018], which has prompted the emergence of popular alternative metrics (Net Reclassification Index NRI, integrated discrimination improvement IDI) to evaluate the added predictive ability of a new marker [Pencina et al., 2008]. However, these metrics have been criticized as being unsafe since they can improve with the addition of non-informative markers [Hilden and Gerds, 2014, Pepe et al., 2014].

All in all, the first step in a good evaluation of mechanistic models would be that the **standards recommended for the evaluation of biomarkers can be applied in the same way to mechanistic models** that have certain applications or validation based on prognostic or predictive values. As these topics are well covered in the relevant literature, we will subsequently focus on a **specific vision of incremental value of biomarkers that is more specific to mechanistic models**.

## 7.2 Processing of biological information

Mechanistic models, and their outputs in particular, have so far been considered and evaluated as biomarkers. A comprehensive appreciation requires that they be seen as **information processing tools** in relation to the biological data they use. In this section, we will focus on a toy example to introduce some concepts. We will thus speak in general terms of the clinical value of this model, understood in the sense of a prognostic or predictive value depending on the application. The next section (7.3) will extend the same analyses to published models. The purpose of these two sections is to question the way in which mechanistic models process information. These **qualitative questions have been written essentially for those who design mechanistic models**. For the sake of technical simplification, the statistical tools chosen for illustration are therefore simpler than those presented in the previous section.

### 7.2.1 Information in, information out

Indeed, the mechanistic models presented in this thesis (Figures 3.6, 5.9 and 6.6) can be schematically represented by Figure 7.1: inputs $X$ (often omics data) are processed through a mechanistic model (here the grey box) to result in an output $Y$. These models can thus be assimilated to a mathematical transformation, often non-linear, of $X$ in $Y$. Thus, when validating the biological or clinical relevance of $Y$, either by calculating a correlation with the ground truth or by using it to stratify survival curves, only the univariate value of $Y$ is checked. This is an important step and a prerequisite for a well-constructed model. On the other hand, it is not sufficient information to understand how the model works. Indeed, the inputs $X$ probably also have a value: *e.g.*, if the mechanistic model uses different inputs, each of which has a prognostic value, the fact that the output also has a prognostic value does not necessarily indicate the relevance of the model. In short, **measuring only the output value of the model does not**

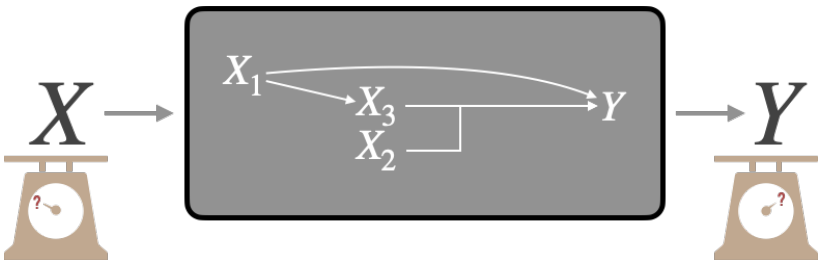



Figure 7.1: **Evaluation of a mechanistic model.** Adapted from Figure 1.6.

**necessarily reveal the model's ability to make sense of the data it uses**.

Therefore, the question of the incremental value of the model can be explained as follows: what does the output of the model represent in relation to the inputs? If we restrict ourselves to cases where the absolute biological/clinical value of $Y$ is positive, we can then identify two families of situations. First we can imagine a situation where the model has improved the value of the inputs: the output would then have a higher value than the inputs (better biological validation, better pronostic value etc.), or in any case a complementary value, a value not present in the inputs. This would correspond to the **capture by the model of emerging or non-linear effects**. For the sake of simplification, we will here assimilate the two in the sense that a non-linear effect resulting from the interaction between certain variables was indeed not predictable from the components taken individually, and therefore emergent. Note, for example, that the identification and capture by statistical models of non-linear components of treatment response is important in the ability to generalize findings from preclinical models to human tumours [Mourragui et al., 2020]. In the second situation, the output does not capture emergent properties but summarizes, totally or partially, the information present in the inputs. This would correspond to a **knowledge-informed dimensionality-reduction**. Even in the latter case, the scientific value of the model as a tool for understanding is not necessarily questioned. The analyses presented below are simply intended to supplement the understanding of models and how they process information.

### 7.2.2 Emergence of information in artificial examples

These questions can be illustrated using a very simple artificial model represented in Figure 7.2. On the one hand there are two latent biological

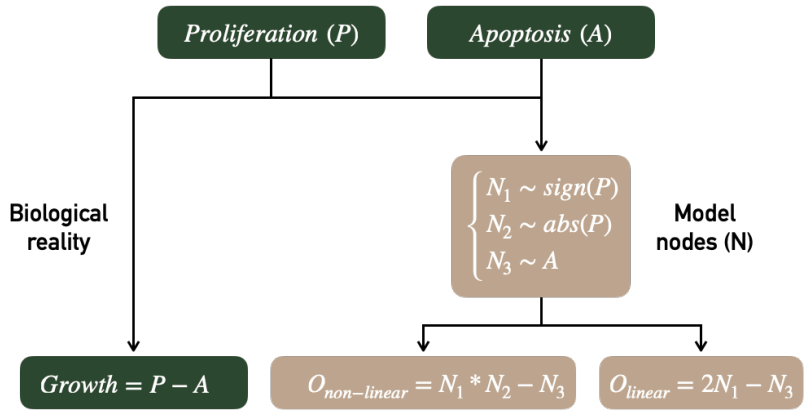



Figure 7.2: **Evaluation of a mechanistic model.** Adapted from Figure 1.6.

variables called *Proliferation (P)* and *Apoptosis (A)* resulting in our biological ground truth, *Growth*. On the other hand, the modeler has access to three different random variables $N_1$, $N_2$ and $N_3$ respectively associated with the sign of $P$, the absolute value of $P$ and the value of $A$. Two mechanistic models are defined, one linear (with its output $O_{linear}$) and one non-linear (with its output $O_{non-linear}$). We note that the two outputs are sufficiently well defined to be correlated with *Growth* but only the non-linear model makes use of $N_2$ by multiplying it with $N_1$.

The ability of models to use inputs to create or summarize information through outputs will be studied using the **explained variation metric** $R^2$. If a linear model is defined as $y_i = \beta_0 + \beta_1 x_i + e_i$, linear coefficients $\beta$ are estimated by minimizing the sum of squared differences between predicted and real values of $y$. The fitted model is written $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ and $R^2$ also called coefficient of determination is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Therefore $R^2$ measures the proportion of variation in $y$ that is explained by the regressors. A different way of expressing this is to say that $R^2$ compares the null model without covariate (observations are compared to their mean) to the linear model with covariates. By extension, it has been proposed to use R2 to assess the effect of adding a new biomarker to a pre-

viously established model[1] [Schemper, 2003]. In order to avoid overfitting, it is possible to calculate the adjusted $R^2$ that corrects with the number of regressors or to fit the model on training data and calculate the $R^2$ on validation data. The latter option was chosen using cross validation and averaging over the $R^2$ obtained in the different folds. Metrics with an interpretation similar to $R^2$ have been defined for logistic regressions or survival analysis [Choodari-Oskooei et al., 2012]. In the case of regressions with several variables $x_i$, it is possible to **decompose $R^2$ into different components associated with each of the variables**. This decomposition is carried out here by averaging over orderings according to the method proposed by Lindeman [1980] and applied in R code by Grömping et al. [2006]. The precise formulas are detailed in appendix C.1.1.

Here is an example of schematic reasoning that can be carried out with $R^2$ about the two models in Figure 7.2. We will **denote $R^2_{X_1+X_2}$ the $R^2$ corresponding to the linear model** $Growth = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ (written more compactly $Y \sim X_1 + X_2$, by analogy to its implementation in $R$). Using only the outputs of the models to predict *Growth*, explained variations are $R^2_{O_{non-linear}} = 0.455$ and $R^2_{O_{linear}} = 0.379$. The mechanistic models are thus correctly defined since the mechanistic output partly recover the biological read-out. However, the inputs of the model also have an important predictive value since $R^2_{N_1+N_2+N_3} = 0.514$. How can we understand the relationship between these values? First, the model including the $N_i$ inputs and the output $O_{linear}$ as regressors show identical performances with

$$R^2_{N_1+N_2+N_3+O_{linear}} = 0.514 = R^2_{N_1+N_2+N_3},$$

which means that $O_{linear}$ has no incremental value compared to a linear combination of the inputs. This is perfectly obvious from a statistical point of view since the two models are equivalent:

$$\begin{aligned} Growth &= \beta_0 + \beta_1 N_1 + \beta_2 N_2 + \beta_3 N_3 + \beta_4 O_{linear} \\ &= \beta_0 + (\beta_1 + 2\beta_4) N_1 + \beta_2 N_2 + (\beta_3 - \beta_4) N_3 \end{aligned}$$

The purpose of this example is to explicitly underline what is done implicitly in the study of certain mechanistic models. The complexity of

[1] An unpublished note by Frank Harrell details and illustrates the possibilities and limitations of $R^2$ for this type of analysis (link)

the described mechanisms sometimes hides more or less linear combinations of inputs that may make it possible to obtain meaningful biomarkers but without incremental value by construction. On the other hand, $O_{non-linear}$ has allowed to extract an emergent information which improves the global prediction when combined linearly with the inputs:

$$R^2_{N_1+N_2+N_3+O_{non-linear}} = 0.586 > R^2_{N_1+N_2+N_3}.$$

We can go further in understanding by breaking down the $R^2$. In Figure 7.3A and B (left columns), $R^2$ of the inputs' models ($Growth = \beta_0+\beta_1 N_1+\beta_2 N_2+\beta_3 N_3$) are decomposed to show that $N_1$ and $N_3$ contribute most to the prediction in a linear model. By using the same strategies for decomposing the $R_2$ and calculating the incremental $R^2$, it is also possible to **decompose the $R^2$ of $O_{linear}$ and $O_{non-linear}$ according to its origin: its component $N_1$ (0.22 in Figure 7.3A) is the proportion of $R^2$ that is also explained by $N_1$, so it can be interpreted as being the part of the value of $N_1$ captured by $O$.** In the non-linear case, we can see in the decomposition that $O_{non-linear}$ has an additional created component (0.07), it is the non-linear component that is not shared with any of the inputs.

In conclusion, if these two models generate meaningful outputs that are correlated with the biological read-out *Growth*, the analysis of their information processing classifies them into two different categories outlined in the previous sub-section. The linear model summarizes some of the information present in the inputs, without creating any. It can be likened to a relevant dimensionality reduction. The output of the non-linear model also fails to avoid some information losses, but at the same time it extracts new non-linear information. Thus, in combination with the inputs, it provides incremental value measured by the increase in total $R^2$. Note that $R^2$ is used here as one tool among others to illustrate the reflection on personalized mechanistic models as information processing tools. The point to remember is not technical but rather methodological: these **mechanistic models using on omics data cannot be evaluated for themselves but must be evaluated in comparison with the data they use in order to better explain the way they process information**. Following this rationale of model selection, other tools such as the Akaike Information Criterion (AIC) have been proposed and could allow to quantify if the reduction of dimension carried out by the models (from many omics inputs to one mechanistic output) allows a more parsimonious description of biology than the direct use of inputs [Kirk et al., 2013].

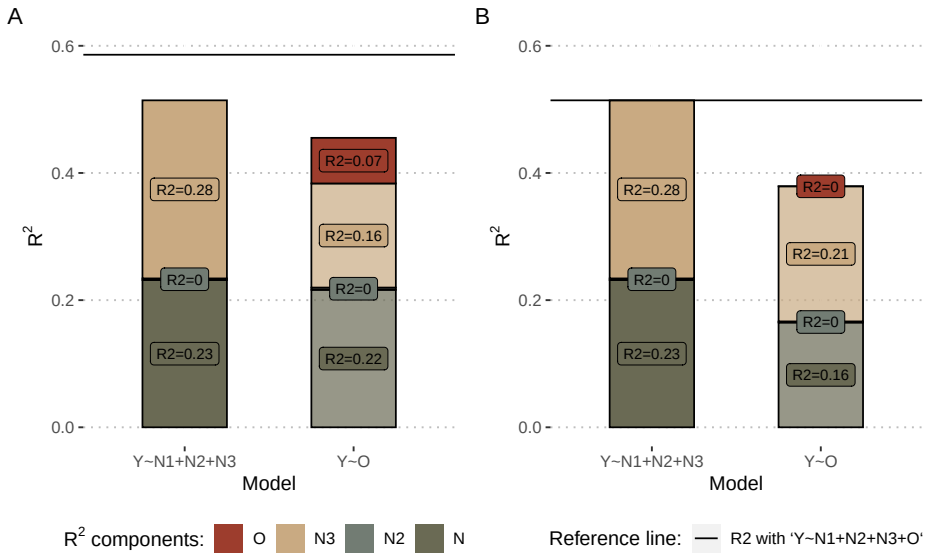



Figure 7.3: **Decomposition of $R^2$ for inputs and output of example models.** (A) Results for the non-linear model inputs and output $O_{non-linear}$ as defined in Figure 7.2: the left column represents the $R^2$ decomposition of model $Growth = \beta_0 + \beta_1 N_1 + \beta_2 N_2 + \beta_3 N_3$ and the right column the $R^2$ decomposition of $Growth = \beta_0 + \beta_4 O_{non-linear}$. (B) Same with the linear model and the corresponding $O_{linear}$. For both (A) and (B), colors represent the origin of $R^2$ contribution according to the decomposition. In particular, for right colums (model$Y \sim O$), the red share represent the proportion of the $R^2$ of the regressor $O$ that does not come linearly from the inputs, and therefore its emerging part. The horizontal reference line corresponds to the maximal $R^2$ obtained from the model $Growth = \beta_0 + \beta_1 N_1 + \beta_2 N_2 + \beta_3 N_3 + \beta_4 O$

## 7.3 Reanalysis of mechanistic models of cancer

Using the tools presented above, it is possible to deepen the analysis of some mechanistic models already presented in this thesis.

### 7.3.1 ODE model of JNK pathway by Fey et al. [2015]

One of the first applications of personalized mechanistic models to cancer is the one proposed by Fey et al. [2015] regarding JNK pathways in patients

with neuroblastomas. This work has been described in section 3.4.2 and is recalled in Figure 7.4. The evaluation of the mechanistic models in the original paper was performed by assessing the clinical value of the inputs (RNA levels of ZAK, MKK4, MKK7, JNK and AKT genes) and outputs ($H$, $A$ and $K_{50}$) separately by comparing them with survival data. The outputs were binarized to optimize the separation between the curves in a log-rank test. In this section we propose to **quantify the value of the output in relation to those of the inputs**, leaving the output continuous, using the tools described in the previous section. In the context of survival data, different measures called $R^2$ by analogy have been described in the literature. The one used thereafter was described by Royston and Sauerbrei [2004], its detailed definition is given in Appendix C.1.2 and its properties have been studied and validated in previous studies using simulated data [Choodari-Oskooei et al., 2012]. $R^2$ is not the preferred tool for survival data and is only used here to allow a qualitative description in line with the previous ones without introducing new tools. A formal and rigorous analysis should favour the tools presented at the beginning of the chapter.

Thus, the $R^2$ of the output $H$ is 0.39 while that of the combined inputs is 0.60. We can see from the decompositions that $H$ derives most of its the value from ZAK, MKK4 and AKT (Figure 7.5A, right column), which were already the largest contributors in the combined evaluation of the inputs (Figure 7.5A, left column). However, $H$ also includes an emerging non-linear share ($R^2 = 0.08$) that was not explained by the linear combination of inputs. Thus, incorporating $H$ with the inputs in a survival prediction model does indeed allow to observe an added value with a global $R^2$ of 0.68. In addition, the authors in the original study stressed the importance of positive feedback from JNK to MKK7 (Figure 7.4A). In its absence, we find that the value of $H$ is almost reduced to zero, since not only its non-linear part (Figure 7.5, red share), but also its parts derived from inputs, disappear. Analyzing the other outputs of the model ($A$ and $K_{50}$) reveals similar but less dramatic trends underlining the importance of this feedback which allows the model to capture a clinically relevant behaviour, assimilated by the authors to the capacity of cells to trigger apoptosis in case of stress. In the case of this model, the analyses provide a better understanding of how the model works with respect to survival prediction: **the outputs partly summarize clinical information already present in the inputs but also reveal relevant emerging information**.

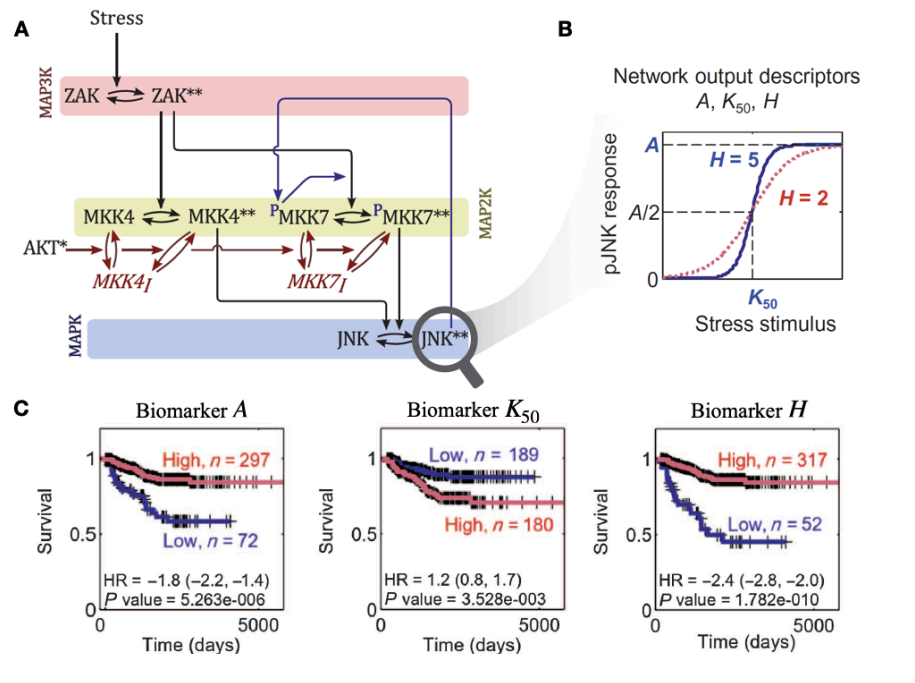



Figure 7.4: **Mechanistic modeling of JNK pathway and survival of neuroblastoma patients, as described by Fey et al. [2015].** (A) Schematic representation (as a process description [Le Novere, 2015]) for the ODE model of JNK pathway. (B) Response curve (phosphorylated JNK) as a function of the input stimulus (Stress) and characterization of the corresponding sigmoidal function with maximum amplitude $A$, Hill exponent $H$ and activation threshold $K_{50}$. (C) Survival curves for neuroblastoma patients based on binarized $A$, $K_{50}$ and $H$; binarization thresholds having been defined based on optimization screening on calibration cohort.

### 7.3.2 Personalized logical models: BRAF inhibition in melanoma and colorectal cancers

Similarly, it is appropriate to assess the relevance of the personalized logical models presented so far. Unlike the models of the previous sub-section, however, they integrate a much larger number of variables and the decomposition of $R^2$ is no longer accessible, because of its computational cost, which increases exponentially with the number of variables. If we focus on the example best suited to these models, that of BRAF inhibition sensitivity, we can however reformulate the question more simply. Given that the most important predictor of the answer is the status of the BRAF mutation

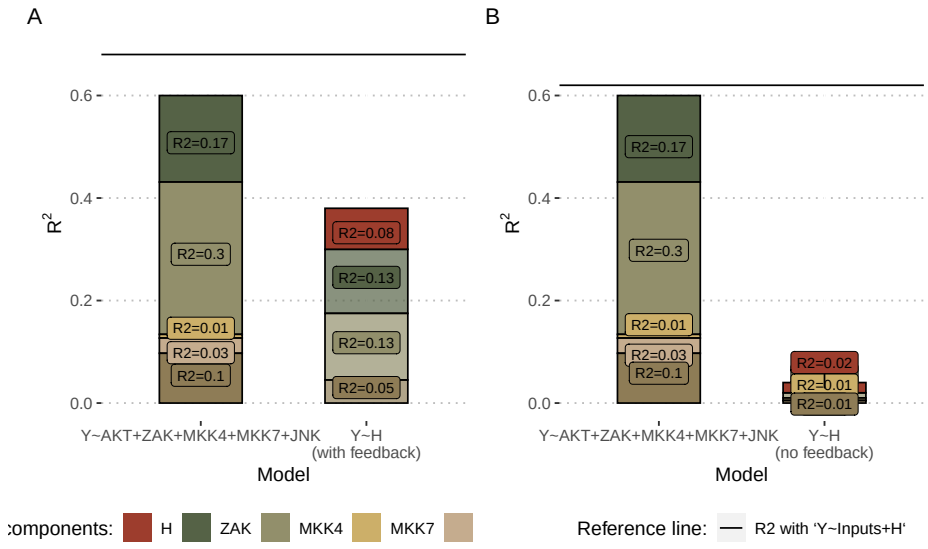



Figure 7.5: **Decomposition of $R^2$ for inputs and output for ODE model in Fey et al. [2015].** (A) Results for the Fey model inputs and output $H$ as defined in Figure 7.4A and B. (B) Same using the model without positive feedback between JNK and MKK7. Colors represent the origine of $R^2$ contribution. In particular, for right colums (model$Y \sim H$), the red share represent the proportion of the $R^2$ of the output $H$ that does not come linearly from the inputs, and therefore its emerging part.

itself, **do the personalized models allow us to do better or provide additional information?** In the case of CRISPR data, the $R^2$ of BRAF alone is 0.75, the $R^2$ of the personalized scores from the models is 0.73, while the combination of the two increases the $R^2$ to 0.83. In the absence of a precise decomposition, this gain can come either from the contribution of the other variables used in the model (the RNA levels of CRAF for example) or from the emergence of non-linear effects. In both cases, these figures are another way of expressing the remarks in section 6.2.4.1: thanks to the integration of other data and their organization in a framework based on literature knowledge, the model provides a more precise and complete vision of the response mechanisms. As positive as it is, this increase in $R^2$ remains modest, illustrating that the **main interest of these models is not necessarily a pure gain in predictive performance. Rather, it lies in their explanatory capacity and in their ability to support the investigation of mechanisms** such as in section 6.2.4.1. In a complementary way, one could imagine extending these analyses to other nodes of the model and not only to its output in order to dissect even more precisely

the information processing within the model.

Chapter 8

# Clinical evidence generation and causal inference

*"Maudit*
*soit le père de l'épouse*
*du forgeron qui forgea le fer de la cognée*
*avec laquelle le bûcheron abattit le chêne*
*dans lequel on sculpta le lit*
*où fut engendré l'arrière-grand-père*
*de l'homme qui conduisit la voiture*
*dans laquelle ta mère*
*rencontra ton père!"*
*Robert Desnos (La colombe de l'arche, 1923)*

The previous chapter introduced some tools to evaluate and quantify the value of mechanistic models, and in particular their outputs, with simple statistical tools. The latter, such as $R^2$, are by no means specific to medical applications. One of the particularities of mechanistic cancer models, on the other hand, is the possibility of simulating treatments that imitate therapeutic interventions. Before tackling more precise questions, this chapter will therefore introduce certain clinical or statistical

methods used to evaluate the effect of different types of treatments on patients. A more specific issue related to the evaluation of mechanistic models will be explored in the next chapter using these methods.

**Scientific content**

This short chapter introduces the framework of causal inference based on the literature and the description of causal inference in the preprint Béal and Latouche [2020].

## 8.1 Clinical trials and beyond

### 8.1.1 Randomized clinical trials as gold standards

When it comes to evaluating the effect of a therapeutic intervention, the reference method in most cases in modern medicine is the randomized clinical trial, which will be described now in its simplest version. Without loss of generality, the rationale for this approach can be detailed for one drug, which will be referred to as $A$ in the remainder of the chapter (Figure 8.1). The patients who can benefit from this drug, and therefore those eligible for the clinical trial, are first of all defined (specific disease, characteristics, etc.). Then, they are randomly separated into two distinct groups, one receiving the new treatment to be evaluated ($A = 1$) and the other generally receiving the treatment considered as standard of care, or a placebo if no validated treatment is available ($A = 0$). A predefined treatment response criterion $Y$ (viral load, tumor size, etc.) is then compared for the two groups to quantify the average treatment effect (ATE):

$$ATE = E[Y|A = 1] - E[Y|A = 0]$$

Thus it will be possible to say, for example, that "compared to patients who received the standard treatment, those treated with the new drug have a 20% lower tumor volume". In this example, **randomly choosing how the two groups of patients, treated and untreated, are constituted ensures *a priori* that the two groups are comparable**. Indeed, it should be verified that the untreated patients were not on average suffering from more advanced cancers that are more likely to proliferate and grow. In this case, the difference in outcome between the groups could simply

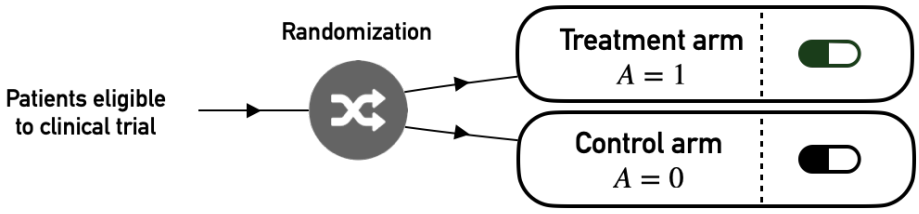



Figure 8.1: **Principles of randomized clinical trials.**. This trial evaluates the impact of treatment $A$.

come from a difference in initial composition and not from a difference derived from therapeutic interventions. Random assignment of treatments therefore offers minimum guarantees concerning the characteristics of the two subgroups.

### 8.1.2 Observational data and confounding factors

The problem of comparability between the two groups is reinforced when the data used does not come from a randomized clinical trial. In the remainder of this thesis these data will be called **observational data**. This means that in the available data, some patients were treated with the new drug ($A = 1$) and others received the reference treatment ($A = 0$). However, the assignment of treatment was not decided by the observer. This assignment was therefore made according to a protocol unknown to the observer which has no guarantee that the two groups are in fact comparable.

The situation can be illustrated with a simple simulated example involving a confounding variable $C$ in addition to the treatment variable $A$ and the outcome variable $Y$. If $Y$ represents tumor volume and $A$ the treatment to be evaluated, $C$ could be a biomarker of cancer agressiveness. 1000 patients have been simulated for all variables in two different settings represented in Figures 8.2 and 8.3. In the first case (Figure 8.2), the outcome $Y$ is positively correlated to $C$ (more agressive tumors have bigger volume) and decreased when $A = 1$ (treatment decreases tumor volume). $C$ **has no influence on** $A$. The causal relationships between the variables and the associated coefficients used to simulate data are summarized in the directed acyclic graphs (DAG) in Figure 8.2A. The observed relations between variables in simulated data are shown in Figure 8.2B, C and D. In particular, the **theoretical influence of** $A$ **on** $Y$ **is recovered in the observed data** since $E[Y|A = 1] - E[Y|A = 0] = -5.05$

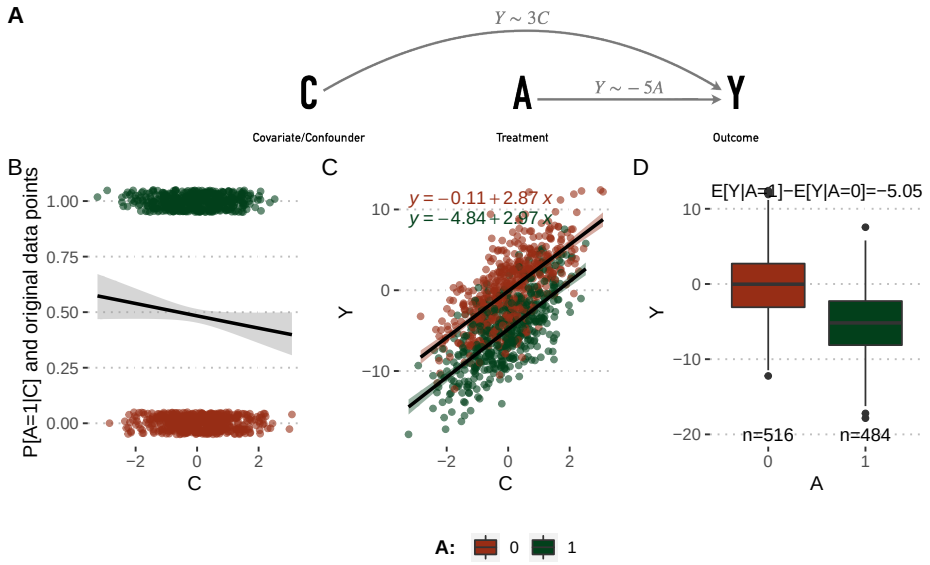



Figure 8.2: **Analysis on observed data without confounder.** (A) Directed acyclic graphs with causal relations between variables and parameters used to simulate data. (B) Influence of $C$ on $A$ in observed simulated data. (C) Same with $C$ and $Y$. (D) Same with $A$ and $Y$.

In the second case (Figure 8.3), $C$ has an influence on $Y$: the more aggressive the tumor, the more likely the patient is to be treated with the new drug. In this case the **simultaneous influence of $C$ on $A$ and $Y$ makes it a real confounder**. The direct observation of the differences in outcomes between treated and untreated patients reveals only a small benefit of the new treatment which does not correspond to the underlying reality used in these simulations since the theoretical causal influence of $A$ on $Y$ remained the same as in the previous case. The **confounding factor prevents the nature of the causal link between A and Y from being simply inferred**.

## 8.2 Causal inference methods to leverage data

Despite these difficulties, some statistical methods have been developed to derive estimates with a causal interpretation from observational data, under precise assumptions. This work will focus on the potential outcomes framework [Rubin, 1974]. We will first describe briefly the fundamentals of

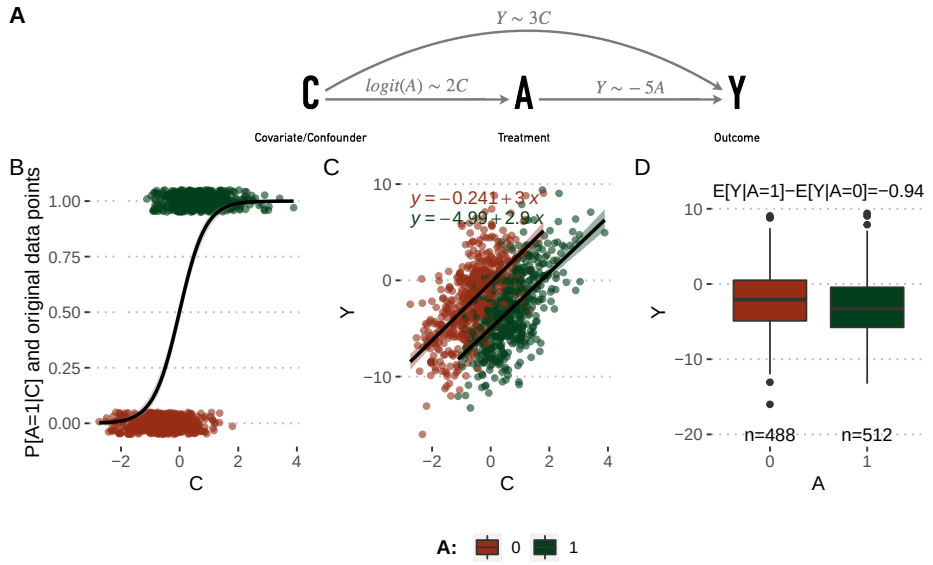



Figure 8.3: **Analysis on observed data with confounder.** (A) Directed acyclic graphs with causal relations between variables and parameters used to simulate data. (B) Influence of $C$ on $A$ in observed simulated data. (C) Same with $C$ and $Y$. (D) Same with $A$ and $Y$.

this framework and different methods that are part of it.

### 8.2.1 Notations in potential outcomes framework

First of all, the notations used in this and the next chapter are defined as follows. We will use $j = 1, ..., N$ to index the individuals in the population. $A_j$ and $Y_j$ correspond respectively to the actual treatment received by individual $j$ and the outcome. In the most simple case, treatment takes values in $\mathcal{A} = \{0, 1\}$, 1 denoting the treated patients and 0 the control ones. $Y_j$ corresponds to the patient's response to treatment. In the case of cancer it may be a continuous value (*e.g.*, size of tumor), a binary value (*e.g.*, status or event indicator), or even a time-to-event (*e.g.*, time to relapse or death). Only the first two cases will be discussed later. Finally, it is necessary to take into account the possible presence of confounders influencing both $A$ and $Y$ and denoted $C_j$ for individual $j$.

The **potential outcomes framework is also described as counterfactual** because it defines variables like $Y_j(a)$ to denote the potential outcome of individual $j$ in case he/she has been treated by $A = a$ which may be different from what we observe if $A_j \neq a$. This definition can be

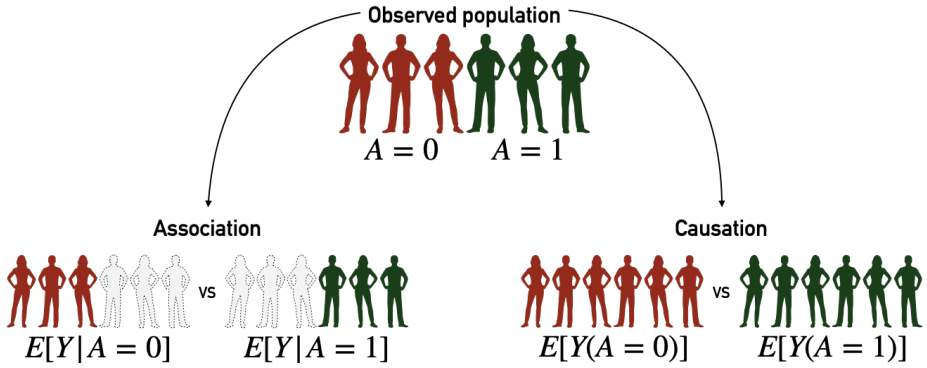



Figure 8.4: **Association, causation and their associated cohorts.** Association analyses are based on observed cohorts and conditional probabilities. Causation analyses are based on counterfactual variables and cohorts.

illustrated at the individual level, for patient $j$, where $A$ is the smoking status (1 for smokers, 0*otherwise*) and $Y$ is the outcome, *e.g.*, cancer status at a given date. If patient $j$ is a smoker then $Y_j = Y_j(A = 1)$. $Y_j(A = 0)$ would be the outcome if this same patient had not been a smoker, all other things being equal. This counterfactual outcome is therefore not observed in the data. These counterfactual variables make it possible to write the causal estimands. For instance, in this context, we can easily compute the difference in outcome between treated patients and control patients (Figure 8.4, left part): $E[Y|A = 1] - E[Y|A = 0]$. However, this difference has no causal interpretation as it does not offer any guarantees as to the confounding factor, as an unbalanced distribution of $C$ can induce biases. Thus we define another estimate: $E[Y(1)] - E[Y(0)]$. In this case, we compare between two ideal cohorts (Figure 8.4, right part), one in which all patients have been treated (possibly contrary to the fact) and one in which all patients have been left in the control arm (once again, possibly contrary to the fact).

### 8.2.2 Identification of causal effects

The next question is whether it is possible to estimate the counterfactual variables $Y(A)$ and under what conditions. The potential outcomes framework explicits **assumptions of consistency, positivity and conditional exchangeability to estimate these counterfactual variables** and therefore infer causal estimates from observational (non-randomized) data [Rubin, 1974, Hernán and Robins, 2020].

*Consistency* means that values of treatment under comparison represent well-defined interventions which themselves correspond to the treatments in the data: if $A_j = a$, then $Y_j(a) = Y_j$.

*Exchangeability* means that treated and control patients are exchangeable, *i.e.*, if the treated patients had not been treated they would have had the same outcomes as the controls, and conversely. Since we usually observe some confounders we define conditional exchangeability to hold if cohorts are exchangeable for same values of confounding $C$. Therefore conditional exchangeability will hold if there is no unmeasured confounding: $Y(a) \perp\!\!\!\perp A|C$.

*Positivity* assumption states that the probability of being administered a certain version of treatment conditional on $C$ is greater than zero: if $P[C = c] \neq 0, P[A = a|C = c] > 0$. Intuitively, this positivity condition is required to ensure that the defined counterfactual variables make sense and do not represent something that cannot exist.

Under these three assumptions, there are different methods and estimators available to evaluate causal effects from observational data. Two of them will be described and applied to the same example as above: the description of the example and the failure of the direct methods are recalled in Figure 8.5A and B and two causal inference methods are illustrated in Figure 8.5C, D and E.

#### 8.2.2.1 Standardization or parametric g-formula

The first method is called standardization or parametric g-formula and it is the one that will be described in more detail in this chapter and the following one. It is based on the following equations:

$$\begin{aligned} E[Y(a)] &= \sum_c E[Y(a)|c] \times P[c] \\ &= \sum_c E[Y(a)|a,c] \times P[c] \quad \text{(exchangeability } Y(a) \perp\!\!\!\perp A|C \text{ )} \\ &= \sum_c E[Y|a,c] \times P[c] \qquad \text{(consistency)} \end{aligned}$$

Thus the average effect of treatment on the entire cohort can be written with standardized means:

$$E[Y(A=1)]-E[Y(A=0)] = \sum_c \Big( E[Y|A=1, C=c] - E[Y|A=0, C=c] \Big) \times P[C=c] \quad (8.1)$$

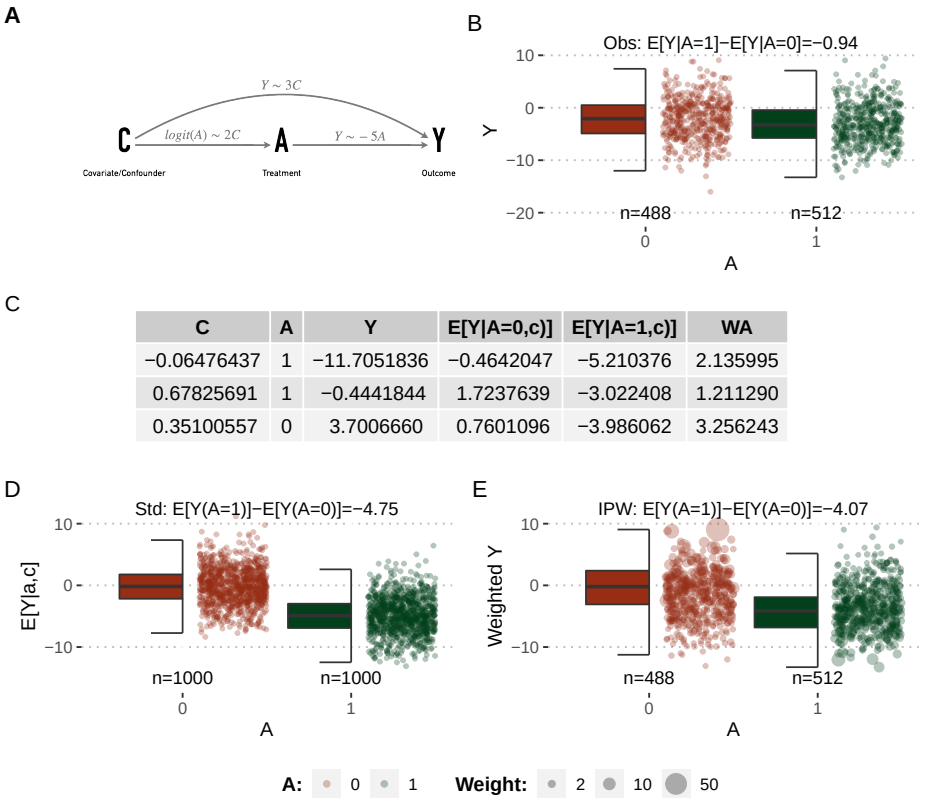
Computationally, non-parametric estimation of $E[Y|A=a, C=c]$ is usually out of reach. Thus, on real-world dataset, $E[Y|A=a, C=c]$ is estimated through **outcome modeling** and explicit computation $P[C=c]$ is replaced by its empirical estimate. The nature of the statistical model used will be specified in the various applications presented. In the simple example depicted in Figure 8.5A, a linear model of the outcome ($Y \sim C+A$) is fitted on observed data. This model is then used to infer $E[Y|A=1, C=c]$ and $E[Y|A=0, C=c]$ for each patient with covariate $C=c$ (Figure 8.5C). By averaging these values over the whole cohort the confounding effect is corrected and the estimator is much closer to the true value $-5$ than the naive estimates (Figure 8.5D and B).

#### 8.2.2.2 Inverse probability weighting (IPW) and propensity scores

Based on the same counterfactual framework, it is possible to build another class of models, called marginal structural models [Robins et al., 2000], from which we derive estimators different from the standardized estimators called inverse-probability-of-treatment weighted (IPW) estimators [Cole and Hernán, 2008]. IP weighting is equivalent to creating a **pseudo-population where the link between covariates and treatment is cancelled**. In the case of binary treatment $A \in 0, 1$, weights are defined for each patient as the inverse of the probability to have received the version of treatment he or she actually received, knowing his or her covariates:

$$W^A = \frac{1}{f[A|C]} \text{ with } f[a|c] = P[A=a|C=c],$$

$f[a|c]$ being called the propensity score, *i.e.*, the probability to have received the treatment $A=a$, given the covariates $C=c$. Again, propensity scores will be estimated in later examples using a parametric model. In this case with a binary treatment $A$, a logistic **treatment model** is used ($A \sim C$) to derive the weights $W^A$ (Figure 8.5C). Note that propensity scores are also useful for positivity investigations since values very close to 0 or




| C | A | Y | E[Y\|A=0,c)] | E[Y\|A=1,c)] | WA |
|---|---|---|---|---|---|
| −0.06476437 | 1 | −11.7051836 | −0.4642047 | −5.210376 | 2.135995 |
| 0.67825691 | 1 | −0.4441844 | 1.7237639 | −3.022408 | 1.211290 |
| 0.35100557 | 0 | 3.7006660 | 0.7601096 | −3.986062 | 3.256243 |

Figure 8.5: **Causal inference methods on a simple example.** (A) Directed acyclic graphs with causal relations between variables and parameters used to simulate data. (B) Association between $A$ and $Y$ from observed data. (C) Some simulated samples/patients with their original variables ($C$, $A$ and $Y$), variables from outcome model ($E[Y|A=0,c)]$, $E[Y|A=1,c)]$) and weights from treatment model ($W^A$). (D) Standardized causal effect of $A$ on $Y$ based on and outcome modeling. (E) IPW causal effect of $A$ on $Y$ based on weights derived from treatment modeling; in this panel weights are taken into account in boxplots and estimations.

1 may indicate (quasi-)violations of positivity. Under the same hypothesis of exchangeability, positivity and consistency, we can derive the modified Horvitz-Thompson estimator [Horvitz and Thompson, 1952, Hernán and Robins, 2020]:

$$E[Y(a)] = \frac{\hat{E}[I(A=a)W^A Y]}{\hat{E}[I(A=a)W^A]}, \tag{8.2}$$

$I$ being the indicator function, such as $I(A = a) = 1$ if $A = a$ and $I(A = a) = 0$ if $A \neq a$. Once again, this method brings estimates closer to the true causal effect by correcting for the influence of the confounder (Figure 8.5E).

#### 8.2.2.3 Limitations and additional methods

These causal inference methods therefore allow to correct some biases due to observed confounders, at the cost of strong hypotheses that it is not possible to verify. The plausibility of these hypotheses, and therefore of the resulting estimates, requires a good knowledge of the context. Furthermore, the estimates are largely based on statical models of outcome or treatment. The correct specification of these models is therefore imperative to ensure unbiased causal estimates. In order to limit the risks of misspecification, some **doubly robust** approaches have also been developed. They require estimating both an outcome model and a treatment model, but the resulting estimates are consistent if at least one of the two models is correctly specified. One of these methodologies among others, called Targeted Maximum Likelihood Estimation (TMLE), will be mentioned in the next section and is detailed in appendix C.2.4.

In summary, evaluating the effect of a treatment requires isolating its impact from that of all confounding factors. This can be done in a randomized clinical trial designed for this purpose. However, there is a great amount of other data available that may not have been generated in this rigorous framework. It is nevertheless possible to draw causal interpretations from them, under certain hypotheses, thus offering insights for *a posteriori* statistical evaluation of specific therapeutic strategies.

CHAPTER 9

# Causal inference for precision medicine

*"Felix qui potuit rerum cognoscere causas."*
*Virgil (Georgics, 29 BC)*

Throughout this manuscript, we first described the complexity of cancer mechanisms, through the diversity of genetic alterations or non-linear signaling pathways. This complexity naturally led to the choice of systemic modeling approaches and in particular mechanistic models whose explicit nature facilitates the study of the effects of new molecular perturbations such as treatments. The simple study of the response to BRAF inhibitors has thus required the consideration of many other genes and pathways.

This final chapter proposes to take the complexity a step further by considering different treatments. The diversity of patients' molecular profiles suggests that the best treatment is not necessarily the same for all patients: this is what is known as precision medicine. This is already a clinical reality in oncology that could be reinforced in the future by the emergence of new computational models of cancer, whether mechanistic or not. **How then can we assess the relevance of these models in their ability to guide patient treatment?**

**Scientific content**

This chapter presents an extension of the causal inference framework to quantify the value of precision medicine strategies. This work is currently under revision and is available as a preprint in Béal and Latouche [2020]. All code is available in the dedicated GitHub repository

## 9.1 Precision medicine in oncology

It is first important to understand what is meant by the concept of precision medicine in the treatment of cancer patients in order to place subsequent questions in a plausible clinical framework.

### 9.1.1 An illustration with patient-derived xenografts

Precision medicine stems from the diversity of treatment responses observed in different tumors. It has already been observed in previous chapters about BRAF inhibition that **different cell lines respond differently to a particular treatment**. A broader analysis of pre-clinical data shows that the same is true for the vast majority of treatments. It would be possible to illustrate this using the same data from cell lines extended to other drugs. However, because of the more directly clinical impact of the issues discussed in this chapter, the analyses presented below will focus on another type of data that is closer to patient data: patient-derived xenografts (PDX).

A **PDX is a tumor tissue that has been removed from a patient and implanted into immunodeficient mice** [Hidalgo et al., 2014]. Unlike cell lines, which are *in vitro* models, PDXs are *in vivo* models that allow cancer cells to evolve in a more realistic microenvironment. In the same way as for cell lines, PDX can be used for drug screening. The data used in this chapter come from a study by Gao et al. [2015] which contains several hundred tumors and more than fifty drugs. Not all drugs have been tested for all tumors; details of the drugs and types of cancer tested are available in the appendix A.2. This dataset was generated following the "one animal per model per treatment" approach ($1 \times 1 \times 1$), the principles of which are summarized in Figure 9.1A. It should be noted that different drug response metrics are computed in the source data, two of which will be used in the analyses. The first one is continuous and called *Best Average Response* in

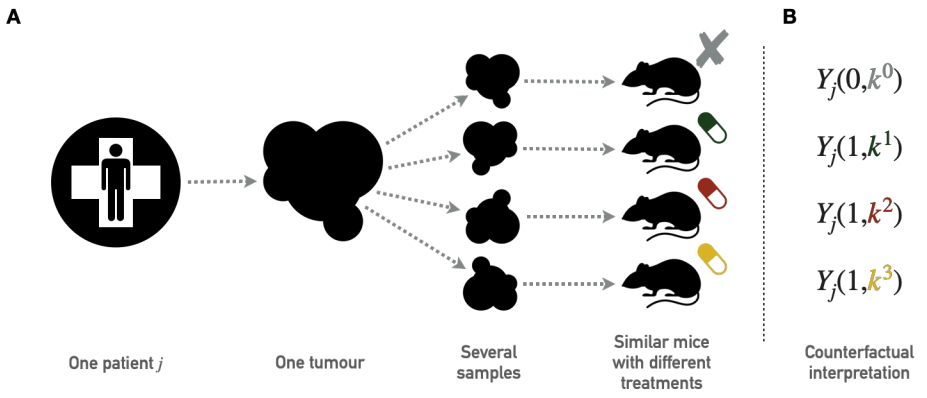



Figure 9.1: **Principles of PDX screening.** (A) Schematic pipeline for PDX screening with tumor biopsies from one patient divided in several pieces later implanted in similar immunodeficient mice. Each mouse is then treated with a different drug; the collection of mice that have received tumor samples from the same patient but have been treated with different drugs therefore gives access to several outcomes for the same tumor of origin. (B) Corresponding counterfactual variables.

the data, it is based on the variation of the tumor volume after treatment, the lower values (and especially negative) corresponding to better responses. The second one is originally categorical and based on a modified Response Evaluation Criteria In Solid Tumors (RECIST) criteria. It was binarized for this study so that the responders have a score of 1 and non-responders 0. The details of the definition and distribution of these metrics are given in appendix A.2.

In order to illustrate the diversity of response to treatment, the database is momentarily restricted to the 4 most widely tested drugs and the 180 tumors (or PDX models) that were evaluated for all four drugs. The four chosen drugs target different pathways: binimetinib (MAPK inhibitor), BKM120 (PIK inhibitor), HDM201 (MDM2 inhibitor) and LEE011 (CDK inhibitor). In Figure 9.2A the 4 treatments show a high variability of response, with a slight advantage for BKM120 and binimetinib on average over all tumors. However, **each of the treatments was found to be the most effective of the 4 for a significant proportion of tumors** (Figure 9.2B), with binimetinib and BKM being the best treatment for one-third of tumors each and LEE011/HDM201 sharing the remaining one-third of tumors. It thus appears that in view of the molecular diversity of tumors and the increasing number of treatments available, it does not seem

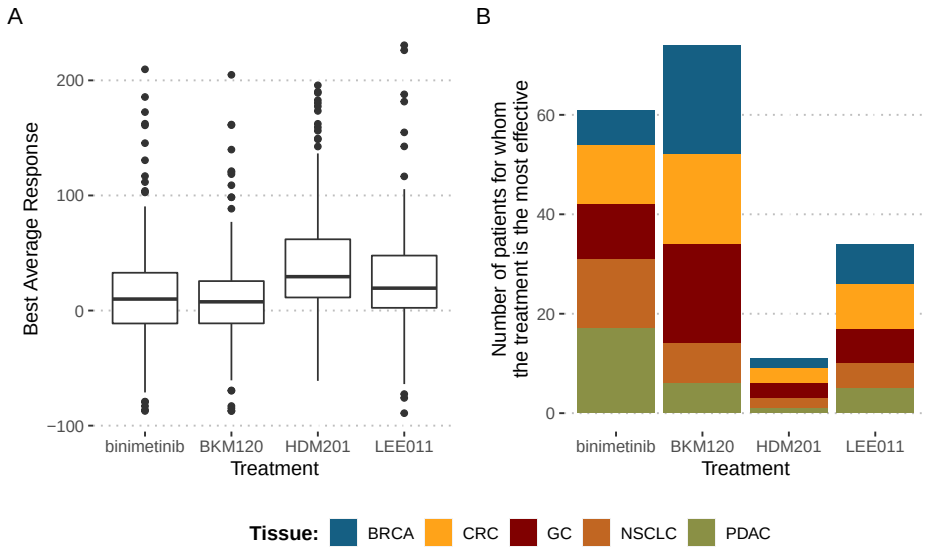



Figure 9.2: **Differences in drug response for 4 drugs and 180 tumors: a call for precision medicine.** (A) Distribution of treatment response for the 4 different drugs, each with all 180 tumors. (B) Number of times each of the 4 drugs is the most effective for a given tumor, distribution by tissue of origin.

advisable, according to these preclinical data, to treat all tumors with the same gold-standard treatment. Furthermore, the tissue of origin of the tumors in this example does not appear to be the main determinant of tumor preference for certain treatments.

### 9.1.2 Clinical trials and treatment algorithms

These remarks can be extended to patients. Thus, precision medicine (PM) consists in **assigning the most appropriate treatment to each patient according to his or her characteristics, usually genomic alterations for cancer patients** [Friedman et al., 2015, De Gramont et al., 2015]. At the individual level, targeted treatments have provided relevant solutions for patients with specific mutations [Abou-Jawde et al., 2003]. Putting together these various treatments, some precision medicine strategies can be defined. Based on the genomic profile of the patient, the treatment most likely to be successful is chosen. If the information available is reliable, **precision medicine can thus be reduced to a treatment algorithm that takes as input the molecular characteristics of the patient's tumor and outputs a recommendation of treatment**. An example of

| Targets | Molecular alterations | Molecularly targeted agents |
|---|---|---|
| **KIT, ABL1/2, RET** | Activating mutation† or amplification* | Imatinib 400 mg qd PO |
| **PI3KCA, AKT1**<br>**AKT2,3, mTOR,**<br>**RAPTOR, RICTOR**<br>**PTEN**<br><br>**STK11**<br>**INPP4B** | Activating mutation or amplification<br>Amplification<br>Amplification<br>Homozygous deletion or heterozygous deletion + inactivating mutation or heterozygous deletion + IHC confirmation<br>Homozygous deletion or heterozygous deletion + inactivating mutation<br>Homozygous deletion | Everolimus 10 mg qd PO |
| **BRAF** | Activating mutation or amplification | Vemurafenib 960 mg bid PO |
| **PDGFRA/B, FLT3** | Activating mutation or amplification | Sorafenib 400 mg bid PO |
| **EGFR** | Activating mutation or amplification | Erlotinib 150 mg qd PO |
| **ERBB2/HER2** | Activating mutation or amplification | Lapatinib 1000 mg qd PO +<br>Trastuzumab 8 mg/kg IV<br>followed by 6 mg/kg IV q3w |
| **SRC**<br>**EPHA2, LCK, YES1** | Activating mutation or amplification<br>Amplification | Dasatinib 70 mg bid PO |
| **ER, PR** | Protein expression >10% | Tamoxifen 20 mg qd PO<br>(or letrozole 2·5 mg qd PO if contra-indication) |
| **AR** | Protein expression >10% | Abiraterone 1000 mg qd PO |

Figure 9.3: **An example of a precision medicine treatment algorithm: the SHIVA clinical trial.** Specific molecular alterations and their associated treatments, as proposed in the SHIVA clinical trial [Le Tourneau et al., 2015].

such a treatment algorithm from the SHIVA clinical trial by Le Tourneau et al. [2015] is shown in Figure 9.3 where different treatments are associated with different alterations. In this case, the treatment algorithm can be considered as an aggregation of the medical knowledge accumulated on the individual biomarkers.

### 9.1.3 Computational models to assign cancer treatments

The treatment algorithm example in Figure 9.3 could, however, be more complex. Indeed, previous chapters have stressed, for example, that being mutated for the BRAF gene is not the only predictor of response to an inhibitor of BRAF (here Vemurafenib). The same is true for most treatments that could benefit from more global and systemic analyses, taking into account more variables and their interactions. This complexity would require the use of computational methods.

It is on this point that this chapter links to the previous ones. **Some of the cancer models studied throughout this thesis, or their future developments, could be interpreted as treatment algorithms**. Indeed, a model capable of predicting the response to a single treatment does not necessarily allow the inference of precision medicine strategies. On the other hand, a model capable of predicting a patient's response to different

treatments is also capable of indicating which one is the best. Such models would then move from systems biology to systems therapeutics [Hansen and Iyengar, 2013], taking patients' genomic features as inputs and outputting a treatment recommendation. In theory, mechanistic models seem to be suitable for this purpose since their explicit representation of genes and proteins makes it possible to simulate the effect of different therapeutic interventions. However, the feasibility of designing and calibrating such a model has yet to be demonstrated. Other types of models are being studied that could achieve these goals. For example, some recent approaches propose the use of deep learning to provide a computational tool for predicting the growth of cells [Ma et al., 2018], or even the sensitivity of cell lines to different treatments [Manica et al., 2019].

In short, if no computational model is sufficiently developed to date to replace the clinician, the emergence of this type of tool is likely in the medium term. This raises the question of **how to assess the clinical value of the precision medicine strategies (and corresponding treatment algorithms) derived from these models**. For the sake of generality, this question will be addressed more broadly in the following without reference to models as a possible source of the treatment algorithm: how to evaluate the clinical impact of a precision medicine strategy and the treatment algorithm? The methods presented will indeed be the same, whether the algorithm evaluated comes from a model or from the knowledge of clinicians as in Figure 9.3. In the spirit of this thesis, the question nevertheless finds its origin in the first hypothesis related to models.

## 9.2 Emulating clinical trials to evaluate precision medicine algorithms

### 9.2.1 Objectives and applications

The question then arises of how to quantify the clinical benefit provided by these treatment algorithms. Some **precision medicine clinical trials** have been proposed, demonstrating both the feasibility of collecting information about mutations [Le Tourneau et al., 2015] or RNA [Rodon et al., 2019] in real-time and the clinical benefit that can be expected from these approaches for some patients [Coyne et al., 2017]. However, the increasing abundance of genomic data and biological knowledge make it progressively easier to establish new algorithms for precision medicine, either directly based on physician knowledge or provided by computational mod-

els [Hansen and Iyengar, 2013]. For practical reasons it is not possible to propose a real clinical trial for each new precision medicine algorithm or for any variants, comparing standard of care with new algorithm-based therapeutic strategies.

Therefore, this work provides a method to assess the clinical impact of proposed PM treatment algorithm based on already generated data, **emulating precision medicine clinical trials and analyzing them in the causal inference framework** [Hernán and Robins, 2016]. First we will define the causal estimates of the precision medicine effects (later referred to as causal estimates) we want to assess, and the corresponding ideal clinical trials one would like to perform. Next, we will define the notations and the causal framework we use to infer the causal effects from observational data with multiple versions of treatment, based on the previous work by VanderWeele and Hernan [2013]. The main principles of the potential outcome framework having been introduced in the previous chapter, an extension to the case of precision medicine will be described, focusing on the multiplicity of treatment versions, *i.e.*, targeted drugs. Then we will apply these methods to simulated data in order to investigate the different biases of the candidate methods. An example scenario will be presented and a RShiny interactive application has been developed to further explore other user-defined settings. Finally, the analysis of data from patient-derived xenografts (PDX) makes it possible both to apply the methods to pre-clinical situation and to have data approximating the counterfactual responses, thus enabling further validation of the proposed estimation methods.

### 9.2.2 Target trials for precision medicine: definition of causal estimates

We first specify the precision medicine effects that are to be estimated. These effects will finally be estimated based on observational data through the causal framework and target trial emulation [Hernán and Robins, 2016]. In this context the notion of **target trial refers to the real clinical trial whose estimates are sought to be reproduced through causal inference**. Thus, if we think in terms of clinical trials, we are not trying to prove or quantify the superiority of one treatment over another but rather to evaluate the clinical utility of a precision medicine strategy assigning treatments based on genomic features of patients. This is therefore closer to the well-studied biomarker-based designs for clinical trials [Freidlin et al., 2010]. In a way, it is a matter of extending these unidimensional

biomarker-based designs to multidimensional strategies that allow a choice between quite a number of different treatments. The potentially large number of treatments thus prompts us to draw more inspiration from scalable biomarker-strategy designs than biomarker-stratified designs [Freidlin et al., 2010]. We can draw a methodological parallel with some trials like the Tumor Chemosensitivity Assay Ovarian Cancer study in which a biochemical assay guides the choice of preferred chemotherapy for patients in a panel of twelve different treatments [Cree et al., 2007]. More recently, some clinical trials have been proposed that include precision medicine strategies, particularly in oncology [Von Hoff et al., 2010, Le Tourneau et al., 2015, Flaherty et al., 2020].

On the basis of these clinical examples, we propose three different target trials and their corresponding causal estimates, the clinical relevance of which may vary according to medical contexts. Each target trial contains a **precision-medicine directed arm** in which patients are treated in accordance with the precision medicine algorithm recommendations but they are differentiated from each other by **alternative control arms** (Figure 9.4). Causal effects will be estimated solely on patients eligible for the assignment of a personalized treatment, *i.e.*, those for whom the treatment algorithm is able to recommend a drug.

#### 9.2.2.1 First causal effect ($CE_1$): comparison with a single standard

The first possible target trial is to compare the precision medicine arm with a control arm in which **all patients have been treated with the same single treatment**. This could classically be the current standard of care applied to all patients (*e.g.*, chemotherapy cancer treatment).

#### 9.2.2.2 Second causal effect ($CE_2$): comparison with physician's assignment of drugs

Then, in order to propose a more comprehensive clinical assessment, we propose a second causal effect, **comparing the PM arm with the current clinical practice**, *i.e.*, the assignment of the same targeted treatments by physicians in the absence of the algorithm. This implicitly means comparing two PM strategies: the one derived from the algorithm and the one that corresponds to current physician's knowledge. Unlike the former, the latter may not be perfectly deterministic depending on the heterogeneity of medical knowledge or practices. This way of defining $CE_2$ by focusing

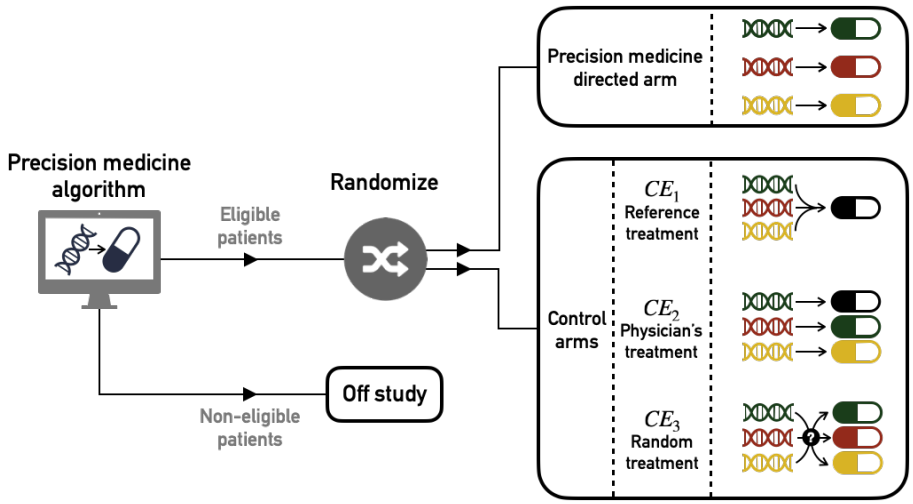



Figure 9.4: **Target trials to estimate causal effect of precision medicine (PM) algorithm versus different controls.** Patients are first screened according to their eligibility for the algorithm: based on their genomic characteristics patients are recommended a specific treatment (eligible) or not (no eligible). Then eligible patients are randomized and assigned either to PM-directed arm or to one of the alternative control arms ($CE_1$, $CE_2$ or $CE_3$)

on the doctor's assignment of the same treatments stems from our question of interest: to quantify the relevance of the algorithm itself. Another possibility would have been to compare the precision medicine arm with the doctor's treatments, allowing him to use treatments other than those of the PM arm, such as the gold-standard one described in $CE_1$. But the differences between the arms could then be biased by the use of treatments with different overall efficacy, changing the focus of the question. We will therefore stick to the first definition, which is more focused on the relevance of the algorithm.

#### 9.2.2.3 Third causal effect ($CE_3$): comparison with random assignment of drugs

Finally, we define the $CE_3$ effect **comparing the PM arm with a control arm using exactly the same pool of treatments assigned randomly**. In this case, we measure the ability of the PM algorithm to assign treatments effectively based on genomic features of patients. This comparison has already been considered in the context of biomarker-based clinical trials

[Sargent et al., 2005]. Although this comparison with random assignment is methodologically relevant, it may not make sense from a clinical point of view if the common clinical practice already contains strong indications (or contraindications) for some patient-treatment associations.

## 9.3 Causal inference methods and precision medicine

### 9.3.1 A treatment with multiple versions

The statement of the potential outcomes framework implicitly implies the uniqueness of the versions of the treatment [Rubin, 1980] or at least the treatment variation irrelevance [VanderWeele, 2009]. **In the precision medicine case, the multiplicity of treatment versions is inherent**: a given treatment status may encompass several drugs since a patient may be associated with several molecular agents based on his or her genomic characteristics. $A$ can be seen as a compound treatment [Hernán and VanderWeele, 2011] or a treatment with multiple versions [VanderWeele and Hernan, 2013].

Therefore, we define a variable $K_j$ denoting the version of treatment administered to individual $j$. If $A_j = a$ is the arm to which the patient is assigned, $K_j^a$ is the molecule received, the version of treatment $A = a$ (*e.g.*, a specific anti-cancer drug) and $K_j^a \in \mathcal{K}^a$, the set of versions of treatment $A = a$. In our precision medicine problem, $A = 0$ will denote control patients and $A = 1$ the patients treated with an anti-cancer drug of the precision medicine pool. $\mathcal{K}^1 = \{k_1^1, ..., k_P^1\}$ is the set of $P$ possible targeted treatments for $A = 1$ patients. For the sake of simplicity we will assume that there is only one treatment version for $A = 0$ controls, $\mathcal{K}^0 = \{k^0\}$. We also need to define other counterfactual variables like $K_j^a(a)$, the counterfactual version of treatment $A = a$ if the subject had been given the treatment level $a$. Thus, we finally write the counterfactual outcome as $Y_j(a, k^a)$ for individual $j$ when treatment $A$ has been set to $a$, using $k^a$ as the version of treatment $a$, with $k^a \in \mathcal{K}^a$. Causal relations between variables $C$, $A$, $K$ and $Y$ are depicted in the causal diagram in Figure 9.5. It should be noted that $A$ has no direct influence on $Y$, its only effect is entirely mediated by $K$, which is the real treatment in the pharmacological sense.

In this context, we can also define the assignment of a version of treatment for patients eligible for precision medicine algorithm. It is important

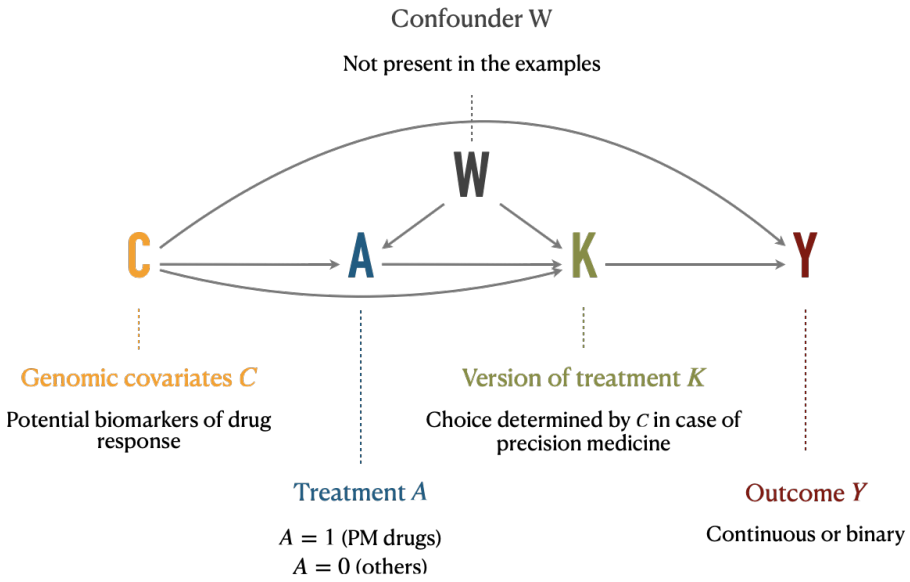



Figure 9.5: **Causal diagram illustrating relations between variables under multiple versions of treatment.** Treatment $A$, version of treatment $K$, outcome $Y$, and confounding variables $C$ and $W$ are placed in a causal digram, along with their interpretation in the precision medicine application.

to note that not all patients are necessarily eligible for the precision medicine strategy. Indeed, the treatment assignment algorithm relies on targetable alterations to establish its recommendations. In the absence of these, no recommendation can be offered to the patient. We denote $\mathcal{C}^{PM}$ the set of eligible patient profiles and consequently define the **drug assignment algorithm as the function $r$ which associates to each $C$ a precision medicine treatment version** $K$ such as:

$$\forall j \in [[1, N]], \text{ if } C_j \in \mathcal{C}^{PM}, r(C_j) \in \mathcal{K}^1$$

### 9.3.2 Causal inference with multiple versions

Consequently, the multiplicity of versions prevents direct application of the framework as described in section 8.2.2. The **theoretical framework has however been extended to causal inference under multiple versions of treatment** and some identifiability conditions and properties have been studied, especially in the seminal article by VanderWeele and

Hernan [2013]. One of the first required adaptation to identify some causal effects is to distinguish between confounders $C$ and $W$ (Figure 9.5). $W$ indicates a collection of covariates that may be causes of treatment $A$ or version of treatment $K$ but are not direct causes of $Y$. These covariates are of special interest for causal effects identification under multiple versions of treatment. $C$ indicates all other covariates. In our precision medicine settings, the genomic features of patients may define the eligibility to precision medicine and therefore affect $A$. They may also be used to define the version of treatment $K$. And finally they can influence the response to treatment $Y$. Thus, the genomic features of patients are a typical example of type $C$ confounders. All causal relationships are summarized in Figure 9.5. Please note that no $W$ variable is present in the applications provided later because all the covariates considered in this situation were likely to influence $A$, $K$ and $Y$ and therefore belonged rather to the covariates of type $C$. However all subsequent formulas and definitions have been derived taking into account $W$.

We summarize here some general observations from VanderWeele and Hernan [2013] regarding the extension of the framework to multiple versions before discussing specific estimates of interest of our precision medicine settings in the next section. These two sections will be based exclusively on the method called **standardization or parametric g-formula** described in section 8.2.2.1. The adaptation of other methods to precision medicine will be discussed more briefly in section 9.3.4. First of all, the identifiability conditions have to be adapted. The *consistency* assumption for instance is extended to $K$:

$$\text{if } A_j = a, \text{then } K_j^a(a) = K_j^a$$

.

Then, the *conditional exchangeability* or no-unmeasured confounding assumptions, may be stated in two different ways, either without or with versions of treatment:

$$Y(a) \perp\!\!\!\perp A|(C, W) \tag{9.1}$$

$$Y(a, k^a) \perp\!\!\!\perp \{A, K\}|C \tag{9.2}$$

If equation (9.1) holds, we can derive a new version of the standardised estimator with multiple versions of treatment [VanderWeele and Hernan, 2013]:

$$E[Y(a)] = E[Y(a, K^a(a))] = \sum_{c,w} E[Y|A = a, C = c, W = w] \times P[c, w] \tag{9.3}$$

Specifically, it should be noted that we need to add $W$ in the set of covariates that must be taken into account in standardization, and we need *positivity* to hold for $C$ and $W$, *i.e.*, $0 < P[A = a|C = c, W = w] < 1$. Detailed proof of equation (9.3) is provided in appendix C.2.1. Equation (9.3) paves the way to overall treatment effect assessment since $E[Y(1, K^1(1))] - E[Y(0, K^0(0))]$ would estimate the effect of treatment $A = 1$ compared to $A = 0$ with current versions of treatment.

Conversely, estimating a treatment effect for a given unique version of treatment $E[Y(a, k^a)]$ would require to check the exchangeability with regard to versions $K$ and therefore to hold equation (9.2) true [VanderWeele and Hernan, 2013]:

$$E[Y(a, k^a)] = \sum_{c} E[Y|A = a, K^a = k^a, C = c] \times P[c] \tag{9.4}$$

Similarly, we can define $G^a$ a random variable for versions of treatment with conditional distribution $P[G^a = k^a|C = c] = g^{k^a,c}$ and assuming the equation (9.2) to be true we can derive the following formula and its formal proof in appendix C.2.2:

$$E[Y(a, G^a)] = \sum_{c,k^a} E[Y|A = a, K^a = k^a, C = c] \times g^{k^a,c} \times P[c] \tag{9.5}$$

In this case, to allow estimation of the right-hand side of the equation, positivity will be defined as $0 < P[A = a, K^a = k^a|C] < 1$.

### 9.3.3 Application to precision medicine

In the context of the potential outcomes framework extended to treatments with multiple versions, it is therefore possible to apply equations (9.3) and

(9.5) in order to define and estimate the precision medicine causal effects previously described in section 9.2.2.

$A = 0$ corresponds to control patients with $\mathcal{K}^0 = \{k^0\}$ and $A = 1$ to patients treated with a targeted treatment. It is important to notice that from this point on we systematically restrict ourselves to patients eligible for the precision medicine algorithm, *i.e.*, to individuals $j$ such as $C_j \in \mathcal{C}^{PM}$.

#### 9.3.3.1 $\text{CE}_1$ estimation

$\text{CE}_1$ is a comparison between the precision medicine arm and a single version control arm:

$$\text{CE}_1 = E[Y(1, r(C)] - E[Y(0, k^0)] \tag{9.6}$$

In details, $E[Y(1, r(C)]$ can be derived from equation (9.5) in the case where $g^{k^a,\ c} = 1$ if $k^a = r(c)$ and $g^{k^a,\ c} = 0$ otherwise:

$$E[Y(1, r(C)] = \sum_c E[Y|A = 1, K^1 = r(c), C = c] \times P[c]$$

Then, $E[Y(0, k^0)]$ and $E[Y(1, k^1_{ref})]$ can be derived from equation (9.4):

$$E[Y(0, k^0)] = \sum_c E[Y|A = 0, C = c] \times P[c]$$

Alternatively, if one wants to use as control only one of the treatments used in the PM arm the previous estimate could be replaced by the following one:

$$E[Y(1, k^1_{ref})] = \sum_c E[Y|A = 1, K^1 = k^1_{ref}, C = c] \times P[c]$$

It should be noted that $\text{CE}_1$, like $\text{CE}_2$ and $\text{CE}_3$ presented later, depends on the PM algorithm of interest $r$. $\text{CE}_i$ could therefore also be written $\text{CE}_1(r)$.

#### 9.3.3.2 $CE_2$ estimation

Then, $CE_2$ is written using $K^1(1)$ the PM targeted treatment that would have been assigned to the patient by the physician if the patient had been allocated in arm $A = 1$ with PM targeted treatments:

$$CE_2 = E[Y(1, r(C)] - E[Y(1, K^1(1))] \tag{9.7}$$

$E[Y(1, K^1(1))]$ is derived from equation (9.3):

$$E[Y(1, K^1(1))] = \sum_{c,w} E[Y|A = 1, C = c, W = w] \times P[c, w]$$

#### 9.3.3.3 $CE_3$ estimation

Defining $G^1$ as the random distribution of versions of treatment $k^1 \in \mathcal{K}^1$, $CE_3$ expresses as:

$$CE_3 = E[Y(1, r(C)] - E[Y(1, G^1)] \text{ with } P[G^1 = k_i^1] = \frac{1}{|\mathcal{K}_{PM}^1|}, \tag{9.8}$$

$|.|$ denoting the cardinality of the set. In this formula, $E[Y(1, G^1)]$ can be derived from equation (9.5):

$$E[Y(1, G^1)] = \frac{1}{|\mathcal{K}_{PM}^1|} \times \sum_{c,k_i^1} E[Y|A = 1, K^1 = k_i^1, C = c] \times P[c]$$

### 9.3.4 Alternative estimation methods

For the sake of simplicity and brevity, we only detailed the standardization in previous sections. However, other popular candidate methods can be used. Estimators based on the **inverse probability weighting (IPW)** and **targeted maximum likelihood estimation (TMLE)** will also be computed in the following sections. IPW has the particularity of not trying to model the outcome but rather the process of assigning treatments. Its theoretical bases have been described in section 8.2.2.2 and the details of its adaptation to multiple versions of treatment is provided in appendix C.2.3.

Table 9.1: **Intercepts and linear coefficients in the linear models specified to simulate data**

| Response variable | Intercept | Lin. coeff. $Y \sim C_1$ | Lin. coeff. $Y \sim C_2$ |
|---|---|---|---|
| $Y(0, k^0)$ | 0 | 0 | 15 |
| $Y(1, k_1^1)$ | -25 | -15 | 10 |
| $Y(1, k_2^1)$ | 0 | 0 | -20 |

The TMLE methods are of a different nature [Van der Laan and Rose, 2011]. They combine an outcome model and a treatment model in order to obtain a doubly robust estimate, *i.e.*, an estimate that is robust to a possible misspecification of either model. Moreover, the estimation is done in several steps in order to optimize the equilibrium bias-variance, not for the overall distribution of the data but specifically for the causal effect of interest. These methods also have the particularity of being very often used with machine learning algorithms to fit the outcome or treatment models, instead of the parametric models classically used in standardization and IPW methods. A more detailed description of TMLE properties and the choices that have been made to adapt it to the problem of precision medicine are available in appendix C.2.4.

### 9.3.5 Code

The methods detailed above have been implemented in R and applied to simulate data and PDX data. The code is provided in the form of R notebooks (simulated data and PDX data) as well as in the form of an RShiny interactive application (simulated data only). All of these files are available in the dedicated GitHub repository.

## 9.4 Application to simulated data

The proposed methods are first tested on simulated data in order to check the performance of the estimators in finite sample sizes.

### 9.4.1 General settings

Using the R package *lava* based on latent variable models, we simulate a super-population of 10,000 patients with variables $C$, $A$, $K$ and $Y$ as in Figure 9.5. We first define two independent binary variables $C_1$ and $C_2$, rep-

resenting mutational status of genomic covariates, with a mutation prevalence of 40%. By analogy with the PDX data, $Y$ represents the evolution of tumor volume and a low value (*a fortiori* negative) corresponds to a better response. $Y$ is therefore defined as a continuous gaussian variable. For each counterfactual variable of response $Y(a, k^a)$, we specify the intercept and the linear regression coefficients regarding influence of $C_i$ as described in Table 9.1. Lower intercepts correspond to better responses/more efficient drugs. Similarly, a negative regression coefficient between $Y(a, k_i^a)$ and $C_j$ means that the gene $C_j$ improves the response to $k_i^a$. So all in all, $k_1^1$ has the best basal response (lowest intercept). $C_1$ (resp. $C_2$) improves the response to $k_1^1$ (resp. $k_2^1$). The treatment algorithm of precision medicine is in line with these settings since patients mutated in $C_1$ (regardless their $C_2$ status) are recommended to take $k_1^1$ and patients mutated for $C_2$ only are recommended to take $k_2^1$. Patients without mutations are not eligible for precision medicine and not taken into account in the computations. Since $k_1^1$ has the best basal response we assume it is assigned with greater probability by the physician and implement the following distribution of observed treatments:

$$P[K = k_1^1] = 0.5 \text{ and } P[K = k_2^1] = P[K = k^0] = 0.25$$

A super-population of 10,000 patients is then generated. 1,000 cohorts of 200 patients are sampled without replacement within this super-population which, with the prevalences defined for the mutations, corresponds to an effective sample size of about 130 patients eligible for the PM algortithm. The causal effects $\text{CE}_1$, $\text{CE}_2$ and $\text{CE}_3$ are computed based on different methods on the sub-cohort eligible for precision medicine:

- **True effects**, using all simulated counterfactuals for all patients
- **Naive effects**, using observed outcomes only for both PM and control arms
- **Corrected effects**: using observed outcome and standardized estimators (Std), inverse probability weighting (IPW) and targeted maximum likelihood estimators (TMLE).

### 9.4.2 Simulation results

First, the distribution of data in the super-population of 10,000 patients can be observed in Figure 9.6A, illustrating the different relations and differences described above. In particular, $Y(1, k_1^1)$ (resp. $Y(1, k_2^1)$) is lower

for $C_1$-mutated (resp. $C_2$-mutated) patients. It can also be seen that the response to precision medicine ($Y(1, r(C))$) differs according to the groups: patients mutated for $C_1$ only have the best response, followed by patients mutated for both $C_1$ and $C_2$ and patients mutated for $C_2$ only. There is therefore a heterogeneity of responses to PM which encourages to take into account the groups of patients and their PM versions. The right side of Figure 9.6A shows the deterministic assignment of the recommended PM treatment ($r(C)$) to each patient profile and the unbalanced distribution of observed treatments ($K$) with a predominance of $k_1^1$.

In the first target trial, true $CE_1$ estimates in the sampled cohorts are distributed around -40 (Figure 9.6B), confirming the **superiority of the PM arm over the control arm** as defined in the simulation parameters. Not all methods of estimating the causal effect perform equally well. The so-called naive estimate and the one based on IPW show a net bias. The over-representation of the most advantaged patients by PM tends to cause these methods to overestimate the benefit of PM, as can also be seen in the deviation plots. The same trends are observed for $CE_2$ and $CE_3$ (Figure 9.6C and D) where the differences are even more drastic. The **mean absolute error of the naive method is thus divided by more than 2 when using standardized estimates or the TMLE**.

In order to further dissect the influence of simulation parameters on estimation performances, a slightly different simulation scenario with equal probabilities of observed treatments has been studied:

$$P[K = k^0] = P[K = k_1^1] = P[K = k_2^1] = \frac{1}{3}$$

In this case, the random and balanced assignment of the observed treatments logically removes the systematic biases of the naive method by providing them with more randomized data. However, the corrections made by the proposed methods of causal inference, and in particular standardization and TMLE, still **reduce the variances in the estimates** due to the heterogeneity of the effects of precision medicine as a function of molecular profiles. Randomly, some sampled cohorts are indeed found with an association between $C$ and observed $K$, thus generating a confounding effect that the causal methods partially correct.

The simulated data allow us to imagine an almost infinite number of scenarios depending on the number of biomarkers taken into account in the algorithm, the number of different treatments, the dependencies of their responses or the distribution of treatments observed. In order to allow

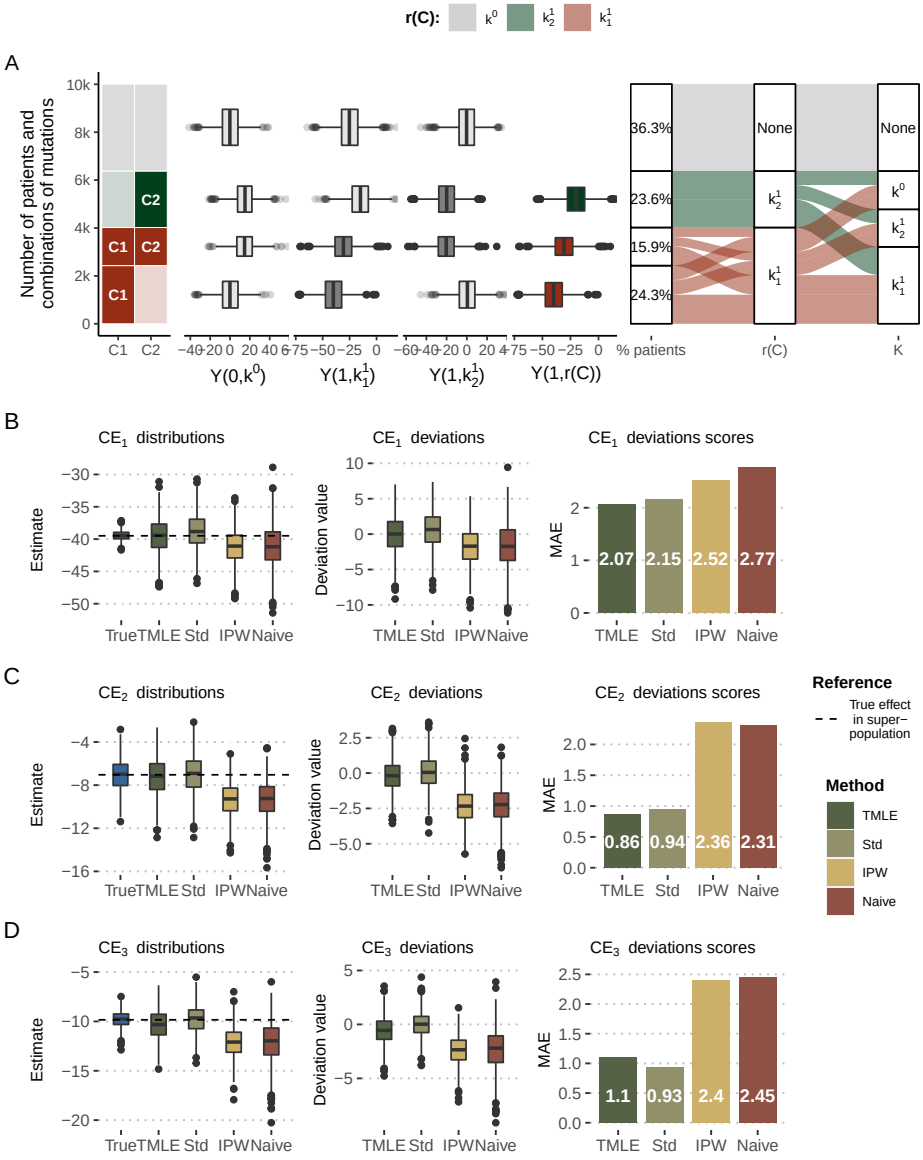



Figure 9.6: **Causal effects of precision medicine strategy with simulated data.** (A) Main variables and relations in the simulated superpopulation. From left to right: categories of patient based on their mutations; responses to $k^0$, $k_1^1$, $k_2^1$ and precision medicine $K = r(C)$; repartition of patients regarding their precision medicine drug and their assigned treatment in observed data. (B) Distribution and deviation of $CE_1$ estimates based on different methods, deviation scores being computed based on mean absolute error (MAE). (C) Same for $CE_2$. (D) Same for $CE_3$.

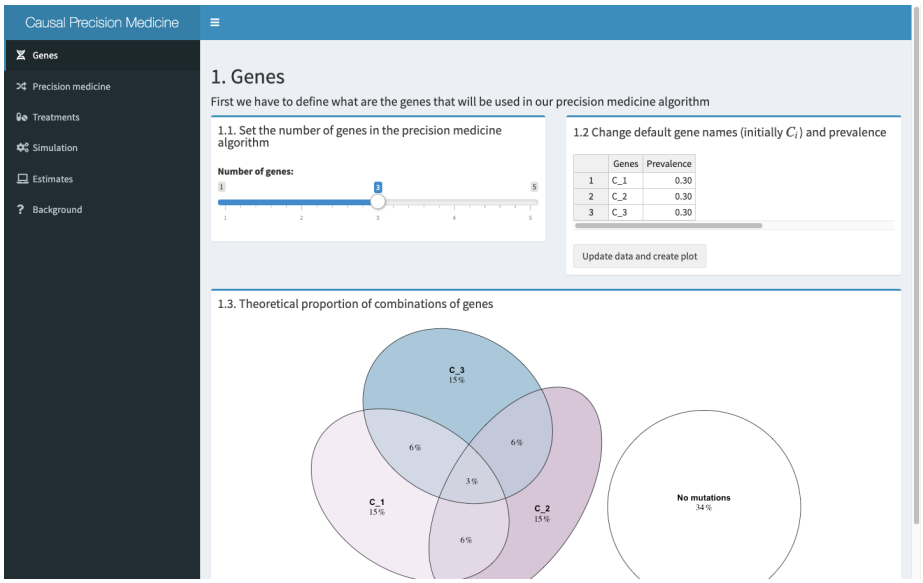



Figure 9.7: **RShiny interactive application to investigate various simulation scenarios of precision medicine evaluation.** It is possible to run the application locally with the source R file or online with the version hosted on the shinyapps.io server

easy exploration of these scenarios without having to master the underlying R code, an **interactive RShiny application has been developped**. It can be accessed by locally running the R source file or by using the online version embedded in Figure 9.7. Readers with the ability to run the application locally are encouraged to favor this option because the hosting of the online application is limited to a maximum amount of time per month. The application allows certain additional analyses not presented in this manuscript, in particular the linking of biases observed in the sampled cohorts according to their composition (prevalence of mutations, treatments, etc.). It is thus possible to trace the origin of the biases.

## 9.5 Application to PDX

The method is then applied to public data from patient-derived xenografts [Gao et al., 2015], described in section 9.1.1 and appendix A.2. One of the major interests of this type of data in the context of this chapter is to provide access to treatment response values otherwise considered as hypothetical (or counterfactual). It is indeed possible to have the response of the same tumor (or more precisely of distinct samples from the same tumor) to different

treatments, thus representing **proxies for counterfactual variables**, as described in Figure 9.1. Availability of these data provides a unique ground truth to assess the validity of proposed causal estimates in a pre-clinical context.

Based on the analysis accompanying the published data [Gao et al., 2015], some biomarkers of treatment response have been selected and resulted in **an example of treatment algorithm**: binimetinib (MEK inhibitor) is recommended to KRAS/BRAF mutated tumors, and BYL719 (alpha-specific PI3K inhibitor, also known as Alpelisib) to PIK3CA mutated tumors. PTEN is also included as a covariate because of its detrimental impact on the response to these two treatments. LEE011 drug (a cell cycle inhibitor also known as Ribociclib) is chosen as the reference drug treatment ($k^0$). Among the sequenced tumors, 88 are eligible for this precision medicine algorithm (*i.e.*, mutated for BRAF, KRAS or PIK3CA) and have been tested for all 3 drugs of interest, thus ensuring the availability of all corresponding responses. The following analyses will focus exclusively on this sub-cohort for which a descriptive analysis is provided in Figure 9.8A. As expected BRAF/KRAS-mutated tumors have a better response to binimetinib and PIK3CA-mutated tumors have a better response to BYL719 (Figure 9.8B). In addition, it can be noted that these biomarkers have deleterious cross-effects.

The analysis settings are similar to the ones used for simulated data. 1,000 different cohorts of 70 tumors (out of 88) are sampled without replacement assuming each time that only the response to one of the treatments is known for each tumor, reproducing the classical clinical situation. The **distribution of the observed treatments was defined randomly**:

$$P[K = k^0] = P[K = k_1^1] = P[K = k_2^1] = \frac{1}{3}$$

It should be noted that, contrary to analyses based on simulated data, all the statistical models used for standardization (outcome model), for the IPW (treatment model) and for the TMLE are no longer generalized linear models (GLM) but **random forests (RF)**. Indeed, it was observed that the performance of GLM-based methods was lower than that of the naive method, supporting the importance of relevant model specification consistent with real data. The RF algorithms then allow to limit misspecification due to the largely non-linear nature of the data. Random forests were chosen for their speed and versatility, especially in view of their ability to handle multinomial classification as well.

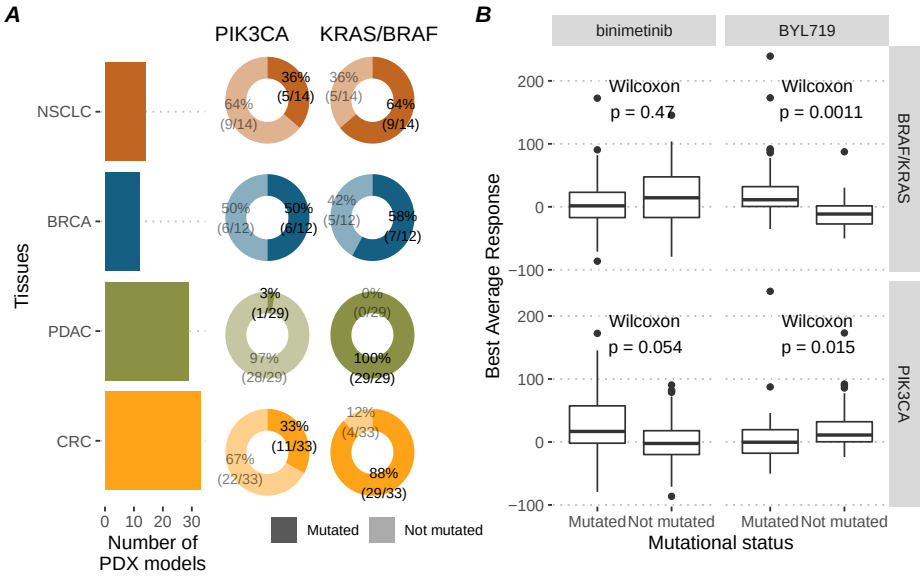



Figure 9.8: **Description of the 88 PDX models cohort.** (A) Tissue of origin and prevalences of the drug biomarkers. (B) Drug response to precision medicine targeted treatments in the 88 PDX models cohort depending on the mutational status of biomarkers

The results of estimations are then presented in Figure 9.9. In the presence of randomly assigned and balanced observed treatments, none of the methods (including the naive one) has significant systematic bias. On the other hand, **more sophisticated methods, and in particular TMLE, allow to reduce the gap between estimates and true values**, as visible on the mean absolute errors in Figure 9.9 right column. An additional analysis using the binary version of outcome $Y$ is presented in Béal and Latouche [2020] with similar trends and conclusions: standardized, IPW and TMLE estimates are closer than naive methods to the true values from PDX. It supports the validity of the extension of the method to binary outcomes. In the same way as before with the simulated data, it would be possible to study the impact of non-random assignment of the observed treatments, which could systematically bias the results of the naive methods.

## 9.6 Limitations and perspectives

In synthesis, this work proposes a conceptual framework for evaluating a precision medicine algorithm, taking advantage of data already generated using adapted causal inference tools. However, in a clinical context, these

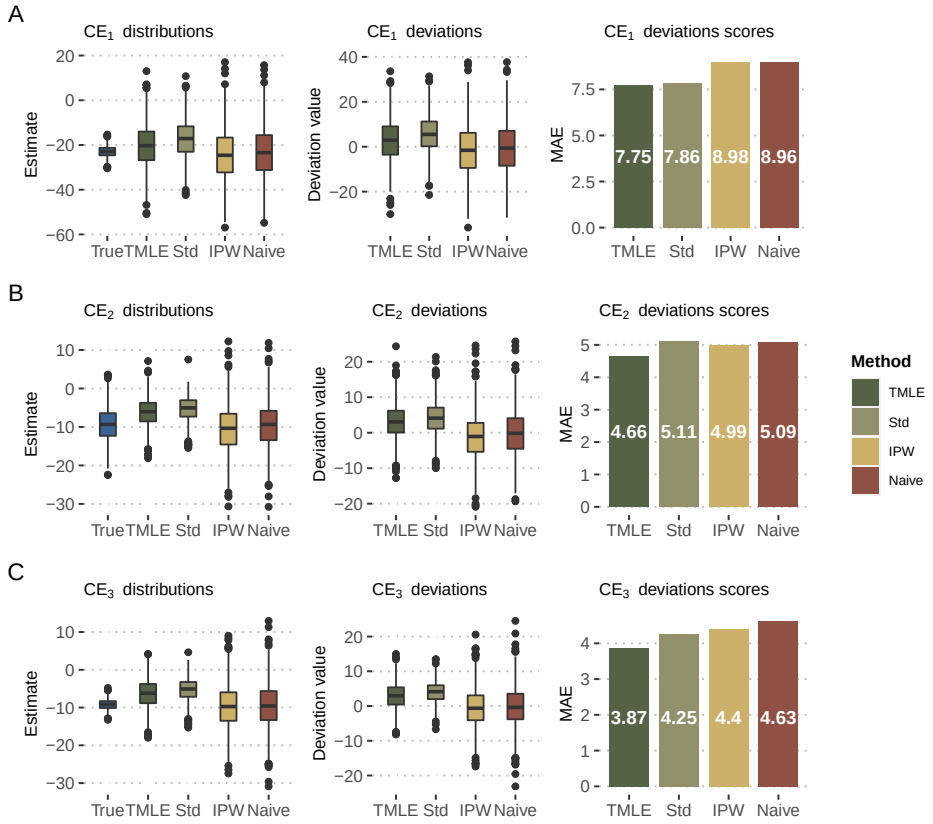



Figure 9.9: **Causal estimates with PDX data.** Distribution and deviation of $CE_1$ (A), $CE_2$ (B) and $CE_3$ (C) estimates based on different methods as in Figure 9.6B.

data were not generated in a purely observational manner. Patients were cared for and treated by physicians who probably took into account some of their characteristics. However, the reasoning, formalized or not, behind the physicians' decisions does not correspond to that which a new investigator might want to test. In the eyes of this new investigator, the data can therefore be considered as observational in that they do not correspond to the randomization he would have liked to have carried out. The possibility for this new investigator to estimate the impact of his PM algorithm using the proposed estimators depends, however, on the consistency, exchangeability and positivity hypotheses.

The hypothesis of consistency has been made more plausible by taking into account the treatment versions, which makes it possible to explicit the heterogeneity of the molecules administered. Exchangeability remains ques-

tionable. The simulations and calculations described above underline the importance of taking into account at least the genomic covariates used in the processing algorithm. The inclusion of additional covariates is likely to be necessary in many real-world applications. **Positivity, on the other hand, can be violated in a much more obvious way in certain situations**. Thus, equation (9.5) requires positivity to be extended to versions of treatment: $0 < P[A = a, K^a = k^a|C] < 1$. If the assignment of the observed treatments was done on a deterministic basis with respect to the variables used by the treatment algorithm, each patient's molecular profile will have been treated with a single drug, thus preventing any subsequent causal inference within the defined framework. The eventual use, by the boards of physicians in charge of assigning the observed treatments, of variables different from those used by the algorithm could then make it possible to verify the positive condition. But these variables would represent unmeasured confounding factors. It is therefore **essential to have an in-depth knowledge of the rationales at work in the assignment of the observed treatments**.

We developed a user-friendly application that extends the scope of the simulations and makes possible to study and quantify the impact of different situations, including possible (quasi-)violations of positivity or unmeasured confounding. It is thus a **tool for empirically framing cases where this causal inference is reasonable or not**. The analysis of the PDX data provides an illustration and proof of feasibility for these methods on pre-clinical data, closer to the human clinical data generally of interest. Beyond feasibility, this implementation leads to some remarks. Firstly, the improvement of causal inference methods compared to naive estimation of PM effects is conditioned in this case to the use of flexible and non-linear learning algorithms. This underlines the **importance of a proper specification of the outcome and treatment models** whose imperfection, especially when trained on small samples, could explain the modesty of the results compared to the simulated data. The particular nature of the PDX data design used should also be kept in mind: each tumor is tested only once for each drug, which may lead to greater variability of results due to tumor heterogeneity [Gao et al., 2015]. Some studies, with smaller numbers of tumors and treatments, propose to form groups of several mice for each treatment-drug combination [Hidalgo et al., 2014]. The use of these mean effects could contribute to more accurate data. In spite of these limitations, which may diminish their ability to provide values with counterfactual interpretation, **PDX data are thus a dataset of interest for studying and validating methods of causal inferences about treatment re-**

**sponse**. It can also be noted that the very nature of these data, due to the multiplicity of drugs tested for each tumor, can provide a framework in which the constraints of positivity are singularly alleviated. Even if all drugs were not tested on all patients, considering each tumor-drug combination as a different unit increases the coverage of the data. It is then necessary to take into account the clustered nature of the data, each tumor being present several times.

Finally, beyond the pre-clinical data presented here, the theoretical framework developed in this chapter should be more directly applicable to data from clinical trials if these data do not violate the requirements of positivity. If it is necessary to consider several trials, the heterogeneity of practices must be taken into account. The use of different drug lists from one trial to another or from one medical centre to another could also provide an example of confounding factor $W$, included in the theoretical framework presented here but not used in applications.

# Conclusion

The aim of this thesis was to trace **a path to link the biological knowledge of cancer to the clinical impact through mechanistic models**. Among the many possible orientations, it was chosen to take the opposite side of the data-intensive machine learning methods. The main approach proposed uses a qualitative logical formalism and **integrates the data by interpreting them** rather than by optimizing the parameters with respect to a particular objective. As a result, the resulting personalized mechanistic models have proven to be **more of an interpretive than a predictive tool**. Their versatility and low data requirements nevertheless allow them to be applied to a wide range of questions, particularly concerning the response to treatments that their mechanistic nature facilitates. This seemingly limitless versatility can, however, prove to be a trap because, while all kinds of applications are theoretically possible, the need to rely on detailed biological knowledge and appropriate data limits its scope.

In the case of mechanistic molecular signaling models, this interpretive nature of the models is confirmed by statistical analyses. The main value of these models is to provide an understandable framework for extracting relevant biological information in the context of current biological knowledge. The **ability of these models to detect emerging non-linear information is also proven, but is rarer and of relatively smaller magnitude**. Given the influx of biological knowledge and data, computational models of cancer, with various formalisms, are nevertheless multiplying, particularly with medical aims. In the context of cancer, their use to recommend personalised treatment for each patient is a possible horizon. The evaluation of these models could then become increasingly acute. This thesis proposes the **adaptation of causal inference methods in order to simulate their evaluation in clinical trials** and to come as close as possible to medical evaluation standards.

In a word, cancer models still have a bright future ahead of them. Mech-

anistic models will continue to be attractive because of their ability not only to predict but, more importantly, to explain. However, the transparency of their mechanisms should not prevent them from being rigorously evaluated statistically. It is not enough for them to explain, they must also be well understood.

APPENDIX A

# About datasets

## A.1 Cell lines

Several analyses in previous chapters are based on data derived from cell lines. Among the different databases, the ones used in the thesis are briefly described below. Please refer to corresponding references for additional details.

### A.1.1 Omics profiles

The omics profiles of cancer cell lines have been downloaded from Cell Model Passports [van der Meer et al., 2019] containing genotypic and phenotypic information about more than 1,000 cell lines. Among the available data used in this thesis are the exome sequencing, copy number variations and RNA-sequencing.

### A.1.2 Drug screenings

Information about response to treatments is retrieved from Genomics of Drug Sensitivity in Cancer Database (GDSC, Yang et al. [2012]). In order to allow detailed analyses at the level of cancer types, we will restrict ourselves here to tissues represented by at least 20 cell lines and highlighted in dark grey in Figure A.1A. Most of the 663 cell lines in this subcohort have a complete profile with all omics data (mutations, CNA and expression) and

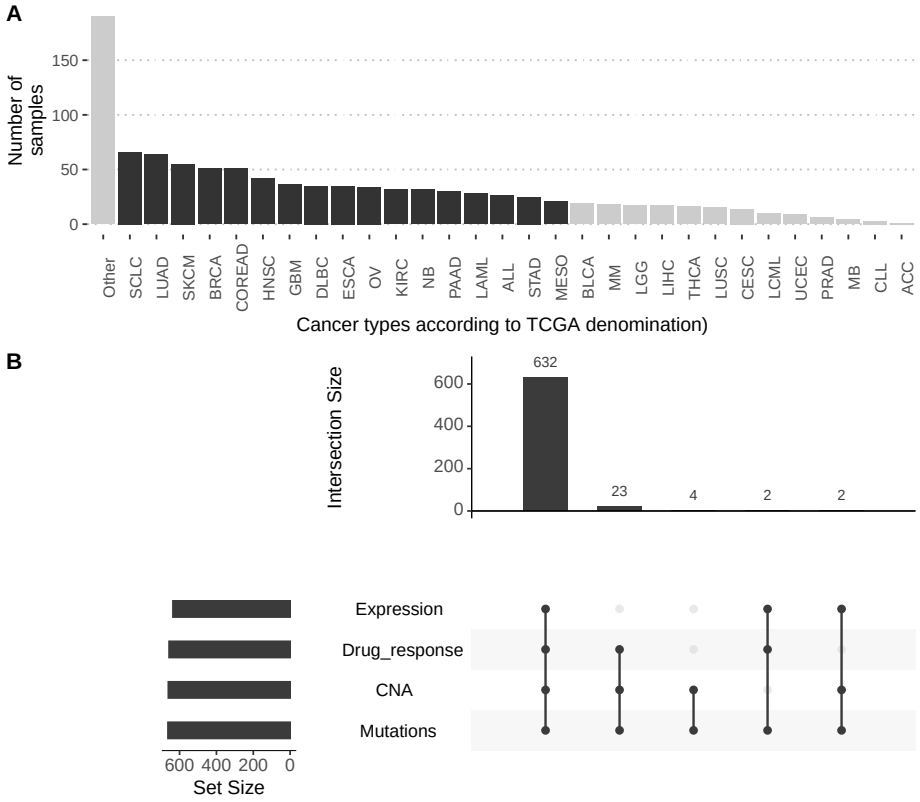



Figure A.1: **Distribution of cancer types and data types in GDSC-associated dataset.** (A) Distribution of cell lines per cancer types, highlighting the ones selected in this thesis with more than 20 cell lines. (B) Availibility of data for the 663 selected cell lines in 17 different cancer types.

drug responses. However, not all cell lines have necessarily been tested for all drugs.

The cell lines are treated with increasing concentration of drugs and the viability of the cell line relative to untreated control is measured. The dose-response relative viability curve is fitted and then used to compute the half maximal inhibitory concentration ($IC_{50}$) and the area under the dose-response curve (AUC) [Vis et al., 2016], both being represented in Figure A.2. Since the $IC_{50}$ values are often extrapolated outside the concentration range actually tested, we will focus on the AUC metric for all validation with drug screening data. AUC is a value between 0 and 1: values close to 1 mean that the relative viability has not been decreased, and lower values correspond to increased sensitivity to inhibitions. In cases where the ranges

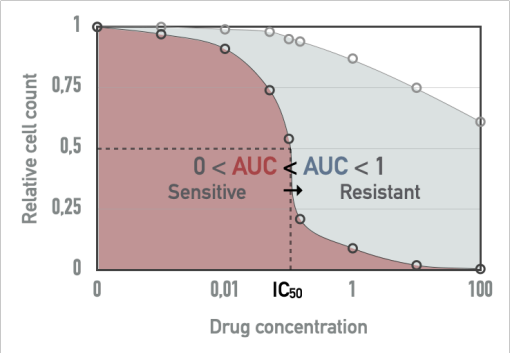



Figure A.2: **Drug screening metrics in cell lines.** Based on a tested drug concentration range, $IC_{50}$ and area under the dose-response curve (AUC) can be computed. For a given drug, red AUC corresponds to a more sensitive cell line than blue AUC.

of concentrations tested for different drugs vary, comparison of their AUC values does not have a simple and straightforward interpretation.

### A.1.3 CRISPR-Cas9 screening

On top the previous drug response characterization, some CRISPR-Cas9 screenings have been performed on cancer cell lines. Very basically, this involves using single-guide RNAs (sgRNAs) to direct the targeted inhibition of certain genes. Conceptually, screening is not very different from drug screening since it allows the sensitivity of cell lines to the inhibition of certain targets to be studied. However, this technology makes it possible to target many more different genes since it is based on RNA guide synthesis and not on the existence of drugs with an affinity for the target of interest. Schematically, sreening is therefore broader (thousands of genes), less biased (any gene can be targeted *a priori*) and more precise (much lower off-target effect).

Among the various databases available, the ones used in this thesis have been downloaded from Cell Model Passports and come from Sanger Institute [Behan et al., 2019] and Broad Institute [Meyers et al., 2017]. Both databases present CRISPR inhibition results for thousands of genes for a few hundred cell lines among those presented in the previous section. The Sanger dataset for instance includes 324 cell lines, and 238 in common with the subcohort previously described in the previous section and in Figure A.1.

Among the different metrics, the examples presented in this thesis will focus on scaled Bayesian factors to assess the effect of CRISPR targeting of genes. These scores are computed based on the fold change distribution of sgRNA [Hart and Moffat, 2016]. The highest values indicate that the targeted gene is essential to the cell fitness.

## A.2 Patient-derived xenografts

Another type of data exists, halfway between cell lines and patients, and that is patient-derived xenografts (PDX). Each patient tumour is divided into pieces later implanted in several immunodeficient cloned mice treated with different drugs, thus providing access to sensitivities to several different drugs for each tumour.

### A.2.1 Overview of PDX data from Gao et al. [2015]

The PDX dataset used in this thesis is the one published by Gao et al. [2015]. The original dataset contains 281 different tumours of origin (sometimes called PDX models, in the sense of a biological model) and 63 tested drugs, not all drugs having been tested for all tumours and some drugs have been tested with tissue-specific patterns (Figure A.3). 192 of these tumours have also been characterized for their mutations, copy-number alterations and mRNA. More detailed analyses of this dataset are available in the dedicated Github repository, in the file *Analysis_PDX.Rmd* and its corresponding HTML report.

### A.2.2 Drug response metrics

#### A.2.2.1 A continuous outcome

The first drug response metric used in this article is called *Best Average Response*. For each combination tumour/drug, the response is determined by comparing tumor volume change at time $t$, $V_t$ to tumor volume at time $t_0$, $V_{t_0}$. Several scores are computed:

$$\text{Tumour Volume Change } (\%) = \Delta Vol_t = 100\% \times \frac{V_t - V_{t_0}}{V_t}$$

$$\text{Best Response} = min(\Delta Vol_t), t > 10d$$

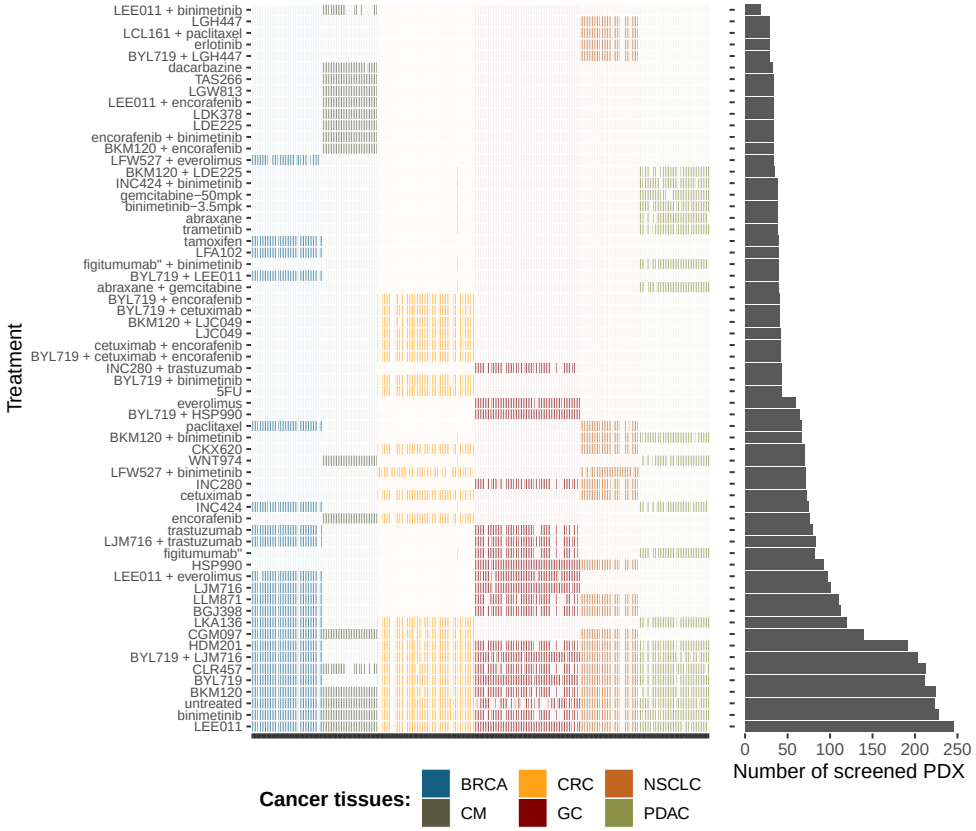



Figure A.3: **Comprehensive overview of tumours and drugs screened in PDX dataset from Gao et al. [2015].**

$$\text{Average Response}_t = mean(\Delta Vol_i, 0 \leq i \leq t)$$

$$\text{Best Average Response} = min(\text{Average Response}_t), t > 10d$$

We will mainly focus on *Best Average Response*. This metric "captures a combination of speed, strength and durability of response into a single value" [Gao et al., 2015]. Qualitatively, lower values correspond to more efficient drugs.

#### A.2.2.2 A binary outcome

Thresholds of *Best Response* and *Best Average Response* are also defined, inspired by RECIST criteria [Therasse et al., 2000], in order to classify response to treatment into 4 categories: Complete Response (CR), Partial Response (PR, Stable Disease (SD) and Progressive Disease (PD). We designed a binary response status by combining the response categories (CR, PR and SD) into a single "responder"" category (1), opposed to the "non-responders" progressive diseases (0).

## A.3 Patients

### A.3.1 METABRIC

METABRIC dataset is large breast cancer dataset with more than 2'000 patients [Pereira et al., 2016]. Mutations, CNA, expression (transcriptomics micro-array) and clinical data are available for a majority of patients (Figure A.4A), with 1'904 patients for whom all the data is available. One of the particular features of these data is to propose a very long clinical follow-up, over more than 10 years (Figure A.4B).

### A.3.2 TCGA: Breast cancer

Another reference database for breast cancer is the one from the TCGA consortium [TCGA et al., 2012]. The cohort is smaller than METABRIC and its clinical follow-up is more limited. In contrast, the omics data are more comprehensive and include RNA sequencing and relative quantification of proteins with RPPA technology (Figure A.5A).

### A.3.3 TCGA: Prostate cancer

Similarly, for prostate cancer, reference can be made to data from the TCGA study [Abeshouse et al., 2015], which has the same type of data but for a smaller number of patients than the breast cancer (Figure A.5B).

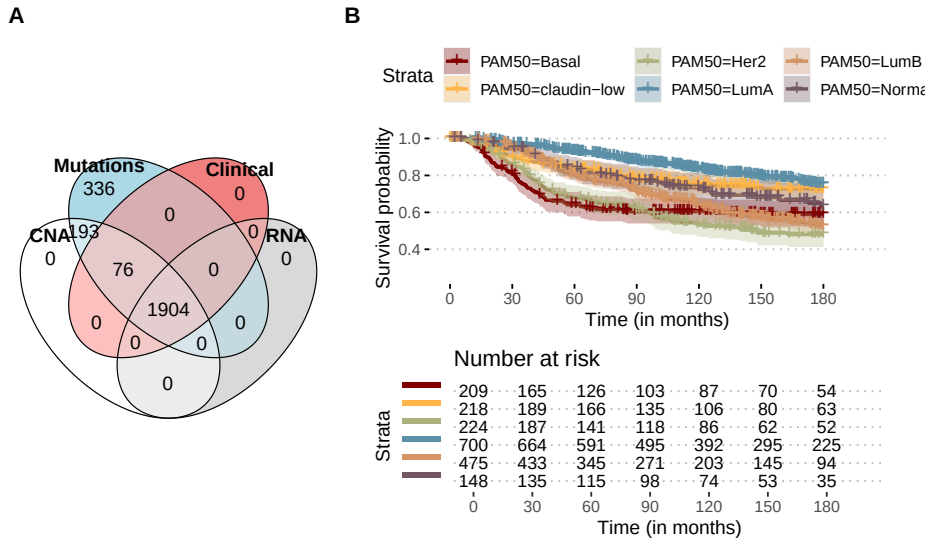



Figure A.4: **Available omics and survival in METABRIC Breast Cancer dataset**. (A) Number of patients for each omics type and their combinations, depicted as a Venn diagram. (B) Overall survival probability for all patients with clinical follow-up, stratified per breast cancer PAM50 subtype; administrative censoring at 180 months.

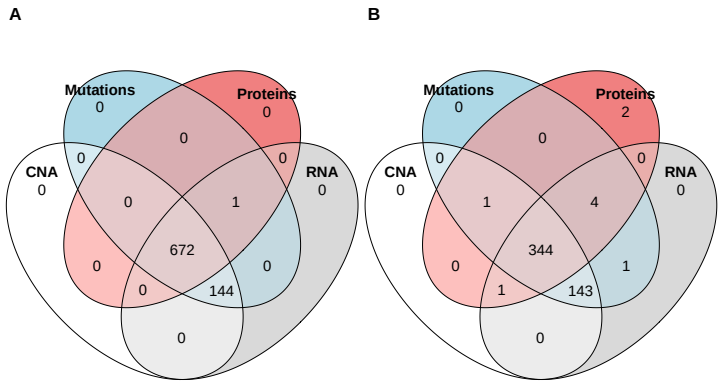



Figure A.5: **Available omics for TCGA Breast and Prostate cancer**. (A) Number of patients for each omics type and their combinations, depicted as a Venn diagram, in TCGA BRCA (Breast Invasive Carcinoma) study. (B) Same for the TCGA PRAD (Prostate Adenocarcinoma) study.

APPENDIX B

# About logical models

Several logical models of cancer are used in this thesis and some additional descriptive elements about them are given below.

## B.1 Generic logical model of cancer pathways

For this thesis, a published Boolean model from [Fumia and Martins, 2013] has first been used to illustrate our PROFILE methodology. This regulatory network summarizes several key players and pathways involved in cancer mechanisms such as RTKs, PI3K/AKT, WNT/$\beta$-catenin, TGF-$\beta$/Smads, Rb, HIF-1, p53 and ATM/ATR. An input node *Acidosis* has been added, along with an output node *Proliferation* used as a readout for the activity of any of the cyclins (*CyclinA*, *CyclinB*, *CyclinD* and *CyclinE*). This slightly extended model contains 98 nodes and 254 edges and its inputs are *Acidosis*, *Nutrients*, *Growth Factors* (GFs), *Hypoxia*, *TNFalpha*, *ROS*, *PTEN*, *p14ARF*, *GLI*, *FOXO*, *APC* and *MAX*. Its outputs are *Proliferation*, *Apoptosis*, *DNA_repair*, *DNA_damage*, *VEGF*, *Lactic_acid*, *GSH*, *GLUT1* and *COX412*.

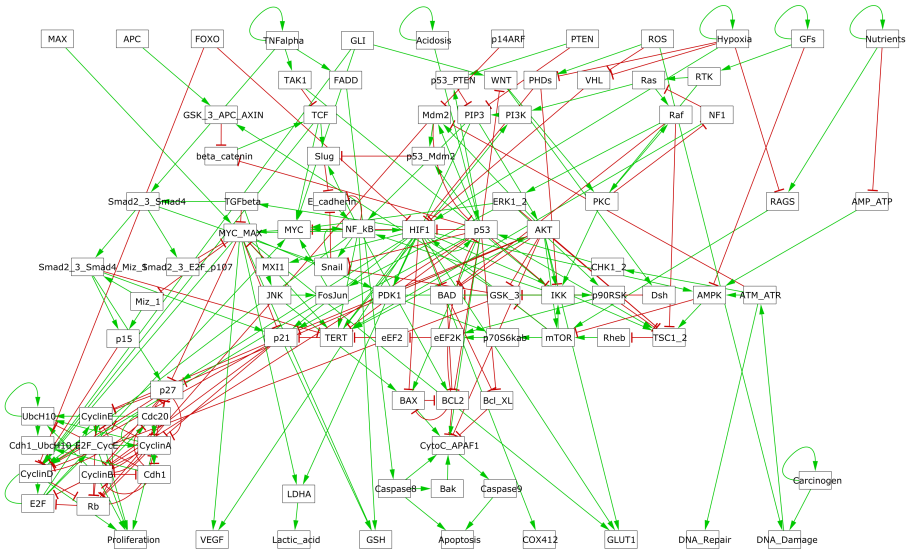



Figure B.1: **GINsim representation of the logical model described in Fumia and Martins [2013].**

## B.2 Extended logical model of cancer pathways

Another logical model of similar size and scope was also used, primarily for the study of treatment responses. This model was built by Loïc Verlingue, a medical oncologist and member of the laboratory and preliminary versions of the model are described in Verlingue et al. [2016b] and Verlingue et al. [2016a]. One of the interests of this model is that it has been designed with a more clinical perspective, notably centred on the response to MTOR inhibitors. In addition, it presents more biological read-outs used for interpretation, and we will use mainly *Proliferation* (also called *G1_S* in the model files to designate the associated stage of the cell cycle), *Apoptosis* and *Quiescence* in particular. In addition, being able to discuss and collaborate directly with the model autor has helped to avoid potential errors in use.

## B.3 Logical model of BRAF pathways in melanoma and colorectal cancer

Here are some details about the regulations represented in Figure 6.4. The MAPK pathway encompasses three families of protein kinases: RAF, MEK,

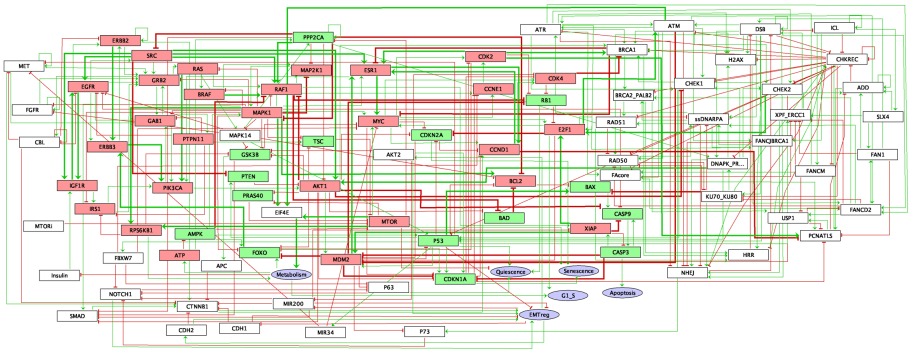



Figure B.2: **GINsim representation of the 'Verlingue' logical model described in Verlingue et al. [2016a].**

ERK. If RAF is separated into two isoforms, CRAF and BRAF, the other two families MEK and ERK are represented by a single node. When BRAF is inhibited, ERK can still be activated through CRAF, and BRAF binds to and phosphorylates MEK1 and MEK2 more efficiently than CRAF [Wellbrock et al., 2004], especially in his V600E/K mutated form. When PI3K/AKT pathway is activated, through the presence of the HGF (Hepatocyte Growth Factors), EGF (Epidermal Growth Factors) and FGF (Fibroblast Growth Factors) ligands, it leads to a proliferative phenotype. The activation of this pathway results in the activation of PDPK1 and mTOR, both able to phosphorylate p70 (RPS6KB1) which then promotes cell proliferation and growth [Consortium, 2019]. There has been some evidence of negative regulations of these two pathways carried out by ERK itself [Lake et al., 2016]: phosphorylated ERK is able to prevent the SOS-GRB2 complex formation through the activation of SPRY [Edwin et al., 2009], inhibit the EGF-dependent GAB1/PI3K association [Lehr et al., 2004] and downregulate EGFR signal through phosphorylation [Lake et al., 2016]. The model also accounts for a negative regulation of proliferation through a pathway involving p53 activation in response to DNA damage (represented by ATM); p53 hinders proliferation through the activation of both PTEN, a PI3K inhibitor, and p21 (CDKN1A) responsible for cell cycle arrest.

We hypothesize that a single network is able to discriminate between melanoma and CRC cells. These differences may come from different sources. One of them is linked to the negative feedback loop from ERK to EGFR. As mentioned previously, this feedback leads to one important difference in response to treatment between melanoma and CRC: $BRAF^{(V600E)}$ inhibition causes a rapid feedback activation of EGFR,

which supports continued proliferation. This feedback is observed only in colorectal since melanoma cells express low levels of EGFR and are therefore not subject to this reactivation [Prahallad et al., 2012]. Moreover, phosphorylation of SOX10 by ERK inhibits its transcription activity towards multiple target genes by interfering with the sumoylation of SOX10 at K55, which is essential for its transcriptional activity [Han et al., 2018]. The absence of ERK releases the activity of SOX10, which is necessary and sufficient for FOXD3 induction. FOXD3 is then able to directly activate the expression of ERBB3 at the transcriptional level, enhancing the responsiveness of melanoma cells to NRG1 (the ligand for ERBB3), and thus leading to the reactivation of both MAPK and PI3K/AKT pathways [Han et al., 2018]. Furthermore, it has been shown that in colorectal cells, FOXD3 inhibits EGFR signal *in vitro* [Li et al., 2017]. Interestingly, SOX10 is highly expressed in melanoma cell lines when compared to other cancer cells. In the model, we define SOX10 as an input because of the lack of information about the regulatory mechanisms controlling its activity. The different expression levels of SOX10 have been reported to play an important role in melanoma (high expression) and colorectal (low expression) cell lines.

Besides a list of formalized biological assertions, retrieved from literature, has been used during the model building to ensure the consitency of the model with some qualitative behaviours. These assertions, listed below, are all verified when the logical model is simulated (details are available on the corresponding GitHub repository):

- BRAF inhibition causes a feedback activation of EGFR in colorectal cancer and not in melanoma [Prahallad et al., 2012]
- MEK inhibition stops ERK signal but activates the PI3K/Akt pathway and increases the activity of ERBB3 [Gopal et al., 2010, Lake et al., 2016]
- HGF signal leads to the reactivation of the MAPK and PI3K/AKT pathways, and resistance to BRAF inhibition [Wroblewski et al., 2013]
- BRAF inhibition in melanoma activates the SOX10/FOXD3/ERBB3 axis, which mediates resistance through the activation of the PI3K/AKT pathway [Han et al., 2018]
- Overexpression/mutation of CRAF results in constitutive activation of ERK and MEK also in the presence of a BRAF inhibitor [Manzano et al. [2016]; johannessen2010cot]
- Early resistance to BRAF inhibition may be observed in case of PTEN

loss, or mutations in PI3K or AKT [Manzano et al., 2016]
- Experiments in melanoma cell lines support combined treatment with BRAF/MEK + PI3K/AKT inhibitors to overcome resistance [Manzano et al., 2016]
- BRAF inhibition (Vemurafenib) leads to the induction of PI3K/AKT pathway and inhibition of EGFR did not block this induction [Corcoran et al., 2012]
- Induction of PI3K/AKT pathway signaling has been associated with decreased sensitivity to MAPK inhibition [Corcoran et al., 2012]

## B.4 Logical model of prostate cancer

In the context of the European project PRECISE (Personalized Engine for Cancer Integrative Study and Evaluation), focused on the integrative study of prostate cancer, an adapted logical model has been built. This prostate cancer model is initially based on the generic structure of the Fumia model presented in section B.1, which has been considerably enriched and extended with genes and mechanisms specific to prostate cancer such as ERG, SPOP or AR. The model contains 133 nodes and 449 edges (Figure B.3) and includes pathways like androgen receptor and growth factor signalling, several signaling pathways (Wnt, NFkB, PI3K/AKT, MAPK, mTOR, SHH), cell cycle, epithelial-mesenchymal transition (EMT), Apoptosis, DNA damage, etc. The model has 9 inputs (EGF, FGF, TGF beta, Nutrients, Hypoxia, Acidosis, Androgen, TNF alpha and Carcinogen presence) and 6 outputs (*Proliferation*, *Apoptosis*, *Invasion*, *Migration*, (bone) *Metastasis* and *DNA repair*).

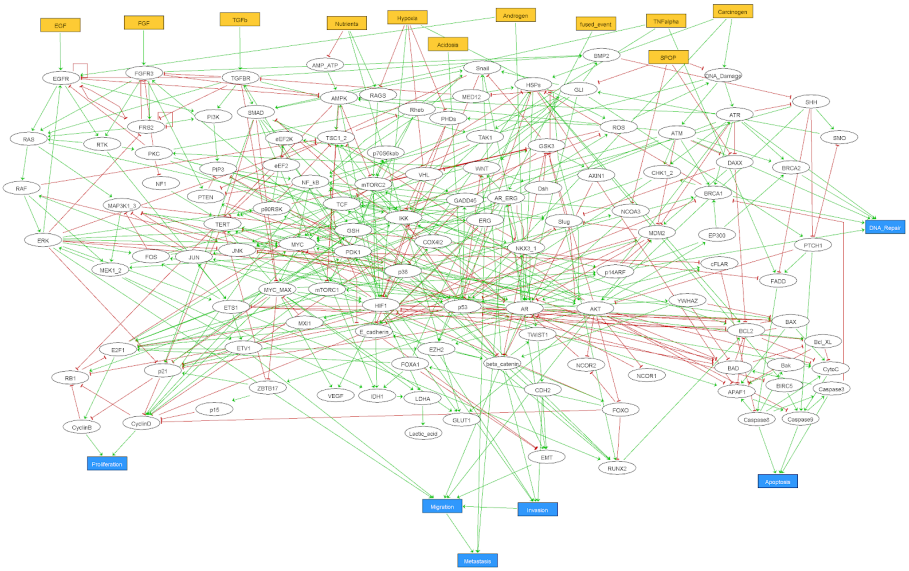



Figure B.3: **GINsim representation of the 'Montagud' logical model of prostate cancer.**

APPENDIX C

# About statistics

## C.1 $R^2$ and beyond

### C.1.1 Decomposition of $R^2$

The decomposition of $R^2$ according to the method of Lindeman [1980] is detailed below. The presentation is taken directly from Grömping et al. [2006].

A linear model is written $y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + e_i$ and the corresponding $R^2$ is:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y_i})^2}$$

Additionally, we define $R^2(S)$ for a model with regressors in set S. The additional $R^2$ when adding the regressors in set $M$ to a model with the regressors in set $S$ is given as:

$$seqR^2(M|S) = R^2(M \cup S) - R^2(S)$$

The order of the regressors in any model is a permutation of the available regressors $x_1, ..., x_p$ and is denoted by the tuple of indices $r = (r_1, ..., r_p)$.

Let $S_k(r)$ denote the set of regressors entered into the model before regressor $x_k$ in the order $r$. Then the portion of $R^2$ allocated to regressor $x_k$ in the order $r$ can be written as

$$seqR^2(\{x_k\}|S_k(r)) = R^2(\{x_k\} \cup S_k(r)) - R^2(S_k(r))$$

All in all, the $R^2$ allocated to $x_k$ after decomposition is:

$$R^2_{decomp}(x_k) = \frac{1}{p!} \sum_{r \text{ permutations}} seqR^2(\{x_k\}|r)$$

### C.1.2 $R^2$ for survival data

Among the different $R^2$ analogues that have been proposed to measure the variation explained by survival models, the one described by Royston and Sauerbrei [2004], called $R^2_D$ appears to be one of the most relevant with respect to the following criteria: independance from sensoring, interpretability and robustness to model misspecification [Choodari-Oskooei et al., 2012].

The description is given below in the context of a Cox proportional hazards (PH) survival model with $n$ individuals with $T_i$ and $C_i$ corresponding respectively to potential death (or relapse) and censoring times, with $i = 1, 2, ..., n$. In this time-to-event setting, $X_i = min(T_i, C_i)$ is the time variables and $\delta_i = I(T_i \leq C_i)$ the status variable, $I$ being the indicator function. The Cox PH model then expresses the hazard function as follows:

$$h(t|X) = h_0(t).\exp(\beta' X)$$

,

with $t$ the time to a death event, $X$ the covariate vector and *beta* the parameter vector. The adapted $R^2$ called $R^2_D$ is given by [Royston, 2006]:

$$R^2 = \frac{D^2/\kappa^2}{D^2/\kappa^2 + \sigma^2_\epsilon}$$

,

with the following component:

- $D$ quantifies the separation of survival curves. It is computed by ordering the estimated prognostic index, $\beta' X$, calculating the expected standard normal order statistics corresponding to these values, dividing the latter by a factor $\kappa$, and performing an auxiliary regression on the scaled scores: the resulting regression coefficient is $D$.
- $\kappa = \sqrt{8/\pi} \approx 1.60$ [Royston and Sauerbrei, 2004]
- $\sigma_\epsilon^2$ is the variance of the error term, $\sigma_\epsilon^2 = \pi^2/6$ for Cox PH models

For a better understanding of this formula, it is interesting to note that in a linear regression model $Y \sim N(\beta' X, \sigma^2)$, it is also possible to write $R^2$ equivalently as follows:

$$R^2 = \frac{\text{Var}(\beta' X)}{\text{Var}(\beta' X) + \sigma^2}$$

This formula underlines the analogy with $R_D^2$, with $D^2/\kappa^2$ being interpreted as an estimate of the variance of the prognostic index $\beta' X$ for the Cox PH model.

## C.2 Causal inference with multiple versions of treatment

This section gathers the demonstrations of the equations present in chapter 9 when they are not present in this chapter and additional details about other estimators based on IPW and TMLE.

### C.2.1 Overall treatment effect with multiple versions of treatment (equation (9.3))

Here is the formal proof for equation (9.3), mostly derived from the proof of Proposition 3 in [VanderWeele and Hernan, 2013].

$$
\begin{aligned}
E[Y(a, K^a(a))] &= E[Y(a)] && K^a \text{ actually received} \\
&= \sum_{c,w} E[Y(a)|c,w] \times P[c,w] \\
&= \sum_{c,w} E[Y(a)|a,c,w] \times P[c,w] && Y(a) \perp\!\!\!\perp A|(C,W) \\
&= \sum_{c,w} E[Y(a, K^a(a))|a,c,w] \times P[c,w] \\
&= \sum_{c,w,k^a} E[Y(a,k^a)|a, K^a(a) = k^a, c, w] \times P[K^a(a) = k^a|a,c,w] \times P[c,w] \\
&= \sum_{c,w,k^a} E[Y(a,k^a)|a, K^a = k^a, c, w] \times P[K^a = k^a|a,c,w] \times P[c,w] && \text{consistency K} \\
&= \sum_{c,w,k^a} E[Y|a, K^a = k^a, c, w] \times P[K^a = k^a|a,c,w] \times P[c,w] && \text{consistency Y} \\
&= \sum_{c,w} E[Y|a,c,w] \times P[c,w]
\end{aligned}
$$

Then, the overall treatment effect can be defined and computed by:

$$E[Y(a, K^a(a))] - E[Y(a^*, K^{a^*}(a^*))]$$

### C.2.2 Treatment effect with predefined distributions of versions of treatment (equation (9.5))

Here is the formal proof for equation (9.5), partially derived from the proof of Proposition 5 in [VanderWeele and Hernan, 2013].

$$
\begin{aligned}
E[Y(a, G^a)] &= \sum_{c} E[Y(a, G^a)|C = c] \times P[c] \\
&= \sum_{c,k^a} E[Y(a,k^a)|G^a = k^a, C = c] \times P[G^a = k^a|C = c] \times P[c] \\
&= \sum_{c,k^a} E[Y(a,k^a)|C = c] \times g^{k^a,c} \times P[c] && \text{since } P[G^a = k^a] = g^{k^a,c} \\
&= \sum_{c,k^a} E[Y(a,k^a)|A = a, K^a = k^a, C = c] \times g^{k^a,c} \times P[c] && \text{with } Y(a,k^a) \perp\!\!\!\perp \{A,K\}|C \\
&= \sum_{c,k^a} E[Y|A = a, K^a = k^a, C = c] \times P[c] && \text{by consistency for Y}
\end{aligned}
$$

### C.2.3 Inverse probability of treatment weighted (IPW) estimators for precision medicine

An extension of IPW methods described in section 8.2.2.2 to multi-valued treatments (only treatment $K$ with different modalities and no $A$) has already been studied and the different formulas and estimators adapted accordingly [Imbens, 2000, Feng et al., 2012], defining in particular a generalized propensity score:

$$f(k|c) = P[K = k|C = c] = E[I(k)|C = c]$$

$$\text{with } I(k) = \begin{cases} 1 & \text{if } K = k \\ 0 & \text{otherwise} \end{cases}$$

and a subsequent estimator:

$$E[Y(k)] = \frac{\hat{E}[I(K = k)W^K Y]}{\hat{E}[I(K = k)W^K]} \text{ with } W^K = \frac{1}{f[K|C]}$$

In our precision medicine settings, to be consistent with the previously defined causal diagram (Figure 9.5), we have both $A$, binary status depending on the class of drugs, and $K$, the multinomial variable for versions of treatments, *i.e.*, the precise drug. Therefore we need to define a slightly different propensity score with joint probabilities:

$$\begin{aligned} f(a, k|c) &= P[A = a, K = k|C = c] \\ &= P[K = k|A = a, C = c].P[A = a|C = c] \\ &= E[I(a, k)|C = c] \end{aligned}$$

$$\text{with } I(a, k) = \begin{cases} 1 & \text{if } A = a, K = k \\ 0 & \text{otherwise} \end{cases}$$

From this we can deduce the estimator:

$$E[Y(a, k)] = \frac{\hat{E}[I(A = a, K = k)W^{A,K} Y]}{\hat{E}[I(A = a, K = k)W^{A,K}]} \text{ with } W^{A,K} = \frac{1}{f[A, K|C]}$$

In all the examples presented in this study and implemented in the code, $\mathcal{K}^0 \cap \mathcal{K}^1 = \emptyset$, it is therefore possible to simplify the joint probabilities since the knowledge of K automatically results in the knowledge of A allowing $P[A = a, K = k|C = c] = P[K = k|C = c]$. The above formulas with the attached probabilities are still necessary in the general case and allow for the derivation of causal effects $CE_1$, $CE_2$ and $CE_3$ previously described.

### C.2.4 TMLE

Targeted maximum likelihood estimation is framework based on a doubly robust maximum-likelihood–based approach that includes a "targeting" step that optimizes the bias-variance trade-off for a defined target parameter. In particular, this method is perfectly compatible with the use of machine learning algorithms for outcome or treatment models. A detailed description of the method and its implementations can be found in Van der Laan and Rose [2011].

The implementation proposed in this article is very similar to the one proposed in a recent tutorial concerning the application to binary processing [Luque-Fernandez et al., 2018]. The specific characteristics of the problem of precision medicine studied here lead to modify this approach. In particular, the outcome and treatment models used in the first steps are modified in the same way as the one explained for the standardized estimators (outcome model) and for the IPW estimators (treatment model). The step of updating the estimates is done on a model similar to Luque-Fernandez et al. [2018].

The algorithm used for the models internal to the TMLE are, as much as possible, the same as those used for the standardised and IPW estimators:

- For simulated data: generalized linear models in all cases except multinomial classification performed through the function *multinom* in *nnet* package.
- For PDX data: random forests for all models. Use of SuperLearner [Van der Laan et al., 2007] is made possible by simple modifications to the code but significantly slows down its execution.

Appendix D

# Résumé détaillé

En vertu de l'article L121-3 du code de l'éducation la langue de rédaction privilégiée en France pour les thèses est le français. Une thèse en anglais doit en conséquence s'accompagner d'un résumé détaillé en français qui est fourni ci-dessous en suivant une progression similaire au manuscrit qui peut alors être titré:

**De la modélisation mécanistique des voies de signalisation dans le cancer à l'interprétation des modèles et de leurs apports : applications cliniques et évaluation statistique**

## D.1 Modélisation et cancer

### La modélisation scientifique, la complexité et l'abstraction

Il importe en premier lieu d'effectuer une brève clarification sémantique et épistémologique concernant les *modèles*, de loin le mot plus fréquent de cette thèse. C'est en effet un terme polysémique, y compris en se restreignant à la seule pratique scientifique. Si les modèles y sont généralement reconnus comme des représentations des phénomènes étudiés, ils peuvent recouvrir des réalités diverses, tantôt objet physique manipulable (Figure 1.1) et tantôt construction purement formelle ou mathématique. C'est à cette deuxième catégorie que nous allons nous attacher tout au long de cette thèse, en la spécifiant davantage encore.

Ainsi, on distinguera par la suite les modèles **mécanistiques** des modèles **statistiques**. Si tous deux sont des modèles formels, ils diffèrent par leur structure et leurs objectifs (Figure 1.6 et Tableau 1.1). Les premiers cherchent à représenter les mécanismes internes du phénomène, quand les deuxièmes cherchent à optimiser la prédiction du phénomène sans présupposer de connaissances particulières. Une illustration de ces différences, appliquée à la modélisation des interactions écologiques entre proies et prédateurs, est proposée en section 1.2.2.

Un point crucial à retenir concernant ces modèles, et en particulier les modèles mécanistiques, est leur nécessaire simplicité. Puisque leur existence découle de la complexité des phénomènes étudiés, les modèles sont par nature des **simplifications de la réalité**. Le recours au bon niveau de détail et la justification des choix effectués sont ainsi beaucoup plus importants qu'une impossible exactitude. Ou comme le disait Paul Valéry~:

> *Ce qui est simple est toujours faux. Ce qui ne l'est pas est inutilisable*

### De la dérégulation de la machinerie cellulaire au cancer

Le cancer, de par sa grande complexité cellulaire et moléculaire, est un terrain de prédilection pour les modèles en tout genre. Les manifestations cliniques de la maladie, caractérisée par une prolifération incontrôlée de cellules, sont connues depuis des siècles. La compréhension des mécanismes biologiques sous-jacents ne s'est elle approfondie qu'à partir de la découverte de l'ADN comme support de l'information génétique, au milieu du XX$^{\text{ème}}$ siècle. Le cancer est aujourd'hui reconnu comme une **maladie génétique** dont l'origine réside essentiellement dans des altérations de l'ADN. C'est la **combinaison de plusieurs altérations**, souvent accumulées au fil du temps, qui permet d'inactiver les différentes protections biologiques qui prémunissent en temps normal contre une prolifération cellulaire excessive.

Ainsi, il apparait de plus en plus indispensable de considérer les informations biologiques de chaque patient (mutations de l'ADN, niveaux d'expression ARN etc.) non plus séparément mais conjointement afin de comprendre le phénomène tumoral. La prise en compte des **réseaux d'interactions** entre gènes, ARN et protéines éclaire également la compréhension : le cancer n'est plus seulement une maladie génétique mais également une maladie de réseaux (Figure 2.8).

Sur le plan technologique, la recherche profite également depuis le début du XXI$^{ème}$ siècle de données beaucoup plus abondantes, issues notamment du séquençage à haut débit, qui permettent une vision plus globale en renseignant sur des milliers de gènes et ce à travers **différents types de données omiques** : génomique (ADN), transcriptomique (ARN), protéomique (protéines) etc. Ces données sont notamment disponibles publiquement pour des **milliers de patients** atteints de cancer (consortium TCGA par exemple) mais aussi pour de **nombreux modèles précliniques** comme des lignées cellulaires provenant de patients.

**Modélisation mécanistique du cancer : d'une maladie complexe à la biologie des systèmes**

L'abondance des données et des relations entre les entités biologiques a rendu nécessaire l'utilisation de méthodes computationnelles pour les comprendre et les modéliser. Ce thèse se focalise sur la **modélisation au niveau moléculaire**, celui des interactions entre gènes, ARN et protéines au sein des cellules. Même en se focalisant sur les modèles mécanistiques qui intègrent la connaissance biologique et biochimique sur ces entités, plusieurs formalismes mathématique existent pour écrire un modèle. Deux formalismes parmi les plus fréquents concentreront l'essentiel des analyses de cette thèse. Le premier est constitué d'**équations différentielles**, il est quantitatif mais mobilise de nombreux paramètres. Le second est le **formalisme logique**, plus parcimonieux mais qualitatif et qui sera décrit en détail plus avant.

L'un comme l'autre ont pour vocation de répliquer le phénomène tumoral étudié (activation d'une voie de signalisation, impact d'une mutations etc.) tout en représentant explicitement les entités biologiques impliquées. De par leur structure complexe, souvent non-linéaire, ils peuvent mettre en évidence des comportements dit **émergents** qui relèvent d'une réponse du système dans son ensemble et ne pouvaient se déduire des entités biologiques prises séparément.

Au-delà de leur utilité intellectuelle et scientifique dans l'étude des phénomènes tumoraux, ces modèles mécanistiques moléculaires du cancer sont parfois utilisés pour des **analyses ou prédictions à portée clinique et médicale** : survie d'un patient en l'absence de traitement (valeur pronostique), réponse d'un patient à un traitement donné (valeur prédictive). Cette thèse se focalise essentiellement sur ces questions d'impact clinique des modèles mécanistiques de cancer, entre biologie des systèmes

et biostatistiques.

## D.2 Des modèles logiques personnalisés de cancer

### Principes de modélisation logique et intégration des données

L'essentiel des modèles mécanistiques présentés dans cette thèse relèvent du formalisme logique. Chaque entité biologique y est représentée par une **variable discrète**, souvent binaire, interprétée comme une abstraction de son activité: 0 si inactive (gène non transcrit, ARN dégradé, protéine en trop faible concentration etc.), 1 si active (gène transcrit, protéine phosphorylée etc.). Ces entités sont reliées entre elles par des règles logiques composées à partir des opérateurs ET (&), OU (|), NON (!). On pourra ainsi définir qu'une entité A doit être activée si B l'est et que, dans le même temps, C ne l'est pas ("B & !C").

Par la suite les modèles logiques seront simulés suivant une **actualisation asynchrone** telle qu'encodée dans le logiciel MaBoSS dont le fonctionnement est résumé en Figure 4.4. En bref, cela signifie que la mise à jour de l'état des variables du modèle (par exemple le passage de l'état 0 à l'état 1 si la règle logique est vérifiée) se fait une variable à la fois et que l'ordre des mises à jour est défini stochastiquement à partir de **constantes de réactions** associés aux variables (algorithme de Gillespie).

La plupart des modèles logiques utilisés par la suite ont été construits à partir de la littérature comme source primaire d'information. Cependant, dans ce formalisme comme dans les autres, les **données biologiques sont de toute première importance** à différentes étapes du modèles, que ce soit pour le définir, la paramétrer ou le valider (Figure 4.5).

### Personnalisation des modèles logiques : méthode et validation pronostique

Les modèles définis à partir de la littérature sont par construction assez génériques et ne permettent pas d'explorer, pour un même cancer, les différences entre individus en termes d'agressivité de la tumeur ou de réponse au traitement. À partir d'un réseau moléculaire générique, une des possibilités est d'utiliser les données biologiques des différentes tumeurs pour spécifier, personnaliser les modèles. Les méthodes existantes proposent pour

la plupart d'**entraîner et d'optimiser les paramètres des modèles** à l'aide d'une fonction d'objectif prédéfinie. Cela requiert cependant des données riches, souvent des données de perturbations qui sont rarement accessibles en dehors de certains modèles précliniques comme les lignées cellulaires.

Une autre approche a été suivie dans cette thèse consistant à personnaliser les modèles en **interprétant biologiquement des données statiques** (par opposition aux données de perturbation) issues d'un seul prélèvement initial. Le formalisme logique est alors mobilisé, sa nature qualitative et parcimonieuse étant adaptée aux données utilisées. Peut-on alors obtenir de ces données des modèles mécanistiques de cancer personnalisés et interprétables cliniquement ? La **méthode PROFILE** conçue durant cette thèse cherche à répondre à cette question **sans entraînement des modèles**.

Son principe peut se diviser en deux méthodes distinctes: **la personnalisation discrète et la personnalisation continue** (Figure 5.5). La première infère l'**effet fonctionnel des altérations génétiques** considérées discrètes comme les mutations ou les altérations du nombre de copies d'un gène. Une mutation connue pour impliquer une perte (resp. un gain) de fonction de l'entité biologique sera traduite en forçant la variable correspondante du modèle à rester à 0 (resp. 1) pour le patient concerné. La seconde méthode consiste elle à interpréter des quantités continues comme peuvent l'être les niveaux d'ARN, de protéines, de phosphorylation etc. Il s'agit alors, **pour chaque variable, de détecter la forme de distribution au sein de la cohorte puis de la normaliser**, c'est à dire de la transformer en une variable continue comprise entre 0 et 1 (Figure 5.4). Cette valeur peut ensuite être utilisée pour définir les constantes de réactions associées à chaque variable du modèle et qui influent sur la rapidité avec laquelle ladite variable sera mise à jour. Les deux types de personnalisations peuvent être combinés, en intégrant par exemple les mutations et les niveaux d'ARN chacun suivant la méthode adaptée. Cette combinaison sera celle utilisée par défaut dans les analyses évoquées ultérieurement.

Une première validation de la méthode est faite en vérifiant la capacité des modèles logiques personnalisés à différencier des tumeurs plus ou moins agressives. Des analyses de survies de patientes atteintes de cancers du sein sont notamment proposées en Figure 5.9 et démontrent la capacité des modèles personnalisés à **stratifier des patientes présentant des pronostics différents**.

**Des modèles logiques personnalisés pour interpréter la réponse aux traitements**

Mais l'intérêt principal des modèles mécanistiques personnalisés réside dans leur représentation explicite des entités et mécanismes biologiques sous-jacents. Une des façons de mettre à profit cet avantage est de s'intéresser à l'**effet de certains traitements sur les modèles**. Il est en effet possible de modéliser l'effet d'un traitement si sa cible et son mode d'action sont suffisamment connus. Cette analyse n'est *a priori* pas possible dans un modèle statistique si le traitement n'est pas compris dans les données d'entraînement.

Une première analyse à large spectre est menée à l'aide de données concernant plusieurs centaines de lignées cellulaires (provenant de différents types de cancer) et de traitements. L'utilisation d'un modèle logique générique n'a cependant pas permis de tirer des enseignements précis de cette analyse. Une étude plus ciblée a donc été menée par la suite pour étudier la **réponse aux inhibiteurs de BRAF des mélanomes et cancers colorectaux**, étant observé que les premiers répondent assez bien au traitement et pas les seconds, en dépit de profils moléculaires assez semblables. Un modèle centré autour de BRAF tout d'abord construit et validé qualitativement. Il est ensuite personnalisé à l'aide des données de mutations et d'ARN issues de lignées cellulaires des deux cancers concernés. La réponse des modèles personnalisés à l'inhibition de BRAF est ensuite évaluée en observant le score de *Prolifération* (variable de sortie phénotypique du modèle) atteint par ces modèles avec et sans inhibiteurs. La réduction de *Prolifération* engendrée par l'inhibition simulée de BRAF est comparée aux résultats expérimentaux issus de l'inhibition de BRAF sur les mêmes lignées cellulaires par un traitement ou par CRISPR/Cas9. Les corrélations significatives entre sensibilités expérimentales et simulées valident la pertinence des modèles personnalisés, en particulier ceux résultant de l'intégration conjointe des données de mutation et d'ARN (Figure 6.6). Les modèles personnalisés apparaissent par ailleurs comme un outil d'investigation qualitative porteur de sens en permettant de mettre en lumière et en contexte certains mécanismes de résistance, liés à CRAF par exemple (Figure 6.7). Une comparaison avec des méthodes d'apprentissage automatique, ici des forêts aléatoires, est effectuée afin de souligner la complémentarité des approches et l'intérêt des modèles logiques personnalisés sans entraînement en présence d'un nombre d'échantillons réduit.

Une application de cette méthode de personnalisation au cancer de la prostate est ensuite présentée. La stratégie est similaire au cas précédent, en insistant sur l'identification de nouvelles cibles thérapeutiques pour une lignée cellulaire en particulier. Certaines cibles identifiées comme efficaces par les modèles personnalisés sont ensuite testées et validées *in vitro*.

## D.3 Quantification statistique de l'impact clinique des modèles

### Flux d'informations dans les modèles mécanistiques de cancer

La valeur clinique, c'est à dire pronostique ou prédictive, des modèles mécanistiques a été analysée jusqu'ici de manière assez simple en se focalisant sur la capacité des sorties du modèle (souvent des variables à interprétation phénotypique comme *Prolifération*) à corréler avec des résultats cliniques. Cependant les données utilisées en entrée des modèles mécanistiques, pour les paramétrer ou les personnaliser, sont souvent déjà des variables pertinentes cliniquement : statut d'une mutation, biomarqueur ARN etc. Il est donc nécessaire de considérer à la fois les variables d'entrée et de sortie du modèle à l'aune des résultats cliniques afin de comprendre **comment le modèle traite et transforme les informations cliniques qu'il ingère**.

Un premier exemple est fourni, fondé sur des données simulées et l'utilisation du pourcentage de la variance exprimée ($R^2$), pour mettre en exergue deux grandes catégories de modèles mécanistiques (Figures 7.2 et 7.3). La première correspond à ceux dont les sorties ont une valeur clinique inférieure aux entrées. Les sorties du modèles, généralement en nombre réduit, peuvent néanmoins constituer une **réduction de dimension** pertinente des entrées, souvent nombreuses. La seconde catégorie est celle des modèles dont les sorties ont une valeur supérieure ou complémentaire aux entrées : ils ont **capturé un comportement émergent** qui a une valeur clinique. Cette analyse est appliquée à des modèles déjà publiés pour illustrer la nature de l'information clinique qu'ils fournissent.

### Essais cliniques et inférence causale

Si l'on se focalise sur la capacité des modèles mécanistiques à orienter le choix des traitements, il devient nécessaire de développer d'autres méthodes d'analyse des modèles, plus proches de la pratique clinique. En la matière, l'évaluation des traitements et stratégies thérapeutiques se fait

souvent à travers des **essais cliniques randomisés** qui permettent de comparer des populations similaires qui ne diffèrent que par l'administration du traitement ou du contrôle. Dans le cas de données non randomisées, que l'on appellera ici **observationnelles**, le choix des patients qui ont reçu le traitement ou le contrôle peut avoir été fait suivant des critères particuliers qui deviennent ainsi des **facteurs de confusion** dans l'analyse : la différence de résultats peut provenir de l'effet du traitement ou d'une répartition déséquilibrée des patients (Figures 8.2 et 8.3).

Il est toutefois possible d'utiliser des méthodes pour corriger certains de ses déséquilibres et ainsi, sous certaines conditions, d'interpréter les différences résiduelles de résultats entre patients traités et contrôles comme résultant de l'effet causal du traitement. Trois implémentations différentes de ces méthodes d'**inférence causale** sont utilisées dans la thèse, chacune fondée sur un modèle statistique différent pour contrôler l'effet des facteurs de confusion: la standardisation (modèle du résultat du traitement), la pondération selon l'inverse de la probabilité (modèle de l'assignation du traitement), le *targeted maximum likelihood estimation* TMLE (combinaison des deux modèles précédents). Ces méthodes dépendent cependant toutes d'hypothèses parfois difficiles à vérifier.

**Inférence causale pour la médecine de précision**

Mais les modèles mécanistiques peuvent théoriquement permettre de choisir parmi plusieurs traitements pour chaque profil moléculaire de patient. Ils sont ainsi assimilables à une **stratégie ou un algorithme de médecine de précision**. Peut-on alors appliquer les méthodes d'inférence causale à l'évaluation de ce genre de stratégies cliniques qui englobent différentes molécules thérapeutiques ? Dans le dernier temps de cette thèse, une extension de ces méthodes est proposée pour s'adapter à des stratégies cliniques comprenant **différentes versions de traitement**. Cela revient à utiliser des données observationnelles pour simuler des essais cliniques comparant un bras où les patients sont traités suivant les recommandations de l'algorithme de médecine de précision (qui peut être un modèle mécanistique) avec différents types de bras contrôle: traitement standard de référence (molécule unique), traitement prévu par le médecin, ou traitement ciblé aléatoirement défini.

Ces méthodes et les estimateurs statistiques associés sont d'abord validés sur des données artificiellement générées afin de mesurer leur capacité à corriger d'éventuels biais causés par l'hétérogénéité de l'effet des traite-

ments ou la répartition inégale des traitements observés par exemple. Une **application interactive RShiny** est également fournie pour permettre l'exploration de très nombreux scénarios de simulations, y compris par un public non familier du langage R (Figure 9.7).

Dans un deuxième temps, une application a été menée avec des données précliniques issues de **xénogreffes dérivées de patients** (PDX). Cela correspond à des tumeurs prélevées chez des patients avant d'être divisées en échantillons implantés chez des souris immunodéprimées identiques qui seront traitées avec des molécules différentes (Figure 9.1). En conséquence, il est possible d'accéder, pour chaque tumeur, à sa réponse thérapeutique face à différents traitements. Ces informations, qui ne sont pas disponibles pour les vrais patients qui ne peuvent recevoir qu'un traitement à la fois, permettent d'**accéder directement dans les données à l'effet de différentes stratégies thérapeutiques pour chaque tumeur**. Ainsi, il a été possible de valider, sur des données précliniques réelles, la supériorité des valeurs issues des nouvelles méthodes d'inférence causale proposées par rapport aux valeurs obtenues suivant des méthodes plus directes.

## D.4 Conclusion

Cette thèse trace un chemin, **de la conception de modèles mécanistiques du cancer à la quantification de leur impact clinique**. La faculté de ces modèles à passer de l'outil théorique à une interprétation clinique repose ici sur leur personnalisation permise par l'intégration des données omiques dans un canevas tissé à partir des connaissances biologiques issues de la littérature. Les modèles logiques personnalisés qui ont été présentés restent essentiellement des supports qualitatifs à l'interprétation clinique des phénomène tumoraux. La réflexion menée par la suite sur leur impact clinique a mis en évidence l'importance d'une évaluation globale qui permette de comprendre l'origine de leur valeur clinique et la nécessité de développer des méthodes statistiques adaptées pour évaluer leurs apports en termes de médecine de précision.

# Bibliography


Nanne Aben, Daniel J Vis, Magali Michaut, and Lodewyk Fa Wessels. Tandem: a two-stage approach to maximize interpretability of drug response models based on multiple molecular data types. *Bioinformatics*, 32(17): i413–i420, 2016.

Adam Abeshouse, Jaeil Ahn, Rehan Akbani, Adrian Ally, Samirkumar Amin, Christopher D Andry, Matti Annala, Armen Aprikian, Joshua Armenia, Arshi Arora, et al. The molecular taxonomy of primary prostate cancer. *Cell*, 163(4):1011–1025, 2015.

Wassim Abou-Jaoudé, Pedro T Monteiro, Aurélien Naldi, Maximilien Grandclaudon, Vassili Soumelis, Claudine Chaouiya, and Denis Thieffry. Model checking to assess t-helper cell plasticity. *Frontiers in bioengineering and biotechnology*, 2:86, 2015.

Wassim Abou-Jaoudé, Pauline Traynard, Pedro T Monteiro, Julio Saez-Rodriguez, Tomáš Helikar, Denis Thieffry, and Claudine Chaouiya. Logical modeling and dynamical analysis of cellular networks. *Frontiers in genetics*, 7:94, 2016.

Rony Abou-Jawde, Toni Choueiri, Carlos Alemany, and Tarek Mekhail. An overview of targeted treatments in cancer. *Clinical therapeutics*, 25(8): 2121–2137, 2003.

Ivan A. Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E. Ramensky, Anna Gerasimova, Peer Bork, Alexey S. Kondrashov, and Shamil R. Sunyaev. A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–249, April 2010. ISSN 1548-7091. doi: 10.1038/nmeth0410-248.

István Albert, Juilee Thakar, Song Li, Ranran Zhang, and Réka Albert. Boolean network simulations for life scientists. *Source code for biology and medicine*, 3(1):16, 2008.

Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. Molecular biology of the cell. garland science. *New York*, 1392, 2007.

Bree B Aldridge, Julio Saez-Rodriguez, Jeremy L Muhlich, Peter K Sorger, and Douglas A Lauffenburger. Fuzzy logic analysis of kinase pathway crosstalk in tnf/egf/insulin-induced signaling. *PLoS Comput Biol*, 5(4): e1000340, 2009.

Douglas G Altman, Berthold Lausen, Willi Sauerbrei, and Martin Schumacher. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *JNCI: Journal of the National Cancer Institute*, 86(11): 829–835, 1994.

Philipp M Altrock, Lin L Liu, and Franziska Michor. The mathematics of cancer: integrating quantitative models. *Nature Reviews Cancer*, 15(12): 730–745, 2015.

Alexander RA Anderson and Vito Quaranta. Integrative mathematical oncology. *Nature Reviews Cancer*, 8(3):227–234, 2008.

Robyn P Araujo and DL Sean McElwain. A history of the study of solid tumour growth: the contribution of mathematical modelling. *Bulletin of mathematical biology*, 66(5):1039–1091, 2004.

Peter Armitage and Richard Doll. The age distribution of cancer and a multi-stage theory of carcinogenesis. *British journal of cancer*, 8(1):1, 1954.

Francisco Azuaje. Computational models for predicting drug responses in cancer research. *Briefings in bioinformatics*, 18(5):820–829, 2017.

Daniela M Bailer-Jones. Scientists' thoughts on scientific models. *Perspectives on Science*, 10(3):275–301, 2002.

Ruth E Baker, Jose-Maria Pena, Jayaratnam Jayamohan, and Antoine Jérusalem. Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biology letters*, 14(5):20170660, 2018.

Albert-Laszlo Barabasi and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101–113, 2004.

Dominique Barbolosi, Joseph Ciccolini, Bruno Lacarelle, Fabrice Barlési, and Nicolas André. Computational oncology—mathematical modelling of drug regimens for precision medicine. *Nature reviews Clinical oncology*, 13(4):242, 2016.

Emmanuel Barillot, Laurence Calzone, Philippe Hupe, Jean-Philippe Vert, and Andrei Zinovyev. *Computational systems biology of cancer*. CRC Press, 2012.

Florentin Baur, Sarah L Nietzer, Meik Kunz, Fabian Saal, Julian Jeromin, Stephanie Matschos, Michael Linnebacher, Heike Walles, Thomas Dandekar, and Gudrun Dandekar. Connecting cancer pathways to tumor engines: A stratification tool for colorectal cancer combining human in vitro tissue models with boolean in silico models. *Cancers*, 12(1):28, 2020.

Jonas Béal and Aurélien Latouche. Causal inference with multiple versions of treatment and application to personalized medicine. *arXiv preprint arXiv:2005.12427*, 2020.

Jonas Béal, Arnau Montagud, Pauline Traynard, Emmanuel Barillot, and Laurence Calzone. Personalization of Logical Models With Multi-Omics Data Allows Clinical Stratification of Patients. *Frontiers in Physiology*, 2019. ISSN 1664-042X. doi: 10.3389/fphys.2018.01965.

Jonas Béal, Lorenzo Pantolini, Vincent Noël, Emmanuel Barillot, and Laurence Calzone. Personalized logical models to investigate cancer response to braf treatments in melanomas and colorectal cancers. *bioRxiv*, 2020.

Fiona M Behan, Francesco Iorio, Gabriele Picco, Emanuel Gonçalves, Charlotte M Beaver, Giorgia Migliardi, Rita Santos, Yanhua Rao, Francesco Sassi, Marika Pinnelli, et al. Prioritization of cancer therapeutic targets using crispr–cas9 screens. *Nature*, 568(7753):511, 2019.

Nicola Bellomo, NK Li, and Ph K Maini. On the foundations of cancer modelling: selected topics, speculations, and perspectives. *Mathematical Models and Methods in Applied Sciences*, 18(04):593–646, 2008.

Sébastien Benzekry, Clare Lamont, Afshin Beheshti, Amanda Tracz, John ML Ebos, Lynn Hlatky, and Philip Hahnfeldt. Classical mathematical models for description and prediction of experimental tumor growth. *PLoS Comput Biol*, 10(8):e1003800, 2014.

Upinder S Bhalla and Ravi Iyengar. Emergent properties of networks of biological signaling pathways. *Science*, 283(5400):381–387, 1999.

Raffaella Bianucci, Antonio Perciaccante, Philippe Charlier, Otto Appenzeller, and Donatella Lippi. Earliest evidence of malignant breast cancer in renaissance paintings. *The Lancet Oncology*, 19(2):166–167, 2018.

Erhan Bilal, Janusz Dutkowski, Justin Guinney, In Sock Jang, Benjamin A Logsdon, Gaurav Pandey, Benjamin A Sauerwine, Yishai Shimoni, Hans Kristian Moen Vollan, Brigham H Mecham, et al. Improving breast cancer survival analysis through competition-based multidimensional modeling. *PLoS computational biology*, 9(5), 2013.

Paul Blanche, Michael W Kattan, and Thomas A Gerds. The c-index is not proper for the evaluation of-year predicted risks. *Biostatistics*, 20(2): 347–357, 2019.

Mehdi Bouhaddou, Anne Marie Barrette, Alan D Stern, Rick J Koch, Matthew S DiStefano, Eric A Riesel, Luis C Santos, Annie L Tan, Alex E Mertz, and Marc R Birtwistle. A mechanistic pan-cancer pathway model informed by multi-omics data interprets stochastic cell fate responses to drugs and mitogens. *PLoS computational biology*, 14(3):e1005985, 2018.

Ivana Bozic, Tibor Antal, Hisashi Ohtsuki, Hannah Carter, Dewey Kim, Sining Chen, Rachel Karchin, Kenneth W Kinzler, Bert Vogelstein, and Martin A Nowak. Accumulation of driver and passenger mutations during tumor progression. *Proceedings of the National Academy of Sciences*, 107 (43):18545–18550, 2010.

Peter Allen Braithwaite and Dace Shugg. Rembrandt's bathsheba: the dark shadow of the left breast. *Annals of the Royal College of Surgeons of England*, 65(5):337, 1983.

Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001a.

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001b.

Marc Buyse, Sherene Loi, Laura Van't Veer, Giuseppe Viale, Mauro Delorenzi, Annuska M Glas, Mahasti Saghatchian d'Assignies, Jonas Bergh, Rosette Lidereau, Paul Ellis, et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *Journal of the National Cancer Institute*, 98(17):1183–1192, 2006.

Helen M Byrne. Dissecting cancer through mathematics: from the cell to the animal model. *Nature Reviews Cancer*, 10(3):221–230, 2010.

Jonas Béal, Elizabeth Rémy, and Laurence Calzone. Modélisation logique et données omiques : de la construction des modèles à la médecine personnalisée. In Elisabeth Rémy and Cédric Lhoussaine, editors, *Approche symbolique de la modélisation et de l'analyse des systèmes biologiques*. ISTE, 2020.

Laurence Calzone, Laurent Tournier, Simon Fourquet, Denis Thieffry, Boris Zhivotovsky, Emmanuel Barillot, and Andrei Zinovyev. Mathematical modelling of cell-fate decision in response to death receptor engagement. *PLoS computational biology*, 6(3), 2010.

Laurence Calzone, Emmanuel Barillot, and Andrei Zinovyev. Logical versus kinetic modeling of biological networks: applications in cancer research. *Current Opinion in Chemical Engineering*, 21:22–31, 2018.

Emma R Cantwell-Dorris, John J O'Leary, and Orla M Sheils. Brafv600e: implications for carcinogenesis and molecular therapy. *Molecular cancer therapeutics*, 10(3):385–394, 2011.

Debyani Chakravarty, Jianjiong Gao, Sarah Phillips, Ritika Kundra, Hongxin Zhang, Jiaojiao Wang, Julia E. Rudolph, Rona Yaeger, Tara Soumerai, Moriah H. Nissan, Matthew T. Chang, Sarat Chandarlapaty, Tiffany A. Traina, Paul K. Paik, Alan L. Ho, Feras M. Hantash, Andrew Grupe, Shrujal S. Baxi, Margaret K. Callahan, Alexandra Snyder, Ping Chi, Daniel C. Danila, Mrinal Gounder, James J. Harding, Matthew D. Hellmann, Gopa Iyer, Yelena Y. Janjigian, Thomas Kaley, Douglas A. Levine, Maeve Lowery, Antonio Omuro, Michael A. Postow, Dana Rathkopf, Alexander N. Shoushtari, Neerav Shukla, Martin H. Voss, Ederlinda Paraiso, Ahmet Zehir, Michael F. Berger, Barry S. Taylor, Leonard B. Saltz, Gregory J. Riely, Marc Ladanyi, David M. Hyman, José Baselga, Paul Sabbatini, David B. Solit, and Nikolaus Schultz. OncoKB: A Precision Oncology Knowledge Base. *JCO Precision Oncology*, (1):1–16, May 2017. ISSN 2473-4284. doi: 10.1200/PO.17.00011.

Paul B Chapman, Axel Hauschild, Caroline Robert, John B Haanen, Paolo Ascierto, James Larkin, Reinhard Dummer, Claus Garbe, Alessandro Testori, Michele Maio, et al. Improved survival with vemurafenib in melanoma with braf v600e mutation. *New England Journal of Medicine*, 364(26):2507–2516, 2011.

Travers Ching, Xun Zhu, and Lana X Garmire. Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS computational biology*, 14(4):e1006076, 2018.

Sung-Hwan Cho, Sang-Min Park, Ho-Sung Lee, Hwang-Yeol Lee, and Kwang-Hyun Cho. Attractor landscape analysis of colorectal tumorigenesis and its reversion. *BMC systems biology*, 10(1):96, 2016.

Babak Choodari-Oskooei, Patrick Royston, and Mahesh KB Parmar. A simulation study of predictive ability measures in a survival model i: explained variation measures. *Statistics in medicine*, 31(23):2627–2643, 2012.

David P. A. Cohen, Loredana Martignetti, Sylvie Robine, Emmanuel Barillot, Andrei Zinovyev, and Laurence Calzone. Mathematical Modelling of Molecular Pathways Enabling Tumour Cell Invasion and Migration. *PLOS Computational Biology*, 11(11):e1004571, November 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004571.

Randall J Cohrs, Tyler Martin, Parviz Ghahramani, Luc Bidaut, Paul J Higgins, and Aamir Shahzad. Translational medicine definition by the european society for translational medicine, 2015.

Stephen R Cole and Miguel A Hernán. Constructing inverse probability weights for marginal structural models. *American journal of epidemiology*, 168(6):656–664, 2008.

Collins. *The Collins English Dictionary*. HarperCollins, 2020. Model.

Samuel Collombet, Chris van Oevelen, Jose Luis Sardina Ortega, Wassim Abou-Jaoude, Bruno Di Stefano, Morgane Thomas-Chollier, Thomas Graf, and Denis Thieffry. Logical modeling of lymphoid and myeloid cell specification and transdifferentiation. *Proceedings of the National Academy of Sciences*, 114(23):5792–5799, June 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1610622114.

UniProt Consortium. Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47(D1):D506–D515, 2019.

Jake R Conway, Alexander Lex, and Nils Gehlenborg. Upsetr: an r package for the visualization of intersecting sets and their properties. *Bioinformatics*, 33(18):2938–2940, 2017.

Ryan B Corcoran, Hiromichi Ebi, Alexa B Turke, Erin M Coffee, Michiya Nishino, Alexandria P Cogdill, Ronald D Brown, Patricia Della Pelle, Dora Dias-Santagata, Kenneth E Hung, et al. Egfr-mediated reactivation of mapk signaling contributes to insensitivity of braf-mutant colorectal cancers to raf inhibition with vemurafenib. *Cancer discovery*, 2(3):227–235, 2012.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

James C Costello, Laura M Heiser, Elisabeth Georgii, Mehmet Gönen, Michael P Menden, Nicholas J Wang, Mukesh Bansal, Petteri Hintsanen, Suleiman A Khan, John-Patrick Mpindi, et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature biotechnology*, 32(12):1202, 2014.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Geraldine O'Sullivan Coyne, Naoko Takebe, and Alice P Chen. Defining precision: the precision medicine initiative trials nci-mpact and nci-match. *Current problems in cancer*, 41(3):182–193, 2017.

Ian A Cree, Christian M Kurbacher, Alan Lamont, Andrew C Hindley, Sharon Love, TCA Ovarian Cancer Trial Group, et al. A prospective randomized controlled trial of tumour chemosensitivity assay directed chemotherapy versus physician's choice in patients with recurrent platinum-resistant ovarian cancer. *Anti-cancer drugs*, 18(9):1093–1101, 2007.

Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258): 561–563, 1970.

Christina Curtis, Sohrab P. Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, Andy G. Lynch, Shamith Samarajiwa, Yinyin Yuan, Stefan Gräf, Gavin Ha, Gholamreza Haffari, Ali Bashashati, Roslin Russell, Steven McKinney, Anita Langerød, Andrew Green, Elena Provenzano, Gordon Wishart, Sarah Pinder, Peter Watson, Florian Markowetz, Leigh Murphy, Ian Ellis, Arnie Purushotham, Anne-Lise Børresen-Dale, James D. Brenton, Simon Tavaré, Carlos Caldas, and Samuel Aparicio. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486 (7403):346–352, April 2012. ISSN 0028-0836. doi: 10.1038/nature10983.

Helen Davies, Graham R Bignell, Charles Cox, Philip Stephens, Sarah Edkins, Sheila Clegg, Jon Teague, Hayley Woffendin, Mathew J Garnett, William Bottomley, et al. Mutations of the braf gene in human cancer. *Nature*, 417(6892):949–954, 2002.

Armand De Gramont, Sarah Watson, Lee M Ellis, Jordi Rodón, Josep Tabernero, Aimery De Gramont, and Stanley R Hamilton. Pragmatic issues in biomarker evaluation for targeted therapies in cancer. *Nature reviews Clinical oncology*, 12(4):197, 2015.

Hidde De Jong. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of computational biology*, 9(1):67–103, 2002.

Thomas S Deisboeck, Le Zhang, Jeongah Yoon, and Jose Costa. In silico cancer modeling: is it ready for prime time? *Nature Clinical Practice Oncology*, 6(1):34–42, 2009.

Antonio Del Sol, Rudi Balling, Lee Hood, and David Galas. Diseases as network perturbations. *Current opinion in biotechnology*, 21(4):566–571, 2010.

Joshua M Dempster, Clare Pacini, Sasha Pantel, Fiona M Behan, Thomas Green, John Krill-Burger, Charlotte M Beaver, Scott T Younger, Victor Zhivich, Hanna Najgebauer, et al. Agreement between two large pan-cancer crispr-cas9 gene dependency data sets. *Nature Communications*, 10(1):1–14, 2019.

Li Ding, Matthew H Bailey, Eduard Porta-Pardo, Vesteinn Thorsson, Antonio Colaprico, Denis Bertrand, David L Gibbs, Amila Weerasinghe, Kuanlin Huang, Collin Tokheim, et al. Perspective on oncogenic processes at the end of the beginning of cancer genomics. *Cell*, 173(2):305–320, 2018.

Eytan Domany. Using high-throughput transcriptomic data for prognosis: a critical overview and perspectives. *Cancer research*, 74(17):4612–4621, 2014.

Yotam Drier, Michal Sheffer, and Eytan Domany. Pathway-based personalized analysis of cancer. *Proceedings of the National Academy of Sciences*, 110(16):6388–6393, 2013.

Renato Dulbecco. A turning point in cancer research: sequencing the human genome. *Science*, 231:1055–1057, 1986.

Federica Eduati, Victoria Doldàn-Martelli, Bertram Klinger, Thomas Cokelaer, Anja Sieber, Fiona Kogera, Mathurin Dorel, Mathew J Garnett, Nils Blüthgen, and Julio Saez-Rodriguez. Drug resistance mechanisms in colorectal cancer dissected with cell type–specific dynamic logic models. *Cancer research*, 77(12):3364–3375, 2017.

Federica Eduati, Patricia Jaaks, Jessica Wappler, Thorsten Cramer, Christoph A Merten, Mathew J Garnett, and Julio Saez-Rodriguez. Patient-specific logic models of signaling pathways from screenings on cancer biopsies to prioritize personalized combination therapies. *Molecular systems biology*, 16(2), 2020.

Francis Edwin, Kimberly Anderson, Chunyi Ying, and Tarun B Patel. Intermolecular interactions of sprouty proteins and their implications in development and disease. *Molecular pharmacology*, 76(4):679–691, 2009.

Mohamed Elati, Pierre Neuvial, Monique Bolotin-Fukuhara, Emmanuel Barillot, François Radvanyi, and Céline Rouveirol. LICORN: learning cooperative regulation networks from gene expression data. *Bioinformatics*, 23(18):2407–2414, September 2007. ISSN 1367-4803. doi: 10.1093/bioinformatics/btm352.

Adrien Fauré, Aurélien Naldi, Claudine Chaouiya, and Denis Thieffry. Dynamical analysis of a generic boolean model for the control of the mammalian cell cycle. *Bioinformatics*, 22(14):e124–e131, 2006.

Ping Feng, Xiao-Hua Zhou, Qing-Ming Zou, Ming-Yu Fan, and Xiao-Song Li. Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in medicine*, 31(7):681–697, 2012.

Dana Ferranti, David Krane, and David Craft. The value of prior knowledge in machine learning of complex network systems. *Bioinformatics*, 33(22): 3610–3618, 2017.

Dirk Fey, Melinda Halasz, Daniel Dreidax, Sean P Kennedy, Jordan F Hastings, Nora Rauch, Amaya Garcia Munoz, Ruth Pilkington, Matthias Fischer, Frank Westermann, et al. Signaling pathway models as biomarkers: Patient-specific simulations of jnk activity predict the survival of neuroblastoma patients. *Sci. Signal.*, 8(408):ra130–ra130, 2015.

Keith T Flaherty, Robert Gray, Alice Chen, Shuli Li, David Patton, Stanley R Hamilton, Paul M Williams, Edith P Mitchell, A John Iafrate, Jeffrey Sklar, et al. The molecular analysis for therapy choice (nci-match)

trial: Lessons for genomic trial design. *JNCI: Journal of the National Cancer Institute*, 2020.

Gary William Flake. *The computational beauty of nature: Computer explorations of fractals, chaos, complex systems, and adaptation*. MIT press, 1998.

Åsmund Flobak, Anaïs Baudot, Elisabeth Remy, Liv Thommesen, Denis Thieffry, Martin Kuiper, and Astrid Lægreid. Discovery of drug synergies in gastric cancer cells predicted by logical modeling. *PLoS computational biology*, 11(8), 2015.

Andrew Cadle Fowler, Anna C Fowler, and AC Fowler. *Mathematical models in the applied sciences*, volume 17. Cambridge University Press, 1997.

Boris Freidlin, Lisa M McShane, and Edward L Korn. Randomized clinical trials with biomarkers: design issues. *Journal of the National Cancer Institute*, 102(3):152–160, 2010.

Adam A Friedman, Anthony Letai, David E Fisher, and Keith T Flaherty. Precision medicine for cancer with next-generation functional diagnostics. *Nature Reviews Cancer*, 15(12):747–756, 2015.

Roman Frigg and Stephan Hartmann. Models in science. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2020 edition, 2020.

Fabian Fröhlich, Thomas Kessler, Daniel Weindl, Alexey Shadrin, Leonard Schmiester, Hendrik Hache, Artur Muradyan, Moritz Schütte, Ji-Hyun Lim, Matthias Heinig, et al. Efficient parameter estimation enables the prediction of drug response using a mechanistic pan-cancer pathway model. *Cell Systems*, 7(6):567–579, 2018.

Herman F Fumia and Marcelo L Martins. Boolean network model for cancer pathways: predicting carcinogenesis and targeted therapy outcomes. *PloS one*, 8(7), 2013.

Mitchell H Gail and Ruth M Pfeiffer. Breast cancer risk model requirements for counseling, prevention, and screening. *JNCI: Journal of the National Cancer Institute*, 110(9):994–1002, 2018.

Hui Gao, Joshua M Korn, Stéphane Ferretti, John E Monahan, Youzhen Wang, Mallika Singh, Chao Zhang, Christian Schnell, Guizhi Yang, Yun Zhang, et al. High-throughput screening using patient-derived tumor

xenografts to predict clinical trial drug response. *Nature medicine*, 21 (11):1318, 2015.

Mathew J Garnett, Elena J Edelman, Sonja J Heidorn, Chris D Greenman, Anahita Dastur, King Wai Lau, Patricia Greninger, I Richard Thompson, Xi Luo, Jorge Soares, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570–575, 2012.

Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, December 1976. ISSN 0021-9991. doi: 10.1016/0021-9991(76)90041-3.

YN Vashisht Gopal, Wanleng Deng, Scott E Woodman, Kakajan Komurov, Prahlad Ram, Paul D Smith, and Michael A Davies. Basal and treatment-induced activation of akt mediates resistance to cell death by azd6244 (arry-142886) in braf-mutant human cutaneous melanoma cells. *Cancer research*, 70(21):8736–8747, 2010.

Luca Grieco, Laurence Calzone, Isabelle Bernard-Pierrot, François Radvanyi, Brigitte Kahn-Perles, and Denis Thieffry. Integrative modelling of the influence of mapk network on cancer cell fate decision. *PLoS computational biology*, 9(10):e1003286, 2013.

Ulrike Grömping et al. Relative importance for linear regression in r: the package relaimpo. *Journal of statistical software*, 17(1):1–27, 2006.

William C Hahn, Christopher M Counter, Ante S Lundberg, Roderick L Beijersbergen, Mary W Brooks, and Robert A Weinberg. Creation of human tumour cells with defined genetic elements. *Nature*, 400(6743): 464–468, 1999.

Steven I Hajdu. A note from history: landmarks in history of cancer, part 1. *Cancer*, 117(5):1097–1102, 2011a.

Steven I Hajdu. A note from history: landmarks in history of cancer, part 2. *Cancer*, 117(12):2811–2820, 2011b.

Steven I Hajdu. A note from history: landmarks in history of cancer, part 3. *Cancer*, 118(4):1155–1168, 2012a.

Steven I Hajdu. A note from history: landmarks in history of cancer, part 4. *Cancer*, 118(20):4914–4928, 2012b.

Steven I Hajdu and Farbod Darvishian. A note from history: landmarks in history of cancer, part 5. *Cancer*, 119(8):1450–1466, 2013.

Steven I Hajdu and Manjunath Vadmal. A note from history: Landmarks in history of cancer, part 6. *Cancer*, 119(23):4058–4082, 2013.

Shujun Han, Yibo Ren, Wangxiao He, Huadong Liu, Zhe Zhi, Xinliang Zhu, Tielin Yang, Yu Rong, Bohan Ma, Timothy J Purwin, et al. Erk-mediated phosphorylation regulates sox10 sumoylation and targets expression in mutant braf melanoma. *Nature communications*, 9(1):1–14, 2018.

Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *cell*, 100(1):57–70, 2000.

Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

J Hansen and R Iyengar. Computation as the mechanistic bridge between precision medicine and systems therapeutics. *Clinical Pharmacology & Therapeutics*, 93(1):117–128, 2013.

Traver Hart and Jason Moffat. Bagel: a computational framework for identifying essential genes from pooled library screens. *BMC bioinformatics*, 17(1):164, 2016.

J. A. Hartigan and P. M. Hartigan. The Dip Test of Unimodality. *The Annals of Statistics*, 13(1):70–84, March 1985. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176346577.

Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome biology*, 18(1):83, 2017.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

Patrick J Heagerty and Yingye Zheng. Survival model predictive accuracy and roc curves. *Biometrics*, 61(1):92–105, 2005.

Patrick J Heagerty, Thomas Lumley, and Margaret S Pepe. Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–344, 2000.

Suzanne Hector, Markus Rehm, Jasmin Schmid, Joan Kehoe, Niamh McCawley, Patrick Dicker, Frank Murray, Deborah McNamara, Elaine W Kay, Caoimhin G Concannon, et al. Clinical application of a systems model of apoptosis execution for the prediction of colorectal cancer therapy responses and personalisation of therapy. *Gut*, 61(5):725–733, 2012.

Michelle Heijblom, Linda M Meijer, Ton G van Leeuwen, Wiendelt Steenbergen, and Srirang Manohar. Monte carlo simulations shed light on bathsheba's suspect breast. *Journal of biophotonics*, 7(5):323–331, 2014.

Laura M Heiser, Anguraj Sadanandam, Wen-Lin Kuo, Stephen C Benz, Theodore C Goldstein, Sam Ng, William J Gibb, Nicholas J Wang, Safiyyah Ziyad, Frances Tong, et al. Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proceedings of the National Academy of Sciences*, 109(8):2724–2729, 2012.

Tomáš Helikar, John Konvalina, Jack Heidel, and Jim A Rogers. Emergent decision-making in biological signal transduction networks. *Proceedings of the National Academy of Sciences*, 105(6):1913–1918, 2008.

Tomáš Helikar, Bryan Kowal, Sean McClenathan, Mitchell Bruckner, Thaine Rowley, Alex Madrahimov, Ben Wicks, Manish Shrestha, Kahani Limbu, and Jim A Rogers. The cell collective: toward an open and collaborative approach to systems biology. *BMC systems biology*, 6(1): 96, 2012.

N Lynn Henry and Daniel F Hayes. Cancer biomarkers. *Molecular oncology*, 6(2):140–146, 2012.

MA Hernán and JM Robins. Causal inference: What if. *Boca Raton: Chapman & Hill/CRC*, 2020.

Miguel A Hernán and James M Robins. Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology*, 183(8):758–764, 2016.

Miguel A Hernán and Tyler J VanderWeele. Compound treatments and transportability of causal inference. *Epidemiology (Cambridge, Mass.)*, 22(3):368, 2011.

C Gordon Hewitt. Conservation of wild life in canada, 1917.

Manuel Hidalgo, Frederic Amant, Andrew V Biankin, Eva Budinská, Annette T Byrne, Carlos Caldas, Robert B Clarke, Steven de Jong, Jos

Jonkers, Gunhild Mari Mælandsmo, et al. Patient-derived xenograft models: an emerging platform for translational cancer research. *Cancer discovery*, 4(9):998–1013, 2014.

Jørgen Hilden and Thomas A Gerds. A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. *Statistics in medicine*, 33(19):3405–3414, 2014.

Steven M Hill, Laura M Heiser, Thomas Cokelaer, Michael Unger, Nicole K Nesser, Daniel E Carlin, Yang Zhang, Artem Sokolov, Evan O Paull, Chris K Wong, et al. Inferring causal molecular networks: empirical assessment through a community-based effort. *Nature methods*, 13(4): 310, 2016.

Susan Galloway Hilsenbeck, Gary M Clark, and William L McGuire. Why do so many prognostic factors fail to pan out? *Breast cancer research and treatment*, 22(3):197–206, 1992.

Jorrit J Hornberg, Frank J Bruggeman, Hans V Westerhoff, and Jan Lankelma. Cancer: a systems biology disease. *Biosystems*, 83(2-3):81–90, 2006.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

Sui Huang, Ingemar Ernberg, and Stuart Kauffman. Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective. 20(7):869–876, 2009.

Guido W Imbens. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710, 2000.

Francesco Iorio, Theo A Knijnenburg, Daniel J Vis, Graham R Bignell, Michael P Menden, Michael Schubert, Nanne Aben, Emanuel Gonçalves, Syd Barthorpe, Howard Lightfoot, et al. A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3):740–754, 2016.

Hemant Ishwaran. Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, 1:519–537, 2007.

François Jacob and Jacques Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology*, 3(3):318–356, 1961.

Holly Janes, Marshall D Brown, Ying Huang, and Margaret S Pepe. An approach to evaluating and comparing biomarkers for patient treatment selection. *The international journal of biostatistics*, 10(1):99–121, 2014.

Katarzyna Jastrzebski, Bram Thijssen, Roelof JC Kluin, Klaas de Lint, Ian J Majewski, Roderick L Beijersbergen, and Lodewyk FA Wessels. Integrative modeling identifies key determinants of inhibitor sensitivity in breast cancer cell lines. *Cancer research*, 78(15):4396–4410, 2018.

Ahmedin Jemal, Elizabeth M Ward, Christopher J Johnson, Kathleen A Cronin, Jiemin Ma, A Blythe Ryerson, Angela Mariotto, Andrew J Lake, Reda Wilson, Recinda L Sherman, et al. Annual report to the nation on the status of cancer, 1975–2014, featuring survival. *JNCI: Journal of the National Cancer Institute*, 109(9):djx030, 2017.

Siân Jones, Xiaosong Zhang, D Williams Parsons, Jimmy Cheng-Ho Lin, Rebecca J Leary, Philipp Angenendt, Parminder Mankoo, Hannah Carter, Hirohiko Kamiyama, Antonio Jimeno, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *science*, 321(5897):1801–1806, 2008.

Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, 40(D1):D109–D114, 2012.

Stuart Kauffman. Homeostasis and differentiation in random genetic control networks. *Nature*, 224(5215):177–178, 1969.

Faiz M Khan, Stephan Marquardt, Shailendra K Gupta, Susanne Knoll, Ulf Schmitz, Alf Spitschak, David Engelmann, Julio Vera, Olaf Wolkenhauer, and Brigitte M Pützer. Unraveling a tumor type-specific regulatory core underlying e2f1-mediated epithelial-mesenchymal transition to predict receptor protein signatures. *Nature communications*, 8(1):1–15, 2017.

Yongsoo Kim, Tycho Bismeijer, Wilbert Zwart, Lodewyk FA Wessels, and Daniel J Vis. Genomic data integration by won-parafac identifies interpretable factors for predicting drug-sensitivity in vivo. *Nature communications*, 10(1):1–12, 2019.

Paul Kirk, Thomas Thorne, and Michael PH Stumpf. Model selection in systems and synthetic biology. *Current opinion in biotechnology*, 24(4): 767–774, 2013.

Hiroaki Kitano. Computational systems biology. *Nature*, 420(6912):206–210, 2002.

Hannes Klarner, Adam Streck, and Heike Siebert. Pyboolnet: a python package for the generation, analysis and visualization of boolean networks. *Bioinformatics*, 33(5):770–772, 2016.

Bertram Klinger, Anja Sieber, Raphaela Fritsche-Guenther, Franziska Witzel, Leanne Berry, Dirk Schumacher, Yibing Yan, Pawel Durek, Mark Merchant, Reinhold Schäfer, et al. Network quantification of egfr signaling unveils potential for targeted combination therapy. *Molecular systems biology*, 9(1), 2013.

Theo A Knijnenburg, Gunnar W Klau, Francesco Iorio, Mathew J Garnett, Ultan McDermott, Ilya Shmulevich, and Lodewyk FA Wessels. Logic models to predict continuous outputs based on binary inputs with an application to personalized cancer therapy. *Scientific reports*, 6(1):1–14, 2016.

Alfred G Knudson. Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences*, 68(4):820–823, 1971.

Tarja Knuuttila and Andrea Loettgers. Modelling as indirect representation? the lotka–volterra model revisited. *The British Journal for the Philosophy of Science*, 68(4):1007–1036, 2017.

S Kopetz, J Desai, E Chan, JR Hecht, PJ O'dwyer, RJ Lee, KB Nolop, and L Saltz. Plx4032 in metastatic colorectal cancer patients with mutant braf tumors. *Journal of Clinical Oncology*, 28(15_suppl):3534–3534, 2010.

Pamela K Kreeger and Douglas A Lauffenburger. Cancer systems biology: a network modeling perspective. *Carcinogenesis*, 31(1):2–8, 2010.

Prateek Kumar, Steven Henikoff, and Pauline C. Ng. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 4(7):1073, June 2009. ISSN 1750-2799. doi: 10.1038/nprot.2009.86.

I Kuperstein, E Bonnet, HA Nguyen, D Cohen, E Viara, L Grieco, S Fourquet, L Calzone, C Russo, M Kondratova, et al. Atlas of cancer signalling network: a systems biology resource for integrative analysis of cancer data with google maps. *Oncogenesis*, 4(7):e160–e160, 2015.

Roman Kurilov, Benjamin Haibe-Kains, and Benedikt Brors. Assessment of modelling strategies for drug response prediction in cell lines and xenografts. *Scientific reports*, 10(1):1–11, 2020.

David Lake, Sonia AL Corrêa, and Jürgen Müller. Negative feedback regulation of the erk1/2 mapk pathway. *Cellular and Molecular Life Sciences*, 73(23):4397–4413, 2016.

Eric S Lander. Initial impact of the sequencing of the human genome. *Nature*, 470(7333):187–197, 2011.

Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. 2001.

Nicolas Le Novere. Quantitative and logic modelling of molecular and gene networks. *Nature Reviews Genetics*, 16(3):146–158, 2015.

Christophe Le Tourneau, Jean-Pierre Delord, Anthony Gonçalves, Céline Gavoille, Coraline Dubot, Nicolas Isambert, Mario Campone, Olivier Trédan, Marie-Ange Massiani, Cécile Mauborgne, et al. Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (shiva): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. *The lancet oncology*, 16(13):1324–1334, 2015.

Celine Lefebvre, Presha Rajbhandari, Mariano J. Alvarez, Pradeep Bandaru, Wei Keat Lim, Mai Sato, Kai Wang, Pavel Sumazin, Manjunath Kustagi, Brygida C. Bisikirska, Katia Basso, Pedro Beltrao, Nevan Krogan, Jean Gautier, Riccardo Dalla-Favera, and Andrea Califano. A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Molecular Systems Biology*, 6:377, June 2010. ISSN 1744-4292. doi: 10.1038/msb.2010.31.

Stefan Lehr, Jorg Kotzka, Haluk Avci, Albert Sickmann, Helmut E Meyer, Armin Herkner, and Dirk Muller-Wieland. Identification of major erk-related phosphorylation sites in gab1. *Biochemistry*, 43(38):12133–12140, 2004.

Gaelle Letort, Arnau Montagud, Gautier Stoll, Randy Heiland, Emmanuel Barillot, Paul Macklin, Andrei Zinovyev, and Laurence Calzone. Physiboss: a multi-scale agent-based modelling framework integrating physical dimension and cell signalling. *Bioinformatics*, 35(7):1188–1196, 2019.

Kun Li, Qunfeng Guo, Jun Yang, Hui Chen, Kewen Hu, Juan Zhao, Shunxin Zheng, Xiufeng Pang, Sufang Zhou, Yongyan Dang, et al. Foxd3 is a tumor suppressor of colon cancer by inhibiting egfr-ras-raf-mek-erk signal pathway. *Oncotarget*, 8(3):5048, 2017.

Richard Harold Lindeman. *Introduction to bivariate and multivariate analysis*. Glenview, Ill: Scott, Foresman, 1980.

Jianfang Liu, Tara Lichtenberg, Katherine A Hoadley, Laila M Poisson, Alexander J Lazar, Andrew D Cherniack, Albert J Kovatich, Christopher C Benz, Douglas A Levine, Adrian V Lee, et al. An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2):400–416, 2018.

AJ Lotka. Principles of physical biology. *Baltimore: Waverly*, 1925.

Miguel Angel Luque-Fernandez, Michael Schomaker, Bernard Rachet, and Mireille E Schnitzer. Targeted maximum likelihood estimation for a binary treatment: A tutorial. *Statistics in medicine*, 37(16):2530–2546, 2018.

Jianzhu Ma, Michael Ku Yu, Samson Fong, Keiichiro Ono, Eric Sage, Barry Demchak, Roded Sharan, and Trey Ideker. Using deep learning to model the hierarchical structure and function of a cell. *Nature methods*, 15(4): 290, 2018.

Matteo Manica, Ali Oskooei, Jannis Born, Vigneshwari Subramanian, Julio Sáez-Rodríguez, and María Rodríguez Martínez. Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders. *Molecular Pharmaceutics*, 2019.

José Luís Manzano, Laura Layos, Cristina Bugés, María de los Llanos Gil, Laia Vila, Eva Martinez-Balibrea, and Anna Martínez-Cardús. Resistant mechanisms to braf inhibitors in melanoma. *Annals of translational medicine*, 4(12), 2016.

Adam A. Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, 7(1):S7, March 2006. ISSN 1471-2105. doi: 10.1186/1471-2105-7-S1-S7.

Nick I Markevich, Jan B Hoek, and Boris N Kholodenko. Signaling switches and bistability arising from multisite phosphorylation in protein kinase cascades. *The Journal of cell biology*, 164(3):353–359, 2004.

Loredana Martignetti, Laurence Calzone, Eric Bonnet, Emmanuel Barillot, and Andrei Zinovyev. Roma: representation and quantification of module activity from target expression data. *Frontiers in genetics*, 7:18, 2016.

Bassirou Mboup, Paul Blanche, and Aurélien Latouche. On evaluating how well a biomarker can predict treatment response with survival data. *Pharmaceutical Statistics*, 2020.

Lisa M McShane, Douglas G Altman, Willi Sauerbrei, Sheila E Taube, Massimo Gion, and Gary M Clark. Reporting recommendations for tumor marker prognostic studies (remark). *Journal of the National Cancer Institute*, 97(16):1180–1184, 2005.

Nicolai Meinshausen, Alain Hauser, Joris M Mooij, Jonas Peters, Philip Versteeg, and Peter Bühlmann. Methods for causal inference from gene perturbation experiments and validation. *Proceedings of the National Academy of Sciences*, 113(27):7361–7368, 2016.

Craig H. Mermel, Steven E. Schumacher, Barbara Hill, Matthew L. Meyerson, Rameen Beroukhim, and Gad Getz. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, 12:R41, April 2011. ISSN 1474-760X. doi: 10.1186/gb-2011-12-4-r41.

Robin M Meyers, Jordan G Bryan, James M McFarland, Barbara A Weir, Ann E Sizemore, Han Xu, Neekesh V Dharia, Phillip G Montgomery, Glenn S Cowley, Sasha Pantel, et al. Computational correction of copy number effect improves specificity of crispr–cas9 essentiality screens in cancer cells. *Nature genetics*, 49(12):1779–1784, 2017.

Matthew Meyerson, Stacey Gabriel, and Gad Getz. Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics*, 11(10):685–696, 2010.

Iain Miller, Mingwei Min, Chen Yang, Chengzhe Tian, Sara Gookin, Dylan Carter, and Sabrina L Spencer. Ki67 is a graded rather than a binary marker of proliferation versus quiescence. *Cell reports*, 24(5):1105–1112, 2018.

Arnau Montagud, Pauline Traynard, Loredana Martignetti, Eric Bonnet, Emmanuel Barillot, Andrei Zinovyev, and Laurence Calzone. Conceptual and computational framework for logical modelling of biological networks deregulated in diseases. *Briefings in Bioinformatics*, 2017. doi: 10.1093/bib/bbx163.

Soufiane Mourragui, Marco Loog, Mark A Van De Wiel, Marcel JT Reinders, and Lodewyk FA Wessels. Precise: A domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors. *Bioinformatics*, 35(14):i510–i519, 2019.

Soufiane Mourragui, Marco Loog, Daniel J Vis, Kat Moore, Anna Gonzalez Manjon, Mark A van de Wiel, Marcel JT Reinders, and Lodewyk FA Wessels. Precise+ predicts drug response in patients by non-linear subspace-based transfer from cell lines and pdx models. *bioRxiv*, 2020.

Á C Murphy, Birgit Weyhenmeyer, Jasmin Schmid, Seán M Kilbride, Markus Rehm, Heinrich J Huber, C Senft, J Weissenberger, V Seifert, M Dunst, et al. Activation of executioner caspases is a predictor of progression-free survival in glioblastoma patients: a systems medicine approach. *Cell death & disease*, 4(5):e629–e629, 2013.

Christoph Müssel, Martin Hopfensitz, and Hans A Kestler. Boolnet?an r package for generation, reconstruction and analysis of boolean networks. *Bioinformatics*, 26(10):1378–1380, 2010.

Aurélien Naldi. Biolqm: a java toolkit for the manipulation and conversion of logical qualitative models of biological networks. *Frontiers in physiology*, 9:1605, 2018.

Aurélien Naldi, Céline Hernandez, Wassim Abou-Jaoudé, Pedro T Monteiro, Claudine Chaouiya, and Denis Thieffry. Logical modeling and analysis of cellular regulatory networks with ginsim 3.0. *Frontiers in physiology*, 9, 2018a.

Aurélien Naldi, Céline Hernandez, Nicolas Levy, Gautier Stoll, Pedro T Monteiro, Claudine Chaouiya, Tomáš Helikar, Andrei Zinovyev, Laurence Calzone, Sarah Cohen-Boulakia, et al. The colomoto interactive notebook: accessible and reproducible computational analyses for qualitative biological networks. *Frontiers in physiology*, 9:680, 2018b.

Nicholas E Navin. The first five years of single-cell cancer genomics and beyond. *Genome research*, 25(10):1499–1507, 2015.

Cancer Genome Atlas Research Network et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061, 2008.

Chiara Nicolò, Cynthia Périer, Melanie Prague, Carine Bellera, Gaëtan MacGrogan, Olivier Saut, and Sébastien Benzekry. Machine learning and

mechanistic modeling for prediction of metastatic relapse in early-stage breast cancer. *JCO Clinical Cancer Informatics*, 4:259–274, 2020.

Nicolas Le Novère, Michael Hucka, Huaiyu Mi, Stuart Moodie, Falk Schreiber, Anatoly Sorokin, Emek Demir, Katja Wegner, Mirit I. Aladjem, Sarala M. Wimalaratne, Frank T. Bergman, Ralph Gauges, Peter Ghazal, Hideya Kawaji, Lu Li, Yukiko Matsuoka, Alice Villeger, Sarah E. Boyd, Laurence Calzone, Melanie Courtot, Ugur Dogrusoz, Tom C. Freeman, Akira Funahashi, Samik Ghosh, Akiya Jouraku, Sohyoung Kim, Fedor Kolpakov, Augustin Luna, Sven Sahle, Esther Schmidt, Steven Watterson, Guanming Wu, Igor Goryanin, Douglas B. Kell, Chris Sander, Herbert Sauro, Jacky L. Snoep, Kurt Kohn, and Hiroaki Kitano. The Systems Biology Graphical Notation. *Nature Biotechnology*, 27(8):735–741, August 2009. ISSN 1546-1696. doi: 10.1038/nbt.1558.

Peter C Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976.

CNAM Oldenhuis, SF Oosting, JA Gietema, and EGE De Vries. Prognostic versus predictive value of biomarkers in oncology. *European Journal of Cancer*, 44(7):946–953, 2008.

M. Ostrowski, L. Paulevé, T. Schaub, A. Siegel, and C. Guziolowski. Boolean network identification from perturbation time series data combining dynamics abstraction and logic programming. *Biosystems*, 149: 139–153, November 2016. ISSN 0303-2647. doi: 10.1016/j.biosystems.2016.07.009.

Loïc Paulevé. Pint: A Static Analyzer for Transient Dynamics of Qualitative Networks with IPython Interface. In *CMSB 2017 - 15th conference on Computational Methods for Systems Biology*, volume 10545 of *Lecture Notes in Computer Science*, pages 370 – 316. Springer, 2017. doi: 10.1007/978-3-319-67471-1_20.

T Pawson and N Warner. Oncogenic re-wiring of cellular signaling pathways. *Oncogene*, 26(9):1268–1275, 2007.

Michael J Pencina, Ralph B D'Agostino Sr, Ralph B D'Agostino Jr, and Ramachandran S Vasan. Evaluating the added predictive ability of a new marker: from area under the roc curve to reclassification and beyond. *Statistics in medicine*, 27(2):157–172, 2008.

Margaret S Pepe, Ziding Feng, Ying Huang, Gary Longton, Ross Prentice, Ian M Thompson, and Yingye Zheng. Integrating the predictiveness

of a marker with its performance as a classifier. *American journal of epidemiology*, 167(3):362–368, 2008.

Margaret S Pepe, Holly Janes, and Christopher I Li. Net risk reclassification p values: valid or misleading? *Journal of the National Cancer Institute*, 106(4):dju041, 2014.

Bernard Pereira, Suet-Feung Chin, Oscar M. Rueda, Hans-Kristian Moen Vollan, Elena Provenzano, Helen A. Bardwell, Michelle Pugh, Linda Jones, Roslin Russell, Stephen-John Sammut, Dana W. Y. Tsui, Bin Liu, Sarah-Jane Dawson, Jean Abraham, Helen Northen, John F. Peden, Abhik Mukherjee, Gulisa Turashvili, Andrew R. Green, Steve McKinney, Arusha Oloumi, Sohrab Shah, Nitzan Rosenfeld, Leigh Murphy, David R. Bentley, Ian O. Ellis, Arnie Purushotham, Sarah E. Pinder, Anne-Lise Børresen-Dale, Helena M. Earl, Paul D. Pharoah, Mark T. Ross, Samuel Aparicio, and Carlos Caldas. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nature Communications*, 7, May 2016. ISSN 2041-1723. doi: 10.1038/ncomms11479.

Livia Perfetto, Leonardo Briganti, Alberto Calderone, Andrea Cerquone Perpetuini, Marta Iannuccelli, Francesca Langone, Luana Licata, Milica Marinkovic, Anna Mattioni, Theodora Pavlidou, et al. Signor: a database of causal relationships between biological entities. *Nucleic acids research*, 44(D1):D548–D554, 2016.

Charles M Perou, Stefanie S Jeffrey, Matt Van De Rijn, Christian A Rees, Michael B Eisen, Douglas T Ross, Alexander Pergamenschikov, Cheryl F Williams, Shirley X Zhu, Jeffrey CF Lee, et al. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proceedings of the National Academy of Sciences*, 96(16):9212–9217, 1999.

Charles M Perou, Therese Sørlie, Michael B Eisen, Matt Van De Rijn, Stefanie S Jeffrey, Christian A Rees, Jonathan R Pollack, Douglas T Ross, Hilde Johnsen, Lars A Akslen, et al. Molecular portraits of human breast tumours. *nature*, 406(6797):747–752, 2000.

Athanasios Polynikis, SJ Hogan, and Mario di Bernardo. Comparing different ode modelling approaches for gene regulatory networks. *Journal of theoretical biology*, 261(4):511–530, 2009.

Angela Potochnik. *Idealization and the Aims of Science*. University of Chicago Press, 2017.

Poulikos I Poulikakos, Yogindra Persaud, Manickam Janakiraman, Xiangju Kong, Charles Ng, Gatien Moriceau, Hubing Shi, Mohammad Atefi, Bjoern Titz, May Tal Gabay, et al. Raf inhibitor resistance is mediated by dimerization of aberrantly spliced braf (v600e). *Nature*, 480(7377):387, 2011.

Gibin G Powathil, Maciej Swat, and Mark AJ Chaplain. Systems oncology: towards patient-specific treatment regimes informed by multiscale mathematical modelling. In *Seminars in cancer biology*, volume 30, pages 13–20. Elsevier, 2015.

Anirudh Prahallad, Chong Sun, Sidong Huang, Federica Di Nicolantonio, Ramon Salazar, Davide Zecchin, Roderick L Beijersbergen, Alberto Bardelli, and René Bernards. Unresponsiveness of colon cancer to braf (v600e) inhibition through feedback activation of egfr. *Nature*, 483(7387): 100–103, 2012.

Misbah Razzaq, Loïc Paulevé, Anne Siegel, Julio Saez-Rodriguez, Jérémie Bourdon, and Carito Guziolowski. Computational discovery of dynamic cell line specific boolean networks from multiplex time-course data. *PLoS computational biology*, 14(10):e1006538, 2018.

E Premkumar Reddy, Roberta K Reynolds, Eugenio Santos, and Mariano Barbacid. A point mutation is responsible for the acquisition of transforming properties by the t24 human bladder carcinoma oncogene. *Nature*, 300 (5888):149–152, 1982.

Elisabeth Remy, Sandra Rebouissou, Claudine Chaouiya, Andrei Zinovyev, François Radvanyi, and Laurence Calzone. A modeling approach to explain mutually exclusive and co-occurring genetic alterations in bladder tumorigenesis. *Cancer research*, 75(19):4042–4052, 2015.

Jason A Reuter, Damek V Spacek, and Michael P Snyder. High-throughput sequencing technologies. *Molecular cell*, 58(4):586–597, 2015.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

Jordi Rodon, Jean-Charles Soria, Raanan Berger, Wilson H Miller, Eitan Rubin, Aleksandra Kugel, Apostolia Tsimberidou, Pierre Saintigny, Aliza Ackerstein, Irene Braña, et al. Genomic and transcriptomic profiling expands precision cancer medicine: the winther trial. *Nature medicine*, 25(5):751, 2019.

Arturo Rosenblueth and Norbert Wiener. The role of models in science. *Philosophy of science*, 12(4):316–321, 1945.

Alessandro Rossi, Michela Roberto, Martina Panebianco, Andrea Botticelli, Federica Mazzuca, and Paolo Marchetti. Drug resistance of braf-mutant melanoma: Review of up-to-date mechanisms of action and promising targeted agents. *European journal of pharmacology*, page 172621, 2019.

Patrick Royston. Explained variation for survival models. *The Stata Journal*, 6(1):83–96, 2006.

Patrick Royston and Willi Sauerbrei. A new measure of prognostic separation in survival data. *Statistics in medicine*, 23(5):723–748, 2004.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5): 688, 1974.

Donald B Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980.

Julio Saez-Rodriguez and Nils Blüthgen. Personalized signaling models for personalized treatments. *Molecular Systems Biology*, 16(1), 2020.

Julio Saez-Rodriguez, Leonidas G Alexopoulos, MingSheng Zhang, Melody K Morris, Douglas A Lauffenburger, and Peter K Sorger. Comparing signaling networks between normal and transformed hepatocytes using discrete logical models. *Cancer research*, 71(16):5400–5411, 2011a.

Julio Saez-Rodriguez, Leonidas G. Alexopoulos, MingSheng Zhang, Melody K. Morris, Douglas A. Lauffenburger, and Peter K. Sorger. Comparing Signaling Networks between Normal and Transformed Hepatocytes Using Discrete Logical Models. *Cancer Research*, 71(16):5400–5411, August 2011b. ISSN 0008-5472, 1538-7445. doi: 10.1158/0008-5472.CAN-10-4453.

Özgür Sahin, Holger Fröhlich, Christian Löbke, Ulrike Korf, Sara Burmester, Meher Majety, Jens Mattern, Ingo Schupp, Claudine Chaouiya, Denis Thieffry, et al. Modeling erbb receptor-regulated g1/s transition to find novel targets for de novo trastuzumab resistance. *BMC systems biology*, 3(1):1, 2009.

Manuela Salvucci, Maximilian L Würstle, Clare Morgan, Sarah Curry, Mattia Cremona, Andreas U Lindner, Orna Bacon, Alexa J Resler, Aine C Murphy, Robert O'Byrne, et al. A stepwise integrated approach to personalized risk predictions in stage iii colorectal cancer. *Clinical Cancer Research*, 23(5):1200–1212, 2017.

Manuela Salvucci, Arman Rahman, Alexa J Resler, Girish M Udupi, Deborah A McNamara, Elaine W Kay, Pierre Laurent-Puig, Daniel B Longley, Patrick G Johnston, Mark Lawler, et al. A machine learning platform to optimize the translation of personalized network models to the clinic. *JCO clinical cancer informatics*, 3:1–17, 2019a.

Manuela Salvucci, Zaitun Zakaria, Steven Carberry, Amanda Tivnan, Volker Seifert, Donat Kögel, Brona M Murphy, and Jochen HM Prehn. System-based approaches as prognostic tools for glioblastoma. *BMC cancer*, 19(1):1092, 2019b.

Yardena Samuels, Zhenghe Wang, Alberto Bardelli, Natalie Silliman, Janine Ptak, Steve Szabo, Hai Yan, Adi Gazdar, Steven M Powell, Gregory J Riggins, et al. High frequency of mutations of the pik3ca gene in human cancers. *Science*, 304(5670):554–554, 2004.

Francisco Sanchez-Vega, Marco Mina, Joshua Armenia, Walid K Chatila, Augustin Luna, Konnor C La, Sofia Dimitriadoy, David L Liu, Havish S Kantheti, Sadegh Saghafinia, et al. Oncogenic signaling pathways in the cancer genome atlas. *Cell*, 173(2):321–337, 2018.

Daniel J Sargent, Barbara A Conley, Carmen Allegra, and Laurence Collette. Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of Clinical Oncology*, 23(9):2020–2027, 2005.

Willi Sauerbrei, Sheila E Taube, Lisa M McShane, Margaret M Cavenagh, and Douglas G Altman. Reporting recommendations for tumor marker prognostic studies (remark): an abridged explanation and elaboration. *JNCI: Journal of the National Cancer Institute*, 110(8):803–811, 2018.

Charles L Sawyers. The cancer biomarker problem. *Nature*, 452(7187): 548–552, 2008.

Michael Schemper. Predictive accuracy and explained variation. *Statistics in medicine*, 22(14):2299–2308, 2003.

Kjetil Søreide. Receiver-operating characteristic curve analysis in diagnostic, prognostic and predictive biomarker research. *Journal of clinical pathology*, 62(1):1–5, 2009.

Zsofia K Stadler, Kasmintan A Schrader, Joseph Vijai, Mark E Robson, and Kenneth Offit. Cancer genomics and inherited risk. *Journal of Clinical Oncology*, 32(7):687, 2014.

Steven Nathaniel Steinway, Jorge Gomez Tejeda Zañudo, Paul J Michel, David J Feith, Thomas P Loughran, and Reka Albert. Combinatorial interventions inhibit tgf$\beta$-driven epithelial-to-mesenchymal transition and support hybrid cellular phenotypes. *NPJ systems biology and applications*, 1:15014, 2015.

Gautier Stoll, Eric Viara, Emmanuel Barillot, and Laurence Calzone. Continuous time boolean modeling for biological signaling: application of gillespie algorithm. *BMC systems biology*, 6(1):116, 2012.

Gautier Stoll, Barthélémy Caron, Eric Viara, Aurélien Dugourd, Andrei Zinovyev, Aurélien Naldi, Guido Kroemer, Emmanuel Barillot, and Laurence Calzone. Maboss 2.0: an environment for stochastic boolean modeling. *Bioinformatics*, 33(14):2226–2228, 2017.

Michael R Stratton, Peter J Campbell, and P Andrew Futreal. The cancer genome. *Nature*, 458(7239):719–724, 2009.

Chong Sun, Liqin Wang, Sidong Huang, Guus JJE Heynen, Anirudh Prahallad, Caroline Robert, John Haanen, Christian Blank, Jelle Wesseling, Stefan M Willems, et al. Reversible and adaptive resistance to braf (v600e) inhibition in melanoma. *Nature*, 508(7494):118–122, 2014.

Kristin R Swanson, Carly Bridge, JD Murray, and Ellsworth C Alvord Jr. Virtual and real brain tumors: using mathematical modeling to quantify glioma growth and invasion. *Journal of the neurological sciences*, 216(1): 1–10, 2003.

TCGA et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61, 2012.

Camille Terfve, Thomas Cokelaer, David Henriques, Aidan MacNamara, Emanuel Goncalves, Melody K. Morris, Martijn van Iersel, Douglas A. Lauffenburger, and Julio Saez-Rodriguez. CellNOptR: a flexible toolkit to train protein signaling networks to data using multiple logic formalisms. *BMC Systems Biology*, 6(1):133, October 2012. ISSN 1752-0509. doi: 10.1186/1752-0509-6-133.

Camille D. A. Terfve, Edmund H. Wilkes, Pedro Casado, Pedro R. Cutillas, and Julio Saez-Rodriguez. Large-scale models of signal propagation in human cells derived from discovery phosphoproteomic data. *Nature Communications*, 6:8033, September 2015. ISSN 2041-1723. doi: 10.1038/ncomms9033.

Andrew E Teschendorff, Ali Naderi, Nuno L Barbosa-Morais, and Carlos Caldas. Pack: Profile analysis using clustering and kurtosis to find molecular classifiers in cancer. *Bioinformatics*, 22(18):2269–2275, 2006.

Patrick Therasse, Susan G Arbuck, Elizabeth A Eisenhauer, Jantien Wanders, Richard S Kaplan, Larry Rubinstein, Jaap Verweij, Martine Van Glabbeke, Allan T van Oosterom, Michaele C Christian, et al. New guidelines to evaluate the response to treatment in solid tumors. *Journal of the National Cancer Institute*, 92(3):205–216, 2000.

René Thomas. Boolean formalization of genetic control circuits. *Journal of theoretical biology*, 42(3):563–585, 1973.

René Thomas and Richard d'Ari. *Biological feedback*. CRC press, 1990.

Collin J. Tokheim, Nickolas Papadopoulos, Kenneth W. Kinzler, Bert Vogelstein, and Rachel Karchin. Evaluating the evaluation of cancer driver genes. *Proceedings of the National Academy of Sciences*, 113 (50):14330–14335, December 2016. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1616440113.

Cristian Tomasetti and Bert Vogelstein. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, 347(6217):78–81, 2015.

Cristian Tomasetti, Luigi Marchionni, Martin A Nowak, Giovanni Parmigiani, and Bert Vogelstein. Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proceedings of the National Academy of Sciences*, 112(1):118–123, 2015.

Scott A Tomlins, Daniel R Rhodes, Sven Perner, Saravana M Dhanasekaran, Rohit Mehra, Xiao-Wei Sun, Sooryanarayana Varambally, Xuhong Cao, Joelle Tchinda, Rainer Kuefer, et al. Recurrent fusion of tmprss2 and ets transcription factor genes in prostate cancer. *science*, 310(5748):644–648, 2005.

Christophe Trefois, Paul MA Antony, Jorge Goncalves, Alexander Skupin, and Rudi Balling. Critical transitions in chronic disease: transferring concepts from ecology to systems medicine. *Current opinion in biotechnology*, 34:48–55, 2015.

John J Tyson, Katherine C Chen, and Bela Novak. Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Current opinion in cell biology*, 15(2):221–231, 2003.

John J Tyson, William T Baumann, Chun Chen, Anael Verdugo, Iman Tavassoly, Yue Wang, Louis M Weiner, and Robert Clarke. Dynamic modelling of oestrogen signalling and cell fate in breast cancer cells. *Nature Reviews Cancer*, 11(7):523–532, 2011.

Carling Ursem, Chloe E Atreya, and Katherine Van Loon. Emerging treatment options for braf-mutant colorectal cancer. *Gastrointestinal cancer: targets and therapy*, 8:13, 2018.

Mark J Van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data.* Springer Science & Business Media, 2011.

Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

Dieudonne van der Meer, Syd Barthorpe, Wanjuan Yang, Howard Lightfoot, Caitlin Hall, James Gilbert, Hayley E Francies, and Mathew J Garnett. Cell model passports—a hub for clinical, genetic and functional datasets of preclinical cancer models. *Nucleic acids research*, 47(D1):D923–D929, 2019.

NG Van Kampen. Stochastic processes in physics and chemistry. 5th, 2004.

Tyler J VanderWeele. Concerning the consistency assumption in causal inference. *Epidemiology*, 20(6):880–883, 2009.

Tyler J VanderWeele and Miguel A Hernan. Causal inference under multiple versions of treatment. *Journal of causal inference*, 1(1):1–20, 2013.

Laura J Van't Veer, Hongyue Dai, Marc J Van De Vijver, Yudong D He, Augustinus AM Hart, Mao Mao, Hans L Peterse, Karin Van Der Kooy, Matthew J Marton, Anke T Witteveen, et al. Gene expression profiling predicts clinical outcome of breast cancer. *nature*, 415(6871):530–536, 2002.

David Venet, Jacques E Dumont, Vincent Detours, et al. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput Biol*, 7(10):e1002240, 2011.

J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *science*, 291(5507):1304–1351, 2001.

Loic Verlingue, Laurence Calzone, Maud Kamal, Nicolas Servant, Lisa Belin, Emmanuel Barillot, and Christophe Le Tourneau. In silico prediction of the clinical response to the mtor inhibitor everolimus using a boolean model: validation from a cohort of the shiva trial, 2016a.

Loic Verlingue, Aurélien Dugourd, Gautier Stoll, Emmanuel Barillot, Laurence Calzone, and Arturo Londoño-Vallejo. A comprehensive approach to the molecular determinants of lifespan using a boolean model of geroconversion. *Aging Cell*, 15(6):1018–1026, 2016b.

Louis Verny, Nadir Sella, Séverine Affeldt, Param Priya Singh, and Hervé Isambert. Learning causal networks with latent variables from multivariate information in genomic data. *PLOS Computational Biology*, 13 (10):e1005662, October 2017. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005662.

Jean-Philippe Vert, Jian Qiu, and William S. Noble. A new pairwise kernel for biological network inference with support vector machines. *BMC Bioinformatics*, 8(Suppl 10):S8, 2007.

Santiago Videla, Julio Saez-Rodriguez, Carito Guziolowski, and Anne Siegel. caspo: a toolbox for automated reasoning on the response of logical signaling networks families. *Bioinformatics*, 33(6):947–950, March 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw738.

Alejandro F Villaverde and Julio R Banga. Reverse engineering and identification in systems biology: strategies, perspectives and challenges. *Journal of the Royal Society Interface*, 11(91):20130505, 2014.

Daniel J Vis, Lorenzo Bombardelli, Howard Lightfoot, Francesco Iorio, Mathew J Garnett, and Lodewyk FA Wessels. Multilevel models improve precision and speed of ic50 estimates. *Pharmacogenomics*, 17(7): 691–700, 2016.

Vito Volterra. Fluctuations in the abundance of a species considered mathematically, 1926.

Daniel D Von Hoff, Joseph J Stephenson Jr, Peter Rosen, David M Loesch, Mitesh J Borad, Stephen Anthony, Gayle Jameson, Susan Brown, Nina Cantafio, Donald A Richards, et al. Pilot study using molecular profiling of patients' tumors to find potential targets and select treatments for their refractory cancers. *Journal of clinical oncology*, 28(33):4877–4883, 2010.

Emily A Vucic, Kelsie L Thu, Keith Robison, Leszek A Rybaczyk, Raj Chari, Carlos E Alvarez, and Wan L Lam. Translating cancer 'omics' to improved outcomes. *Genome research*, 22(2):188–195, 2012.

Frederick YM Wan. *Mathematical models and their analysis*, volume 79. SIAM, 2018.

Jing Wang, Sijin Wen, W. Fraser Symmans, Lajos Pusztai, and Kevin R. Coombes. The Bimodality Index: A Criterion for Discovering and Ranking Bimodal Signatures from Cancer Gene Expression Profiling Data. *Cancer Informatics*, 7:199–216, August 2009. ISSN 1176-9351.

James D Watson and Francis HC Crick. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

Robert A Weinberg. *The biology of cancer*. Garland science, 2013.

Nathan Weinstein, Luis Mendoza, Isidoro Gitler, and Jaime Klapp. A network model to explore the effect of the micro-environment on endothelial cell behavior during angiogenesis. *Frontiers in physiology*, 8:960, 2017.

Michael Weisberg. Three kinds of idealization. *The journal of Philosophy*, 104(12):639–659, 2007.

Claudia Wellbrock, Maria Karasarides, and Richard Marais. The raf proteins take centre stage. *Nature reviews Molecular cell biology*, 5(11): 875–885, 2004.

Ulrike Wittig, Renate Kania, Martin Golebiewski, Maja Rey, Lei Shi, Lenneke Jong, Enkhjargal Algaa, Andreas Weidemann, Heidrun Sauer-Danzwith, Saqib Mir, et al. Sabio-rk—database for biochemical reaction kinetics. *Nucleic acids research*, 40(D1):D790–D796, 2012.

Steven Woodhouse, Nir Piterman, Christoph M Wintersteiger, Berthold Göttgens, and Jasmin Fisher. Scns: a graphical tool for reconstructing executable regulatory networks from single-cell genomic data. *BMC systems biology*, 12(1):59, 2018.

David Wroblewski, Branka Mijatov, Nethia Mohana-Kumaran, Fritz Lai, Stuart J Gallagher, Nikolas K Haass, Xu Dong Zhang, and Peter Hersey. The bh3-mimetic abt-737 sensitizes human melanoma cells to apoptosis induced by selective braf inhibitors but does not reverse acquired resistance. *Carcinogenesis*, 34(2):237–247, 2013.

Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, Dave Beare, James A Smith, I Richard Thompson, et al. Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research*, 41(D1):D955–D961, 2012.

Jorge Gómez Tejeda Zañudo, Maurizio Scaltriti, and Réka Albert. A network modeling approach to elucidate drug resistance mechanisms and predict combinatorial drug treatments in breast cancer. *Cancer convergence*, 1(1):5, 2017.

## RÉSUMÉ

Au delà de ses mécanismes génétiques, le cancer peut être compris comme une maladie de réseaux qui résulte souvent de l'interaction entre différentes perturbations dans un réseau de régulation cellulaire. La dynamique de ces réseaux et des voies de signalisation associées est complexe et requiert des approches intégrées. Une d'entre elles est la conception de modèles dits mécanistiques qui traduisent mathématiquement la connaissance biologique des réseaux afin de pouvoir simuler le fonctionnement moléculaire des cancers informatiquement. Ces modèles ne traduisent cependant que les mécanismes généraux à l'oeuvre dans certains cancers en particulier.
Cette thèse propose en premier lieu de définir des modèles mécanistiques personnalisés de cancer. Un modèle générique est d'abord défini dans un formalisme logique (ou Booléen), avant d'utiliser les données omiques (mutations, ARN, protéines) de patients ou de lignées cellulaires afin de rendre le modèle spécifique à chacun. Ces modèles personnalisés peuvent ensuite être confrontés aux données cliniques de patients pour vérifier leur validité. Le cas de la réponse clinique aux traitements est exploré en particulier dans ce travail. La représentation explicite des mécanismes moléculaires par ces modèles permet en effet de simuler l'effet de différents traitements suivant leur mode d'action et de vérifier si la sensibilité d'un patient à un traitement est bien prédite par le modèle personnalisé correspondant. Un exemple concernant la réponse aux inhibiteurs de BRAF dans les mélanomes et cancers colorectaux est ainsi proposé.
La confrontation des modèles mécanistiques de cancer, ceux présentés dans cette thèse et d'autres, aux données cliniques incite par ailleurs à évaluer rigoureusement leurs éventuels bénéfices dans la cadre d'une utilisation médicale. La quantification et l'interprétation de la valeur pronostique des biomarqueurs issus de certains modèles méchanistiques est brièvement présentée avant de se focaliser sur le cas particulier des modèles capables de sélectionner le meilleur traitement pour chaque patient en fonction des ses caractéristiques moléculaires. Un cadre théorique est proposé pour étendre les méthodes d'inférence causale à l'évaluation de tels algorithmes de médecine de précision. Une illustration est fournie à l'aide de données simulées et de xénogreffes dérivées de patients.
L'ensemble des méthodes et applications décrites tracent donc un chemin, de la conception de modèles mécanistiques de cancer à leur évaluation grâce à des modèles statistiques émulant des essais cliniques, proposant ainsi un cadre pour la mise en oeuvre de la médecine de précision en oncologie.

## MOTS CLÉS

Modélisation, Cancer, Modèle mécanistique, Biostatistiques, Inférence causale, Médecine de précision.

## ABSTRACT

Beyond its genetic mechanisms, cancer can be understood as a network disease that often results from the interactions between different perturbations in a cellular regulatory network. The dynamics of these networks and associated signaling pathways are complex and require integrated approaches. One approach is to design mechanistic models that translate the biological knowledge of networks in mathematical terms to simulate computationally the molecular features of cancers. However, these models only reflect the general mechanisms at work in cancers.
This thesis proposes to define personalized mechanistic models of cancer. A generic model is first defined in a logical (or Boolean) formalism, before using omics data (mutations, RNA, proteins) from patients or cell lines in order to make the model specific to each one profile. These personalized models can then be compared with the clinical data of patients in order to validate them. The response to treatment is investigated in particular in this thesis. The explicit representation of the molecular mechanisms by these models allows to simulate the effect of different treatments according to their targets and to verify if the sensitivity of a patient to a drug is well predicted by the corresponding personalized model. An example concerning the response to BRAF inhibitors in melanomas and colorectal cancers is thus presented.
The comparison of mechanistic models of cancer, those presented in this thesis and others, with clinical data also encourages a rigorous evaluation of their possible benefits in the context of medical use. The quantification and interpretation of the prognostic value of outputs of some mechanistic models is briefly presented before focusing on the particular case of models able to recommend the best treatment for each patient according to his molecular profile. A theoretical framework is defined to extend causal inference methods to the evaluation of such precision medicine algorithms. An illustration is provided using simulated data and patient derived xenografts.
All the methods and applications put forward a possible path from the design of mechanistic models of cancer to their evaluation using statistical models emulating clinical trials. As such, this thesis provides one framework for the implementation of precision medicine in oncology.

## KEYWORDS

Modeling, Cancer, Mechanistic model, Biostatistics, Causal inference, Precision medicine.