Forthcoming in *European Journal for Philosophy of Science*

Complexity and integration. A philosophical analysis of how cancer complexity can be faced in the era of precision medicine

Giovanni Boniolo, giovanni.boniolo@unife.it
Raffaella Campaner, raffaella.campaner@unibo.it

Abstract

Complexity and integration are longstanding widely debated issues in philosophy of science and recent contributions have largely focused on biology and biomedicine. This paper specifically considers some methodological novelties in cancer research, motivated by various features of tumours as complex diseases, and shows how they encourage some rethinking of philosophical discourses on those topics. In particular, we discuss the *integrative cluster approach*, and analyse its potential in the epistemology of cancer. We suggest that, far from being *the* solution to tame cancer complexity, this approach offers a philosophically interesting new manner of considering
integration, and show how it can help addressing the apparent contrast between a pluralistic and a unitary account.

Keywords

Complexity; Integration; Cancer; Precision medicine

1. Introduction: philosophical issues in past and present cancer research

In 2014, R.A. Weinberg, one of the world's leading molecular oncologists, published a paper to celebrate 40 years of *Cell*, the prestigious journal that "publishes findings of unusual significance in any area of experimental biology". Weinberg's paper, titled 'Coming full circle. From endless complexity to simplicity and back again', is a synthetic but illuminating history of cancer research over the last 40 years. That research started from the "phenomenological chaos that the traditional cancer researchers had been accumulating from more than half a century", and moved to a molecular level with a reductionist approach based on the presupposition that finding out "simple molecular mechanisms" (p. 267) would suffice to win what, in 1971, former

US President R. Nixon called the *War on Cancer*[1]. Since then, journals devoted to cancer research have published an increasingly large number of articles dealing with such molecular mechanisms, which also undoubtedly spurred many philosophers to investigate their nature and role.

Together with most scholars in the field, Weinberg claimed that the reductionist approach resulted in remarkable steps forward in our knowledge of the details of what happens locally in certain cell compartments or in certain infracellular signalling lines. Notwithstanding its oversimplifications, the approach led to significant progress concerning the role of certain viruses in cancerogenesis, the function of oncogenes and oncosuppressors, the impact of genome and epigenome mutations, and many others. Unfortunately, however, the huge efforts of small and large laboratories over the world did not lead to victory in the cancer war, nor did they reach global knowledge of what cancer is, how it develops and how it can be defeated.

It is also worth recalling that a silent change occurred in the main cancer research centres about 15 years ago: the number of bioinformaticians drastically increased, especially because the new sequencing biotechnologies started producing enormous quantities of data at an unprecedented pace, with the immediate need to govern them and understand their meaning[2]. Further major changes gradually took place. New biotechnologies and the 'omics' involved opened the road to an acknowledgement that cancerogenesis and metastatic processes were not easily understandable and that a purely reductionist approach, despite all its positive aspects, could not lead to a genuine comprehension of their nature and behaviour. Greater awareness of the limits and gaps in scientific knowledge on the topic had a major impact on attitudes and expectations concerning treatments and their efficacy. It was soon realized that cancer is an extremely spatially, temporally and hierarchically complex disease, and that a new approach was needed: an approach that could not disregard the enormous databases made available by the deluge of data produced by laboratories.

Analysing this scenario, Weinberg's paper reaches a conclusion that, although rather worrying, is extremely interesting for the purpose of reflecting on complexity and its interpretation in cancer studies. Weinberg clearly stated: "We lack the conceptual paradigms and computational strategies for dealing with this complexity. And equally painful, we don't know how to integrate individual data sets, such as those deriving from cancer genome analyses, with other, equally important data sets, such as proteomics. This is most frustrating, since it is becoming increasingly apparent that a precise and truly useful understanding of the behaviour of

[1] See: https://www.cancer.gov/about-nci/legislative/history/national-cancer-act-1971.
[2] On the rise, growth and success of bioinformatics, especially with respect to the life sciences, see e.g. Perez-Iratxeta, Andrade-Navarro and Wren (2007); Ouzounis (2012); Mehmood, Sehar and Ahmad (2014); Ratti (2016).

individual cancer cells and the tumours that they form will only come once we are able to integrate and then distil these data. So, perhaps ironically, we have come full circle, beginning in a period when vast amounts of cancer research data yielded little insight into underlying mechanisms to a period (1980–2000) when a flurry of molecular and genetic research gave hope that cancer really could be understood through simple and logical reductionist thinking, and finally to our current dilemma. Once again, we can't really assimilate and interpret most of the data that we accumulate. How will all this play out?" (p. 271). That is the serious question raised by many biomedical researchers and highly relevant to grasp how understanding cancer has been evolving. In turn, changes in the understanding the natural history of cancer, especially due to increasing awareness of its heterogeneity in space and in time, affect the design of further studies and the assessment of the impact of different kinds of interventions. These issues are also highly relevant for philosophers of science and prompt some rethinking of a few crucial notions, as we will show.

Weinberg points out that both *conceptual paradigms* and *computational strategies* are needed. As philosophers, we are clearly concerned with the ways in which the former affects the latter, and vice versa. How the design of computational strategies depends on the underlying conceptual paradigms regarding the phenomenon under examination, and how the results can be evaluated as more or less adequate according to the paradigm assumed, are very important matters. Equally important is the analysis of how conceptual paradigms can be framed, in turn, by the computational strategies available. A second set of problems hinted at in Weinberg's quotation has to do with the *integration* of data sets: What exactly is to be integrated, and what does integration require in order to be successfully performed? On which background picture of cancer does integration build upon, and what does it achieve? Thirdly, once we realize that currently we "can't really assimilate and interpret most of the data that we accumulate", we are urged to rethink our conception of what "a precise and truly useful understanding" of the behaviour of tumours should be like, and what role integration might play in that respect. What integration exactly amounts to, how it can *de facto* be performed, and its impact on a "truly useful understanding" to be reached for specific epistemic purposes (in particular, classification, prediction, intervention) are issues which need to be explored in depth.

All these questions stem from the recent history of cancer studies and have both a specific and a more general scope. On the one hand, they call for answers directly dealing with the pressing issues at stake – i.e. cancerogenesis and metastatic processes – given their possible implications for both future research lines and clinical approaches. On the other hand, they are of wider philosophical interest, insofar as they can encourage further reflections on integration

and complexity, both within philosophy of medicine with respect to other complex diseases (e.g. diabetes and psychiatric disorders[3]) and with respect to complex phenomena as addressed by other disciplinary fields. The conceptual understanding of integration, and of modes of data integration, has a direct impact on epistemological practices and outcomes, as causal explanations of disease and prediction of its evolving in time.

To tackle these epistemological concerns, we should not be as pessimistic as Weinberg seems to be. In the oncological field a number of papers have been appearing in biomedical journals proposing new ideas on integration – exactly as Weinberg suggested – to govern, even etiologically and prognostically, the complexity of cancer pathologies. These new perspectives focus on a different kind of integration and present us with a clutch of problems that we should be ready and conceptually equipped to discuss also in philosophical terms. We believe that novel approaches to integration in biomedicine, the meaning attributed to it and the uses to which integrative accounts are put in that context offer interesting challenges to philosophical reflections on the topic.

This paper addresses some of these pressing philosophical questions arising, as indicated, at the frontiers of biomedical research. In order to entertain a genuine and fruitful dialogue with scientists, a proper understanding of what is going on in science is needed. Only on this ground will philosophical work be able to grasp critical features of biomedical research and contribute to their theoretical disentanglement. It is in this spirit that what follows provides a short overview of the philosophical debate on integration as a means of addressing complexity (§ 2), followed by a sketchy outline of the state-of-the-art of research into tumour heterogeneity, biomarkers and stratification: all issues directly affecting our understanding of cancer and its epistemologically problematic features (§ 3). We then present a particular solution, the *integrative-cluster approach*, and analyse its potential in the epistemology of cancer (§ 4). In no way do we claim that this is *the* solution to tame cancer complexity, rather that this is *one possible* and interesting way to tame cancer complexity. Moreover, it is a solution that offers the opportunity to propose, also from a philosophical perspective, a new manner of considering integration and the apparent contrast between a pluralistic and a unitary account. This will be the specific scientific terrain on which our analysis will focus in the last section (§ 5). Certainly, complexity and integration have already been dealt with extensively in the philosophical literature[4], but the rapidly evolving situation in biomedical research and related methodologies demands renewed reflections on what exactly they amount to. More specifically, we will show

[3] See e.g. Lemoine 2017.
[4] See e.g. Mitchell (2003; 2009); Bechtel and Richardson (2010); Hooker (2011); Ladyman, Lambert, Wiesner (2013); Ladyman and Wiesner (forthcoming, 2018).

that the integration offered via the integrative-cluster approach allows us to see the notion of integration itself in a different way. In particular, we will argue that it allows for a unitary framework to address complex diseases, which considers both causal explanatory and predictive aspects. This kind of integration might therefore allow us to grasp the etiological and prognostic features of cancer, in a unitary and – obviously – probabilistic scenario.

2. On complexity and integration

There has been much talk about complexity and complex systems in philosophy and it is not possible to extensively review here all the positions. To pave the way for our analysis, we recall just some core claims allowing a better grasp of integration from an epistemological standpoint.

Although the literature provides no unanimous definition of complexity, accounts of what are deemed "complex systems" converge on a few aspects. Complex systems are constituted by a multiplicity of parts, belonging to several different levels and mutually interacting. Their relations are non-linear, and usually their behaviour is highly sensitive to initial conditions and emerges from the interactions among the parts by virtue of some self-organizing and hierarchical arrangement. Such behaviour does not just result from the aggregated behaviour of the parts involved, which cannot be inter-substituted: specific structural and functional organization of the system is crucial to its working, and established across what are deemed multiple levels (see e.g. Craver 2007, p. 135; Wimsatt 2007, pp. 280-281). Attempts to understand the complexity of natural phenomena have hence been accompanied by a rejection of divide-and-conquer strategies aiming to grasp phenomena by pursuing decompositions, studying parts of systems in isolation and neglecting contextual elements. The parts of a complex system are inter-dependent, and their behaviour is "co-determined by the system's organization" (Kaiser 2013, p. 260), and dependent on certain variations of contextual factors.

While a range of different perspectives and taxonomies of complexity is currently available (Wimsatt 2007, ch. 9; Mitchell 2009; Ladyman, Lambert and Wiesner 2013), it is a shared view that no single epistemic strategy will suffice to grasp it. Given multilevel structures, the heterogeneity of component parts, and high variability of complex systems, it seems that no unitary explanatory or predictive theory or model could be applied, in particular that there are no "simple, universal, and timeless underling laws to explain what there is and how it behaves" (Mitchell 2009, p. 11). A multiplicity of modelling practices and explanatory perspectives seem therefore to be needed to address complex systems, and "integration" is often advocated.

Unfortunately, consensus does not hold on the notion of integration either: there is neither a unique definition nor a unique form of integration, with philosophers of science proposing various versions, labelling them differently[5].

Different forms of integration have been discussed together with their different epistemological implications. It has been stressed that they can encourage, for instance, interactions between different disciplinary fields and cooperation between different lines of investigations and research communities. Moreover, integration can demand the combination of different accounts of the same phenomenon as studied within a single field, but analysed along different descriptive levels or investigated for different epistemic purposes: each account or level of analysis will claim to have some, but not all, relevant information for the construction of an account of the phenomenon at stake. Complexity dictates the pursuit of interactions on different problem domains and problem agendas, and draws upon a variety of tools to foster discussions on the criteria for model adequacy. Integration is usually presented as beneficial in the literature. It should be noted that it is not at odds with unification, but does not coincide with it: unificatory trends search actively for some comprehensive picture, while integration is mostly driven by a problem-oriented approach.

First and foremost, discourses on complexity are meant to stress the multiplicity and heterogeneity of variables involved in the representation of a single phenomenon or set of phenomena, the multiplicity of inter-related levels which are structurally and functionally organized, and the mutual constraints of the system's behaviour due to component parts and vice versa. Accordingly, reflections on complexity and integration have been debated as, amongst others, an "antidote" to reductionism: given the complexity of most biological phenomena, there is no single lowest-level theory from which multilevel knowledge from several different fields can be derived. As mentioned, a plurality of incompletely articulated, partly complementary and partly contradictory views are asked to interact. Such theoretical interactions will have to take place at different scales of components and across different spatial and temporal locations.

In what respects are such reflections on complexity relevant for cancer research and for the attempts to tame its heterogeneity in space and time, in order to propose therapeutic responses and predict patient outcomes? In §1 we stressed how, high expectations notwithstanding, the reductionist approach per se has not proved up to the job to master cancer's complexity. In §3 we will further show how cancer research must deal with: i) highly heterogeneous variables,

[5] On integration in biology, along a few of the different dimensions we have recalled, see, e.g., Leonelli (2008) and (2016) ch. 6; Brigandt (2010; 2013); O'Malley and Soyer (2012); VV. AA. (2013). On the possible benefits of different ways of conceiving integration, and possible epistemic trade-offs, see also Chang (2012), ch. 5, and Plutynski (2013).

acting at different biological levels and at different spatial and temporal scales, and whose organizational principles are still largely opaque; ii) huge amounts of available data; iii) the risk of fragmentation of classificatory practices and research lines. These issues undoubtedly impact on our prospects for an adequate causal explanation and successful prediction of its development. It should be stressed that it is not only an epistemological concern that should drive reflections on complexity in this context. Cancer complexity should first and foremost be tackled for both research and clinical purposes[6]. Rather than an exclusive focus on, for instance, some fine-grained understanding of possible different "shades of complexity", solutions must be envisaged to classify and treat cancer as effectively as possible, and integration plays an essential role in this respect.

As already mentioned above, complexity and the issue of integration have been largely addressed together with forms of pluralism and the need to have a range of multiple approaches, views, methods, standards, ... In the philosophical debate, pluralism has been mostly evaluated positively as bringing epistemic extra-value to the construction of scientific knowledge. Although a plurality of theoretical and/or clinical approaches is doubtless highly beneficial, and the importance, plausibility and usefulness of pluralistic attitudes cannot be neglected in the biomedical framework either, reaching integration in some unitary framework can have the merit of making complexity epistemically more tractable for the benefit of patients. Undoubtedly, if we have a unitary framework of the disease, in our case cancer, and if this unitary framework is tailored to a molecularly specific group of individuals, we may propose i) a more homogeneous treatment to any patient having that particular molecular characterisation, as precision medicine is indeed now attempting to do; ii) a first explanatory account of his/her disease; and iii) a prediction of its outcome.

What follows does not aim to suggest revisions or refinements to any extant philosophical taxonomy, but to analyse what scientific practice in cancer studies takes as the most pressing challenges, what strategies are being *de facto* devised to tackle them, and what epistemological implications can be drawn. The notion of integration we will discuss builds on computational tools as key resources to name clinically meaningful clusters: it is through the identification and clustering of sets of common (molecular, epidemiological, clinical, ...) features – rather than on some combination of a range of different, partial models, or on the search for underlying mechanisms – that cancer subtypes are identified and data integrated in single accounts. Each account is molecularly characterised and clinically tailored, and each one offers a unitary framework, with both explanatory and predictive significance. In this way, the individual patient

[6] On pathways to the clinic, see also Fagan (2017).

with his/her particular molecular and non-molecular profile (i.e., for example, his/her clinical profile or his/her lifestyle) can be both treated in the most accurate adequate manner (of course relatively to the coeval best clinical knowledge) and informed at least of the molecular causes of his/her pathology and of his/her possible future, by being inserted in the appropriate cluster. In other words, in integrative clustering each clusterisation is put forward as a tool to tackle biomedical complexity, as such able on its own to do so with no need to be complemented by/combined with other clusterisations. Far from having just theoretical import, integrative clustering promises to have an active impact on clinical treatments adopted for the single patient and, more in general, on the strategies and guidelines at population level. It is in these respects that the debate on understanding cancer complexity has become strictly interwoven with that concerning so-called precision medicine, as we will show below. This innovative scientific perspective also challenges philosophy of science, in particular approaches to integration and uses to which notions of integration are put, and affects modes of conceiving the natural history of cancer, as we will illustrate.

3. The state-of-the-art: cancer as a complex disease

Over the last few years, there has been much talk about 'precision medicine' and 'personalized medicine'. These locutions have been used to refer to different research strategies and communities, but all connected to the progress of molecular medicine, and related clinical expectations (See Boniolo and Nathan, 2017). The document of the *Precision Medicine Initiative*, launched in 2015 by former US President, Barack Obama,[7] states that precision medicine is an "innovative approach that takes into account individual differences in people's genes, environments, and lifestyles", and that it is about "delivering the right treatments, at the right time, every time to the right person", as Obama himself emphasized in an interview (Kaiser, 2015). Without much surprise, these words and concepts are more or less the same as those we find in the *definiens* of 'personalized medicine' in the position paper of *European Society for Predictive, Preventive and Personalised Medicine*.[8] Over and above different definitions of 'precision medicine' and 'personalized medicine', hereafter we accept the US National Research Council's suggestion to use 'precision medicine' especially when research is at issue. Of the different available interpretations of personalised medicine, we here take the term 'personalized' to imply that treatments and preventions are being advanced specifically for

[7] See, https://obamawhitehouse.archives.gov/precision-medicine (Accessed 30 April 2017). On this initiative, see, for example, Ashley (2015); Collins and Varmus (2015); Kohane, I.S. (2015); Sabatello and Appelbaum (2017). For a first hint on a philosophical analysis, see Tonelli and Shirts (2017).

[8] See the position paper of European Society for Predictive, Preventive and Personalised Medicine (EPMA) by Golubnitschaja et al. (2016).

a given individual[9]. Precision medicine is hence preferred as it focuses on identifying which approaches will be effective for which groups of patients on the basis of genetic, environmental, and lifestyle factors.[10] But what motivates such an approach from an oncological standpoint and what are the epistemological challenges it must face?

3.1. The quest for integration

Even a quick look at the huge amount of scientific, didactic and popular papers on precision medicine clearly reveals that it owes much to and is supported by the amazing advances of computational and information technologies (CIT) and biotechnologies over the last few years. Two biotechnology fields in particular, the new sequencing technologies and the new imaging technologies, have seen this major impact and have provided a decisive boost to precision medicine. Fig. 1 shows the state-of-the-art of sequencing technologies, and how the data analysis permitted by CIT and the integration of outcomes have become central points for research and clinics.

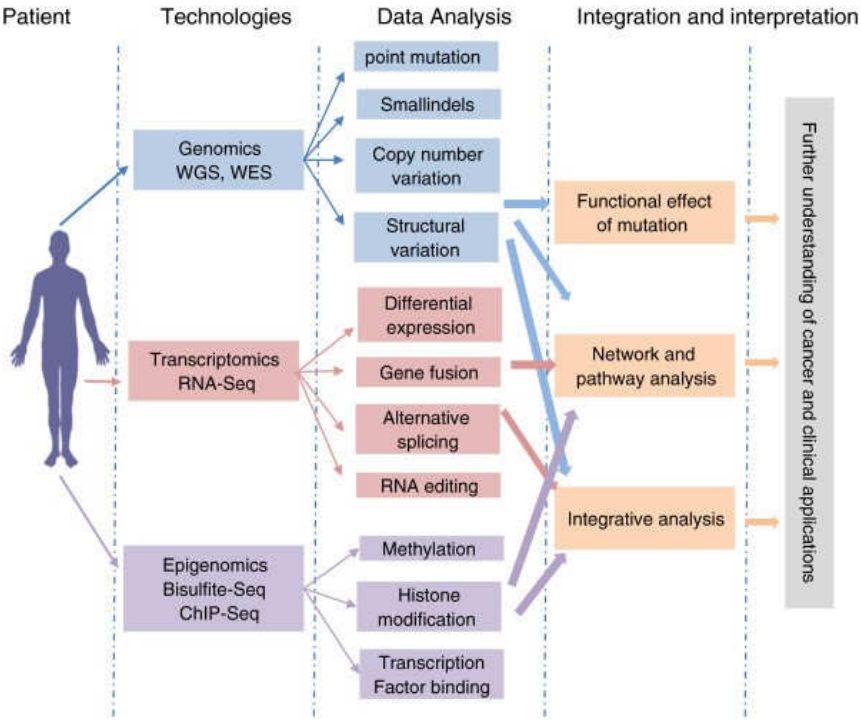

Fig. 1 (From Shyr and Liu, 2013). WGS = Whole Genome Sequencing; WES = Whole-Exome Sequencing; RNA-Seq = RNA Sequencing; Bisulfite-Seq = Bisulfite Sequencing; ChIP-Seq = Chromatin Immunoprecipitation Sequencing.

[9] See, https://ghr.nlm.nih.gov/ Precision Medicine; see also https://www.nih.gov/research-training/allofus-research-program (Accessed 30 April 2017).
[10] See, Nabipour and Assadi (2016); see also Zhang (2015).

The “further understanding of cancer” is under the heading of “integration and interpretation”. How exactly is integration to be interpreted here? *What* do we want to integrate, and *how*? The philosophical literature usually presents integration in the positive. Plutynski (2013), for instance, stresses how integration has become endowed with normative weight, carrying with it the underlying idea that the more integrative the scientific enterprise is, the better, the more holistic and closer to completeness the picture provided is. But what are the partial pictures to be integrated as in the scenario of current cancer research? What strategies adopted in that context are deemed “integrative”? Integration has also to do with the wider picture we have of the target phenomenon, depends on it and, in turn, affects it. Whether we believe integration should, in the end, prove how all different portions of knowledge conspire to create a single account of the phenomenon under investigation or whether, while interacting, they should remain alternative are sensitive issues which are likely to affect the philosophy of cancer as a complex disease also in the long run. They impact on our evaluation of the successfulness of forms of integration and on the uses we put them to. What follows addresses these problems in the light of a novel approach to cancer research, showing how it actually conceives of integration in an attempt to tame cancer complexity, stressing its possible epistemological consequences.

Fig.1 represents the not-yet fully realized integrative programme of comprehensive information flow starting from the patient's genome, transcriptome and epigenome to the clinical decision. It is a flow made possible by the data evaluation pathway with bioinformatics algorithms applied to genomics, transcriptomics and epigenomics sequencing technologies, which permit an analysis of the detected genomic mutations, genetic variants, differential gene expression, fusion transcripts, DNA methylations, transcription binding factors, etc. The figure also pictures the integration of protein expression information into appropriate genes and metabolic/functional networks, which ultimately facilitates mapping the framework for a personalized treatment strategy. Interestingly, the last column on the right hand side of the figure is headed “Integration and Interpretation”, with “Function effect of mutation”, “Network and pathway analysis” and “Integrative analysis” being called to converge into a larger, conceptually very wide, box indicating that “Further understanding of cancer and clinical applications” are to be reached. The steps through which knowledge shall shift from the penultimate to the ultimate column on the right hand side of the figure, thus obtaining an integration and proper interpretation for better understanding, are a core epistemological concern, worthy of deeper investigation.

Parallel to the massive advances of the biotechnologies in the field of 'omics' sequencing, imaging biotechnologies have also made impressive steps forward. If the discovery of x-rays more than a century ago profoundly changed the practice of medicine by enabling us to see inside the living body, molecular imaging is now probing deep inside the body to reveal its inner workings. And this is the other side of precision medicine. Unlike conventional imaging studies, which produce primarily structural pictures, molecular imaging visualizes how the body is functioning and what is occurring at the cellular and molecular levels. This has opened the door to a better understanding of the pathways of disease, the design of new drugs, improved therapeutic decision-making, and monitoring the patient's response to treatment. Molecular imaging allows non-invasive assessment, and quantification is especially desirable when following patients over time. Of course, to assess cellular function noninvasively, it is important to identify biomarkers that are specific to a disease or cellular process that we wish to measure[11]. Here again the question of integration springs up, no longer limited to integrating the "omics" results alone, or with the clinical information, but "omics" results and molecular imaging results, as shown in Fig. 2, to obtain the best diagnosis and therapy.

Here, we have two levels of complexity. On the one hand, there is a sort of ontological complexity related to features of a healthy or diseased human body. It is a complexity related to what we are discovering day by day, thanks to the increasingly powerful biotechnologies. Now we are beginning to understand the amazing number of molecules at play belonging to different "omic" levels and their mutual deterministic and, more often, probabilistic causal interactions, characterised by non-linearity, sensitivity to the initial conditions, sensitivity to the cellular and extracellular environmental conditions, sensitivity to temporal (i.e., developmental and evolutionary) and spatial (i.e., location in the body, in the organ, in the tissue) parameters, etc. On the other hand, there is data complexity, emerging from the application of sequencing and imaging biotechnologies. Now we have an increasingly large volume of data, concerning not only a single individual but sometimes thousands of individuals (see, e.g., Strasser, 2017; Leonelli, 2016). Such data have to be interpreted, integrated and thus governed, especially for clinical reasons and especially in the oncological field, in a unitary framework, which should be explanatory if possible, but especially predictive and not only classificatory. We should not forget that, from the specific point of view of the patient struggling against the disease, having several models describing his/her complex abnormal phenomenon is unlikely to be appreciated as beneficial in direct clinical terms. While pluralism can be very fruitful in theoretical terms, a physician, and even more so a patient, can actually feel disoriented by an array of models, and

[11] See e.g. Xue et al. (2013) and Pu et al. (2016a) and (2016b).

related alternative therapeutic choices and prognostic frameworks. What is important for the patient is to have something which reduces the range of alternative scenarios she is presented with, and which could quickly suggest the most effective therapy, given current medical knowledge, and a plausible prognosis.

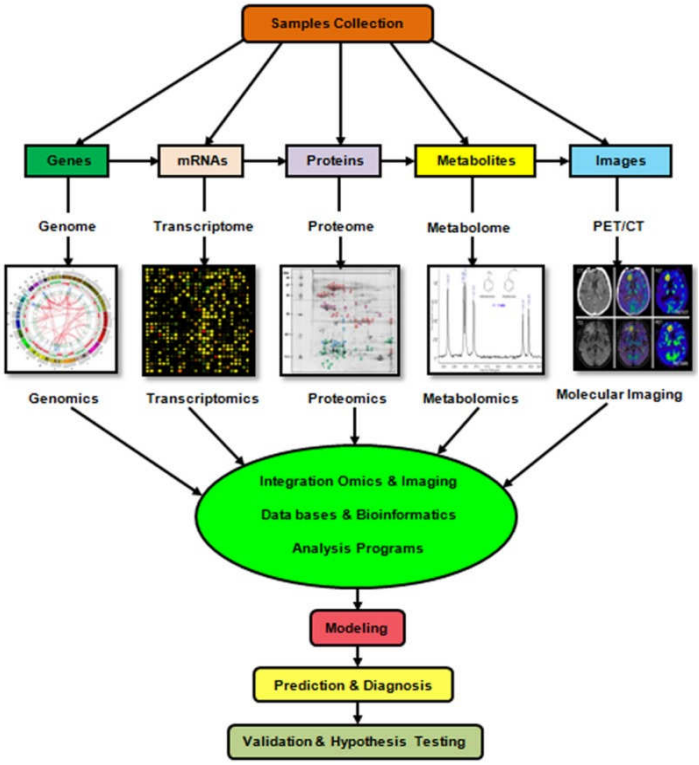

Fig. 2 (From Ghasemi et al., 2016).

3.2 Tumour heterogeneity

The drive towards the molecular analysis of oncological diseases, with the impressive advances of biotechnologies and CIT, has not only given rise to the precision medicine approach, but also made problematic aspects emerge more clearly. On the one hand, it has opened the door to enthusiastic expectations on the possibilities of curing severe pathologies, and on the other, to a deeper awareness of the difficulties in pursuing that goal. In particular, we have become aware of what is known as *tumour heterogeneity*. Tumour heterogeneity means not only that each cancer has to be individualised in a specific patient, but, more importantly, that each cancer affecting a given individual is actually composed of a set of different cancer subpopulations. That is, cells belonging to the same cancer show distinct genetic and phenotypic characteristics (such as gene expression, metabolism, motility, and angiogenic, proliferative, immunogenic, and

metastatic potential) in different space locations and in different time frames. This impressive complexity lies at the centre of an intensive biomedical research programme,[12] and is posing a huge challenge to medicine, and to precision medicine in particular. We have hence understood that any patient's cancer is a particular, specific disease, and that "many cancers" coexist in the same patient's cancer, each with its own histopathological and biological features.

Tumour heterogeneity is multifaceted, comprising: 1) *intertumour heterogeneity*, i.e., variability between tumours arising in the same organ, and *intratumour heterogeneity*, i.e. variability in the same individual tumour (see Burrell, 2013); 2) *spatial heterogeneity*, indicating that different regions of a tumour present different series of genetic aberrations, and *temporal heterogeneity*, referring to the course of disease progression (see Geyer et al., 2010; Torres et al., 2006; Martelotto et al., 2014). Heterogeneity within primary tumours is only one aspect. Cancer could also be thought of as a systemic disease[13]: over time, malignant cancers shed a large number of cells into the bloodstream and lymph vessels; some of these cells find a place in distant sites and develop into metastases. Therefore, to have a proper understanding of cancer heterogeneity we should also understand metastatic tumours, which, as is known, are the most fearsome, since they are responsible for the majority of cancer-related deaths.

Tumour heterogeneity means tumour complexity, and several different models have been proposed to address it, including mathematical models. What is interesting is that more or less any model which tries to cope with such complexity has borrowed its jargon from a range of different biological fields: evolutionary biology, developmental biology, population genetics, ecology, stem cell biology, etc. Unfortunately, none of these models is capable by itself of grasping the heterogeneity (complexity), but only certain aspects of the phenomena at stake (see Boniolo, 2017). Summing up, we are in a typical epistemic situation regarding complexity as tackled by the philosophers, as recalled in §2. However, patients cannot but be most interested in efficacy of treatments, rather than in ontological and/ or epistemic complexity.

In epistemological terms, tumour heterogeneity can be understood in a range of different ways. On the one hand, this raises issues to do with the need to design general models of pathologies, and on the other, with how single cases can be accounted for in clinics. Therefore, attempts to tame complexity in this context cannot neglect their practical, clinical implications. Again, we need a unitary framework that a tumour (a diseased patient) can be assigned to, after some choice on which specific features to ignore or diminish. Concerns related to heterogeneity and single cases are thus related also to the last aspect we shall touch upon to give a proper

[12] See the recent issue of *Nature* (VV. AA., 2013, issue 501) devoted to it.

[13] Note that there many different ways of thinking cancer that have been proposed along the years. No one is unanimously accepted by researchers and clinicians. For philosophical overviews of the different positions, see Bertolaso (2016) and Plutynski (forthcoming, 2018).

sense to the philosophical aspects stemming from state-of-the-art cancer research.

3.3 Biomarkers and stratifications

Biomarkers have played a crucial role with respect to the impact of new sequencing and imaging biotechnologies and the acknowledgment of tumour heterogeneity. In cancer, DNA-based biomarkers (SNPs, chromosomal aberrations, changes in DNA copy number, microsatellite instability, differential promoter-region methylation, etc.), RNA-based biomarkers (over or under-expressed transcripts, microRNAs, etc.) and protein biomarkers (cell-surface receptors, tumour antigens, phosphorylation states, tumour-released peptides into body fluids, etc.) are particularly important[14].

Precision medicine considers biomarkers crucial indicators when trying to answer questions such as: Who has or could have a disease? What is the actual or potential disease? Who could or should be treated, and with what? How could the patient react to the treatment? Biomarkers are taken as a fundamental key to most clinical matters – matters which are strongly affected by each patient's peculiar individual features and, at the same time, by the struggle to *classify* tumours. Speaking of biomarkers means speaking of stratifications among individuals on the basis of being or not being the carrier of one or more of them.

Classification practices in medicine have been changing over time. Patients have been classified on the basis of symptoms, or signs, or other characteristics. With the advent of molecular biology, classifications have gone deeper, to the molecular level. This means producing more precise stratifications of potential and actual patients, but also vastly enlarging their numbers by iteratively creating new ones. In a sense, delving into the molecular level potentially brings with it a fragmentation of possible classifications, and forces us to reflect on which molecular data we should actually focus to avoid too broad a proliferation. This worry has to do with the fact that a deluge of molecular data is produced daily by laboratories, and new problems arise as to how to manage them, in particular how to give them clinical significance and how to clinically validate them for preventive, predictive, diagnostic, prognostic and therapeutic purposes. Alongside enthusiasm for the availability of huge amounts of molecular data, we are called to ask whether they are equally significant, and easily manageable, for all the epistemic and clinical purposes just listed. Concerning validation, there is sort of 'contrapasso': the growing number of possible cancer biomarkers studied in the laboratory is associated with a shrinking number of them being clinically validated, either due to the cost of clinical validation,

[14] See, for example, Lee (2003); Bracht (2009); Koychev et al. (2011); Negm, Verma and Srivastava (2002); Vasan (2006).

or to the time required for the same[15].

To address these points in more detail, let us consider breast cancer classification. The current routine for breast cancer assessment usually comprises a clinical component, involving information gathered by imaging techniques, clinical examinations and biographic narrations, and a morpho-histopathological component, where an analysis is made of the tumour size, grade and lymph node status and the tests regarding the oestrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2).[16] Pathologists are then used to classifying breast cancer into four main subtypes: luminal A (usually ER+ and/or PR+, HER2-, with a low proliferation index); luminal B (ER+ and/or PR+ and high proliferation index); HER2-amplified (ER-, PR- and high levels of HER2); and basal-like cancer (the 'triple negative', i.e. negative with ER-, PR- and HER2-). However, it is now widely recognised that this grouping does not reliably predict how tumours will behave. One possible way out rests on inserting molecular profiling and adding molecular classification. But then we are required to decide *which* molecular classification we should rely on, and how information from the usual path and information from the molecular level are to be brought together. Shifting to the molecular level, in other words, does not per se warrant a single clear-cut classification suitable for clinical purposes.

Summing up, we do not have a single stratification accepted by the entire biomedical community, but several stratifications depending both on the set of biomarkers selected and identified and on their purpose. That is, we have a number of different classifications of diseased patients, each of which is grounded on bottoming out underlying features, but also depends on *a priori* decisions on which biomarkers (or set of biomarkers) are to be considered. Epistemological concerns thus clearly raise questions as to the grounds on which a biomarker (or set of biomarkers) is chosen as the most relevant, and how the purpose for which some biomarkers are identified influences such choice. The proliferation of stratifications is troublesome in a number of respects: at the preclinical level, since we have to fully understand their clinical significance and adoptability; at the clinical level, since we have to manage all of them in order to propose a diagnosis and a therapy; at a philosophical level too, since we are faced with different bio-ontologies, each connected with a different patient stratification and, thus, with a different research enterprise and focus on clinical level. How data belonging to different bio-ontologies can be integrated and which relations can be drawn between different stratifications and classificatory strategies are puzzling aspects.

[15] See, e.g., Goossens et al. (2015); Mordente et al. (2015); Scatena (2015).
[16] The receptor status is identified by immunohistochemistry, which stains the cells based on the presence (ER+, PR+, HER2+) or the absence (ER-, PR-, HER2-) of the receptor itself.

From a philosophical perspective, such a situation encourages discussion of which conclusions we should draw from the availability of several stratifications, and what should drive our choices in terms of taxonomies to be adopted. More in general, we should reflect over which relations hold between ongoing progress in methodologies to investigate diseases and the construction of nosographies – which tend to undergo iterative processes. Classifications are based on empirical detections, but they rely heavily on the methods of investigation and, therefore, on the technological innovations permitting them. Given the ways in which research is progressing, the deeper investigative methods allow us to go, the more stratifications we have. Validating these stratifications both from a research and a clinical standpoint is then problematic, as is establishing how "to make them talk to each other", since they deal with different levels (organs, tissues, cells, molecules) and can be constructed for different uses (biomedical basic and translational research, clinical practice) and epistemic and practical purposes (prevention, explanation, prediction, diagnosis, prognosis, therapy).

All the questions we have been posing in outlining various aspects of current cancer research are closely interrelated. The ways in which we describe cancer as a pathological condition – or, rather, as an array of possible pathological conditions – rely on the different kinds of evidence collected, the different technological resources allowing us to collect them, and on our capacity to integrate different portions of the information acquired. Different descriptions in turn affect classification practices, which impact on diagnoses and hence treatments, as well as tentative explanatory accounts. Explanations, again, affect the ways in which we describe and classify, how we decide which features are to be taken as relevant for the inception, progress and course of the disease we are considering, and how we shall, in the end, carve tumours out of the huge, impressive amount of available data. Descriptions, stratification and classification processes, explanatory and integrative strategies have been revolutionized by novel technologies devised to deal with impressive amounts of data.[17] The quest for integration put forward in research contexts has to do with the availability of different sets of data produced in different research, translational, or clinical contexts, with bio-ontologies, different methods and tools. A demand to integrate forms of diversity arises in research settings and is taken as the preliminary and necessary step to be taken if we want to take proper advantage of the results produced in one field and transpose them to another. What follows dwells on a specific methodology in cancer studies and analyses its import for theoretical approaches to cancer complexity, and for our very conception of how taxonomies of complex diseases can be built and the uses they can be put to.

[17] For some critical remarks over the use of big data, and on limits and drawback of big data science, see, e.g. Boyd and Crawford (2012); Leonelli (2014); Kitchin (2014); Coveney, Dougherty and Highfield (2016)

4. Complexity and integrative clustering

As shown, the advent of new sequencing and imaging biotechnologies has allowed deeper investigation at molecular level yielding what has been called *precision medicine*, but it has also opened the Pandora's box of cancer heterogeneity, enormous quantities of data, and many different ways of stratifying actual and potential cancers and patients. In short, we have *complexity and search for integration at different levels*. How should we tame it, or try to tame it? It seems that the password is *integration by means of CIT tools*.

Many different attempts to integrate data and methods have been advanced and papers on promising integrative approaches for precision medicine are published by leading scientific journals almost weekly. We focus here on *iCluster* (or *integrativeCluster*), an integrative approach which, although by no means the only option currently available[18], is extremely interesting also from a philosophical standpoint, insofar as it stimulates a range of conceptual considerations.

On the one hand, there is no single objective way to evaluate cluster analysis methods, since they depend on the problem-specific information, epistemic purposes and research programmes scientists are interested in. On the other hand, we should not conceive of cluster analysis as somehow arbitrary (Hennig 2015). According to one of the most popular textbooks on statistics, "cluster analysis, also called data segmentation, has a variety of goals. All relate to grouping or segmenting a collection of objects into subsets or 'clusters', such that those within each cluster are more closely related to one another than objects assigned to different clusters. An object can be described by a set of measurements, or by its relation to other objects. In addition, the goal is sometimes to arrange the clusters into a natural hierarchy. […] Central to all of the goals of cluster analysis is the notion of the degree of similarity (or dissimilarity) between the individual objects being clustered. A clustering method attempts to group the objects based on the definition of similarity supplied to it" (Hastie, Tibshirani and Friedman 2008, p. 501). This procedure raises the well-known problem of how to classify or categorize the elements under investigation, and how to order the classification or categorisation obtained, given some choice on the respects and degrees of similarity/dissimilarity. This problem, already encountered above touching on the issue of patient stratification, has been addressed throughout the history of philosophy (Boniolo 2007, Ch. 1**)**. We do not want here to address the metaphysical and ontological *vexata quaestio* of possible natural kinds in the life sciences or that of universals.

[18] A particularly interesting and successful approach proposal on tumour heterogeneity – specifically, in breast cancer - and modes of providing distinctive molecular portraits of each tumour is provided by Perou and Sorlie (see e.g. Perou, Sorlie et al 2000). As we discuss in the following, what characterizes iCluster with respect to this classification and similar ones is that while these latter are based on molecular features, the former is characterized by both molecular and clinical features.

Our purpose, rather, is to shed light on the ways in which the innovative methodologies leading to different ways of defining diseases, and carving them out of incredibly large amounts of data as working entities, *de facto* deeply question our classification practices. As we have seen when presenting strata realised via biomarkers, the epistemologically relevant point is that we *choose* which biomarkers to consider and then build the strata bottom-up, relying on our choices. Depending on the biomarker (or set of biomarkers) chosen, be it at genomic or proteomic, etc., level, we can have different classifications. The same applies to the computational cluster: *we choose* the similarities/dissimilarities on whose basis the clusters and their hierarchies have to be generated. Then, how they are generated and which hierarchies of clusters we obtain depend on the algorithm *we decide* to run over the data[19]. Our epistemic procedures, and evaluations of what counts as most relevant, play a fundamental role. Moreover, this all hints towards a general failure of a reductionist approach based on the focus on a particular mutated gene or protein, or on a particular molecular mechanism[20]. Summing up, algorithms running over huge amounts of data have taken the place of attempts to unravel networks of underlying mechanisms.

With respect to breast cancer, the idea underlying iCluster is to try to integrate databases at genomic and transcriptomic levels by means of *computational statistics*, namely cluster analysis. C. Caldas and his group[21] began analysing about 1,000 samples from breast cancer patients considered homogeneous for treatment purposes and which referred to a follow-up study of about 10 years, hence also having a lot of clinical information on what happened to the patients later on in time[22]. They made a genomic inquiry focused on hereditary characteristics, such as single nucleotide polymorphisms (SNP) and copy number variation (CNV), and on somatic characteristics (copy number alterations, CNA), and a transcriptomic inquiry on how hereditary characteristics could alter gene expression both at the same locus (cis action) and at different loci (trans action).[23] In order to understand the differences between normal and pathological

[19] Of course, this is not the right place to enter technical details on the statistical algorithms that are used. They could be easily retrieved in textbooks on cluster theory, or in the scientific papers adopting statistical models based on it. Our current focus is on the epistemological impact of the adoption of iCluster in dealing with cancer complexity, especially from a classificatory and a prognostic standpoint.

[20] Against the reductionism concerning molecular mechanisms and in favour of a more holist approach based on pathways, see Boniolo, Campaner (2018).

[21] https://www.cruk.cam.ac.uk/research-groups/caldas-group. See Curtis et al. (2012); Ali et al. (2014); Bruna et al. (2016); Pereira et al. (2016); Russnes et al. (2017).

[22] The group obtained about 1,000 frozen breast cancer samples from five tumor biobanks in the UK and Canada. It should be noted that "Nearly all oestrogen receptor (ER)-positive and/or lymph node (LN)-negative patients did not receive chemotherapy, whereas ER-negative and LN-positive patients did. Additionally, none of the HER21patients received trastuzumab. As such, the treatments were homogeneous with respect to clinically relevant groupings." (Curtis et al. 2012, p. 346).

[23] The SNPs are the most common type of genetic variation among people. Each SNP represents a difference in a single DNA nucleotide. CNV is a repetition of sections of the genome the number of repetition varies among people. CNA is a repetition of sections of the genome that has arisen in somatic tissue. *Cis module* and *trasn module* are stretches of DNA that affect the expression, respectively, of nearby and distant genes

conditions, they took blood samples from 500 healthy patients to identify when a copy number was not oncologically pathological. Then they used computer algorithms to search for clusters, based on similarities in CNV, SNP, CNA and gene expression correlated with clinical information. At the end of the computation process, they obtained ten different clusters they called *Integrative Clusters* (*iClusters*, or *IntClusters*).

The problem, at this point, was to validate the ten iClusters, to compare them with other molecular classifications and, most importantly, to understand their clinical validity. The scientists involved succeeded in this task by using a second cohort of about 1,000 breast cancer samples and a third cohort of about 7,500 samples. As shown in the table (Fig. 3), they successfully managed to compare their clusterisations both with other molecular characterisations (e.g., PAM50[24]) and with the clinical outcomes[25]. With respect to clinical matters, they also showed that their integrative classification reflected differences in chemotherapy. This might be seen as a good example of how to link molecular classification to clinical treatment, and to treatment outcomes. And it is this union, rather *this integration*, of molecular information and clinical information that characterises the iCluster approach and renders it different from the many other classifications which rely only on a molecular or on a clinical basis. In order to reach such a result, however, the researchers used a collection of breast cancer studies on patients who received chemotherapy adjuvants and from whom data were available concerning the so-called *pathological Complete Response* (pCR).[26] The iCluster approach also provides a grouping of biomarkers that can be used to test new treatments, with the underlying idea that by means of such testing of treatments we do not just perform trials to establish drug efficacy, but also, in a sense, test the adequacy of disease classification for clinical purposes. If we take this seriously, the very idea of how a complex disease is identified undergoes a significant change, with the responses to treatments impacting back on the definition of the pathology for which they were prescribed and which they were meant to cure. To some extent, classification of the disease is driven not only by the search for its underlying conditions or – more or less remote – aetiological causes, but a look backwards is accompanied also by a look forward in time towards clinical outcomes and the efficacy of drugs and prognoses.

[24] PAM50 (Prosigna®) is a tumour profiling test that helps determine the benefit of using chemotherapy in addition to hormone therapy for some oestrogen receptor-positive (ER-positive) and HER2-negative breast cancers.
[25] An analogous figure, but contemplating also a column explicitly dedicated to prognosis, is Table II in Dawson et al. (2013).
[26] A tumour is said to have had a *pathological Complete Response* (i.e. a pCR) if, after surgery, no residual cancer cells remain.

Table 1 Overview of the Integrative Cluster Subtypes and the Dominating Properties with Regard to Copy Number Driving Events, Biomarkers, Type of DNA Architecture,[46] Dominant PAM50 Subtype, and Clinical Outcome

Integrative cluster group	Copy number driver	Pathology biomarker class	DNA architecture	Dominant PAM50	Clinical characteristics (survival)
1	Chromosome 17/ chromosome 20	ER^+ ($HER2^+$)	Simplex/firestorm (chromosome 17q)	Luminal B	Intermediate
2	Chromosome 11	ER^+	Firestorm (chromosome 11q)	Luminal A and B	Poor
3	Very few	ER^+	Simplex/flat	Luminal A	Good
4	Very few	ER^+/ER^-	Sawtooth/flat	Luminal A (mixed)	Good (immune cells)
5	Chromosome 17 (*HER2* gene)	$ER^-(ER^+)/HER2^+$	Firestorm (chromosome 17q)	Luminal B and HER2	Extremely poor (in pre-Herceptin cohorts)
6	8p deletion	ER^+	Simplex/firestorm (chromosome 8p/ chromosome 11q)	Luminal B	Intermediate
7	Chromosome 16	ER^+	Simplex (chromosome 8q/chromosome 16q)	Luminal A	Good
8	Chromosome 1, Chromosome 16	ER^+	Simplex (chromosome 1q/chromosome 16q)	Luminal A	Good
9	Chromosome 8/ Chromosome 20	ER^+ (ER^-)	Simplex/firestorm (chromosome 8q/ chromosome 20q)	Luminal B (mixed)	Intermediate
10	Chromosome 5, Chromosome 8, Chromosome 10, Chromosome 12	TNBC	Complex/sawtooth	Basal-like	Poor 5-year, good long-term if survival

ER, estrogen receptor; TNBC, triple-negative breast carcinoma.

Fig. 3 (From Russnes et al. (2017).

This approach is very promising, as evidenced by the works adopting it (*mutatis mutandis*) to classify other types of cancer, in particular prostate cancer (five clusters), pancreatic cancer (four clusters), colorectal cancer (four clusters), bladder cancer (five clusters) and melanoma (four clusters).[27]

And what can be done for cancer heterogeneity? Nik-Zainal and colleagues faced this challenge by adopting a similar approach trying to find a unitary framework of such complexity exactly through computational integration[28]. They started from the idea, borrowed from population genetics (see Boniolo 2017), that there is a sort of "most common ancestor", that is, they suggested dividing the somatic mutations occurring over cancer's lifetime into those acquired before the last selective selection, and therefore shared by all cancer cells, and those acquired afterwards. They analysed the genome of 21 breast cancer samples to reconstruct their genomic history, clustering the classes of mutations via computation algorithms. To support their conception of what cancer heterogeneity means in terms of complexity, they built a clustered catalogue of more than 200,000 different mutations occurring over the course of patients' lives.

Thus, at least for breast cancer, we have the integration proposed by Caldas and colleagues, focused on taming tumour complexity at genomic and transcriptomic level but also considering

[27] See, respectively, Ross-Adams (2015); Weddell et al. (2015); Guinney et al. (2015); Robertson et al. (2017); Cancer Genome Atlas Network (2015).

[28] Mutational processes molding the genomes of 21 breast cancers. See Nik-Zainal et al. (2012); Nik-Zainal et al. (2016); Morganella et al. (2016).

the clinical level, and the integration proposed by Nik-Zainal and colleagues, focused on taming tame cancer heterogeneity. Although the two approaches could have a sort of meta-integration[29], let us stay on Caldas' project and elaborate further on its epistemological significance and prospects.

5. Taming complexity through integration: philosophical explorations on cancer research

The sections above have presented a few aspects related to current cancer research having to do with complexity and integration, understood in a manifold way. Issues arise due to new sequencing and imaging technologies, the awareness of cancer heterogeneity, different kinds of methodologies employed and evidence collected, and different possible classifications. Considering cancer heterogeneity, we have seen that each tumour is different from all others and is actually composed of different sub-tumours. Complexity in biomedicine strongly and directly impacts on clinical matters. It is totally unrealistic to start from complex features of cancers to try to find a therapy or propose a prognosis to a specific patient. In order to try to restore the previous, non-pathological course of events in each single patient, we must intervene in such variety, and to do so we need to start from some grouping of cancers. But on which bases? Some form of grouping must be devised to allow a partition in reference classes where any kind of tumour (and thus any patient) can be located, even if not in an absolutely precise way, in the best possible way given current knowledge. In other terms, to address both research and, even more so, clinical matters, "similarities must be found out of dissimilarities" among patients due to the uniqueness of their disease, their clinical story and their "omics", and these similarities should allow classificatory practices to impact effectively on explanatory and predictive issues relevant for each single patient.

In a wider theoretical perspective, everything we have sketchily addressed in the sections above hints at the fact that tumours and the related available data challenge our conception of diseases, and in many senses force us to reshape our epistemic practice when looking for explanatory, predictive and prognostic accounts. To recall the examples analysed above, we can no longer speak in terms of, for instance, breast cancer, but, properly speaking, we should refer to *one of the many possible cancers* affecting the breast. If so, how do we cope with several different diseases of the same kind, both in theoretical and clinical terms? Should we change nosology, and, if so, on which grounds should we do so, given that we can have several different classifications of possible cancers affecting the breast? How do current classificatory practices impact on our understanding of cancer as a complex disease and of its natural history, and how

[29] This is what is happening inside the *Personalised Breast Cancer Project*! See, https://crukcambridgecentre.org.uk/news/personalised-breast-cancer-program-launches-cambridge.

do our classifications, given the deluge of available data, allow for the attribution of each single case to the proper reference class?

We have seen that in the case of integrative clusters, a combination of "information on the genomic and transcriptomic landscapes of […] cancer to refine the molecular classification of the disease" (Dawson et al. 2013, p. 617) is pursued through statistical and computational methods. A *revision* of the disease classification is suggested, *rather than* its *reduction* to some allegedly more fundamental level. Approaches to complexity and, especially, to integration must be rethought and reshaped. iClusters do not address complexity by epistemic decomposition of the system into subsystems and then re-assembly, or by the integration of compatible and complementary explanatory models. What is integrated is not different accounts of the same behaviour, but rather heterogeneous data through clustering procedures based on similarities. What tables like the one in Fig.3 above represent is a form of integration, which is aimed to supply information on the inception of each sub-type of the disease (and, in this sense, to supply causal explanatory information at the level of chromosome mutations), a predictive account of the disease (with respect to biomarkers), and a prognostic picture (see the last column on the right). This yields a unitary framework allowing a new classification that takes into account both the bench level (i.e., the research one based on "omics") and the bedside level (i.e., the clinical one based on patients' situation and follow up), and also serves as a predictive tool. With data integration, what in the end is being suggested is also integration of different epistemic procedures.

Discussing complex systems, Wimsatt has stressed that we need to decide which are the *relevant* components and levels with respect to the *epistemic aim* at stake, plus "we need to know more generally how we order and relate different descriptions of the behaviour of a system, particularly partial descriptions, to construct explanatory accounts of its behaviour" (2007, p. 161). Wimsatt's view both does and does not fit the case at stake here. On the one hand, decisions on what counts as relevant variables and on the epistemic aims to be pursued play a crucial role in clustering: clustering per se is not a domain-independent method, it is used in a variety of contexts and with different goals, and it is then with respect to specific goals and contexts that merits of clusters will be evaluated[30]. On the other hand, not much seems to be left in integrative clustering of the direct construction of different partial, explicit descriptions, to be then combined to form explanatory accounts of the behaviour of the system itself. The focus is not on discourses on the integration of different explanatory levels in the medical context, or on epistemological concerns regarding interactions among levels at various spatio-temporal scales,

[30] On measures and evaluation of the usefulness of clusters for particular tasks, and for a catalogue of clustering problems, see e.g. von Luxburg, Williamson and Guyon (2012).

or the disclosure of different mechanistic sub-systems. Hierarchies and levels are set aside in favour of a range of integrative, therapeutically relevant and predictive clusters, each being "associated with distinct clinical outcomes and providing new insights into the underlying biology and potential molecular drivers" (Dawson et al. 2013, p. 617). Clues on the biological underpinnings are not achieved through the interaction of multiple alternative accounts. Rather, the identification of diseases, and some fixing of cancer type and subtypes, stem from the clusters themselves on the basis of the algorithms chosen.

What is worth stressing is that a totally different, novel idea of "integration" is at stake here, an idea that is not taken as part and parcel of forms of pluralism. Different kinds of data are collected from different, heterogeneous sources and include genomic, epigenomic, transcriptomic, clinical, epidemiological, etc., information. No direct cooperation between different research groups and laboratories is advocated, and sharing of data and collaboration across fields takes a very specific form: integration is achieved by clustering thanks to statistics and computer algorithms. In this perspective, the pathology is not identified on the basis of, for instance, symptoms and signs alone, nor on the basis of a given set of biomarkers fixed *ex ante*, to then progressively add further variables and complementary perspectives. Cancer classification is presented as the outcome of the computational clustering process, which serves as a bridge-tool to navigate our way through the deluge of biological data and follow-up clinical data, and, at the same time, to pursue both explanatory and predictive targets – through focus on etiological and prognostic factors respectively. Cancer types and subtypes are thus carved out of complexity by computational tools; integration is achieved by establishing patterns, which will then be taken as reference points to overcome problems and secure better diagnosis and treatment. The classifications mentioned above based on clustering take all or most of the course of patients' lives into account, considering both mutations and treatment outcomes. Classification itself is grounded on data collected at different spatial and temporal scales, which are reassembled through clustering. Temporal scales, in particular, span long intervals: not only is it assumed that, given the complex features of cancer, the development of the disorder and its dynamics must be followed from the predisposing factors up to the symptoms, but that prognostic elements will also influence classification. Integration here is then not seen as a solution aimed at making different stances converge, in the end, or different models interact fruitfully, possibly as complementary pictures. Rather, integration of data and interpretation of their epistemological significance in outlining aspects of the natural history of the disease go hand in hand and proceed simultaneously.

"Integration" has been widely discussed as providing some sort of forward thinking, which,

as we stressed in § 2, has been presented as highly beneficial with respect to various epistemic aims. What are the steps forward warranted by iClusters? As appears from the above reflections, the forward-looking aspect here is not dictated by some puzzle-like arrangement of partial complementary accounts, but by devising computational tools that address not only, e.g., genetic diversity due to inherited genetic variation or acquired genomic aberrations, but also variances in incidence, in treatment efficacy, and in intermediate prognoses and survival rates, that is, clinical aspects. The importance of proper sub-stratification is highlighted as the road to novel potential therapeutic targets[31]. What is ultimately taken as relevant for the identification of what cancer is are both occurrences back in the patient's life and elements which are significant in terms of treatments and their outcomes. Complexity is addressed by isolating classificatory sets, and concerns shall then regard aspects of cluster validity and their measurements – considering, e.g., such issues as small within-cluster dissimilarities or between-cluster separation[32]. At the same time, the distance between prognosis and classification grows shorter, with the classification being quite far away from attempts at some definitive nosography.
What these new methodologies suggest are shifts in epistemic attitudes as well: instead of digging deeper and deeper, discussing aggregative/non-aggregative, emergent/reducible, decomposable, non-decomposable or nearly-decomposable features, to unravel underlying mechanisms and the like, the relevance of a look at the life-long development of conditions and at the outcomes of treatment strategies is stressed. Also symptoms at later disease stages are called to play a larger role in our epistemic practices, while aetiological factors are called to initiate an explanatory discourse.

Summing up, even if it cannot be considered *the solution* and even less *the definitive solution*, iCluster integration can be taken as a relevant, original and up-to-date standpoint to question more traditional – medical and philosophical – ways to address issues in biomedical research, from problems concerning interactions between different disciplinary fields to the philosophical implications of different ways of grouping entities. Integrative clustering puts forward a different way of coping with integrative needs, and encourages us to reconsider which relations can be established between describing, classifying and explaining a complex disorder once we opt for the identification of patterns through algorithms. Philosophical reflections will then have to reconsider theoretical approaches to integration, and address the genuine epistemological

[31] For instance, in discussing IntCluster3, Dawson et al (2013) state: "The excellent prognosis of this subtype emphasizes the importance of identifying this cluster within the previously defined luminal A intrinsic subtype, as these individuals represent a distinct group that could potentially be spared treatment with systemic chemotherapy" (p. 622), while InCluster4 provides hints as to specific immunological responses, to be exploited for future therapies.
[32] See e.g. Hennig (2017).

meaning of classificatory practices in taming complexity[33]. In turn, iClusters opens up a range of further questions: How stable is the taxonomy we are provided with, and under which circumstances will it be revised? Which definitions and descriptions of cancer could we build up relying on the integrative clusters elaborated, and how fine-grained shall they be? How much does integrative clustering, as an antidote to excessive disciplinary specialization and fragmentation of data patterns, illuminate biological complexity as such? How does this all impact the delivery of healthcare to individuals? And to what extent do cluster-based approaches to identifying dependencies actually draw on existing theories?

In order to answer to these epistemological question, one should carefully consider two aspects: i) every clustering method depends on the quality and the extent of the available data and that since the interactome (i.e., the complete available set of interacting molecules in a human organism) has yet to be completed, the clusters identified by the same iCluster method could evolve in time and with the ever increasing amount of molecular data available; ii) different types of clustering methods would have led to potentially different results, different subgroups and to reclassifying some patients.

All these issues are affected by the fact that, in the cases we are considering, we do not move from the identification of some given entity to the integration of kinds of evidence and/or models over it, we are not clustering around a mechanism or the like. Instead, the entity itself is identified through a computational integration of heterogeneous data. Whereas the traditional cancer nomenclature has been based on organ location, thereby directly designating the affected structure, the unprecedented amount of data collected from increasingly large tumour cohorts has challenged such grounding. What are integrated now are huge sets of data into robust classifiers, to be clinically implemented in patient management. Amongst critical issues we find the need to avoid excessive fracturing of cancer subtypes, and the identification of their precise biological meaning[34]. Clusters should be used for prediction and risk stratification as moves towards precision medicine. One of the next steps to be taken will be to evaluate "within-cluster patient heterogeneity, which would be studied by relating patients' distance to the cluster centre or posterior probabilities of cluster membership to outcomes" (van Smeden, Harrell and Dahly 2018, p. 440), and to discuss criteria upon which some variables are relevant for clustering and some are not. "In real applications, in which the variables have a meaning that is of substantial importance for the clustering task, choosing different variables changes the meaning of the resulting clustering, [...] and whether certain variables 'do not cluster' and whether they then

[33] For some theoretical reflections on the relations between clusters and ways of conceiving natural kinds, reality and truth, see Hennig (2015). Hennig discusses context- and aims-dependence of clustering methods, comparisons and choices among them, and related impacts in practice.
[34] On problems for cancer classification see also Song at al. (2015).

should not be involved in the computation of the clustering of interest depends on the context and the clustering aims" (Hennig 2015, p. 61).

In sum, iCluster is not *the* solution to tackle cancer complexity, but it seems to be a fruitful approach, which, by bringing together molecular and clinical information, effectively provides etiological and prognostic classes of relevance, where any new patient could be inserted and hence have a robust idea of what his/her clinical trajectory is likely to be. Moreover, not only does iCluster play a promising role in classificatory practice, but it is also a reliable starting point, given currently available knowledge, for individualised treatments.

6. Acknowledgments

We would like to thanks the four anonymous referees, whose comments have been very useful to improve early versions of the manuscript.

References

Ali, Raza H..et al. 2014. Genome-driven integrated classification of breast cancer validated in over 7,500 samples. Genome Biology 15: 431.

Ashley, Euan A. 2015. The Precision Medicine Initiative: A new national effort. JAMA 313: 2119–2120.

Bechtel, William, and Robert Richardson. 2010. Discovering complexity. Decomposition and localization as strategies in scientific research. Cambridge: MIT Press.

Bertolaso, Marta. 2016. Philosophy of cancer. A dynamic and relational view. Dordrecht: Springer.

Boniolo, Giovanni. 2007. On scientific representation. From Kant to a new philosophy of science. Houndmills: Palgrave Macmillan, chapter 1.

Boniolo, Giovanni. 2017. Patchwork narratives for tumour heterogeneity. In Logic, methodology and philosophy of science – Proceedings of the 15th International Congress, eds Hannes Leitgeb, Ilkka Niiniluoto, Elliott Sober, Päivi Seppälä, 311-324. London: College Publications.

Boniolo, Giovanni, and Raffaella Campaner. 2018. Molecular pathways and the contextual explanation of molecular functions, Biology & Philosophy 33: 24. doi: 10.1007/s10539-018-9634-2.

Boniolo, Giovanni, and Marco J. Nathan (eds) 2017. Philosophy of molecular medicine. London: Routledge.

Boyd, Danah, and Kate Crawford. 2012. Provocations for a cultural, technological, and

scholarly phenomenon. Information, Communication & Society 15(5): 662-679.
Bracht, Karin. 2009. Biomarker: Indikatoren für Diagnose und Therapie. Pharmazeutische Zeitung, http://www.pharmazeutische-zeitung.de/index.php?id=29346 (Accessed 30 April 2017).
Brigandt, Ingo. 2010. Beyond reduction and pluralism: Toward an epistemology of explanatory integration in biology. Erkenntnis 73: 295-311.
Brigandt, Ingo. 2013. Integration in biology: Philosophical perspectives on the dynamics of interdisciplinarity. Studies in History and Philosophy of Biological and Biomedical Sciences 44: 461-465.
Bruna, Alejandra et al. 2016. A biobank of breast cancer explants with preserved intra-tumor heterogeneity to screen anticancer compounds. Cell 167: 260–274.
Burrell, Rebecca A. et al. 2013. The causes and consequences of genetic heterogeneity in cancer evolution. Nature 501: 338–45.
Cancer Genome Atlas Network. 2015. Genomic Classification of Cutaneous Melanoma, Cell 161:1681-96. doi: 10.1016/j.cell.2015.05.044.
Chang, Hasok. 2012. Is Water H_2O? Evidence, realism and pluralism, Dordrecht: Springer.
Collins, Francis S., Varmus, Harold. 2015. A new initiative on precision medicine. New England Journal of Medicine 372: 793–795.
Conveney, Peter V., Dougherty, Edward, and Roger Highfield. 2016. Big data need big theory too. Phil. Trans. R. Soc. A 374: 20160153. http://dx.doi.org/10.1098/rsta.2016.0153
Craver, Carl. 2007. Explaining the brain. Oxford: OUP.
Curtis, Christina et al. 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 18; 486: 346-52.
Dawson, Sarah-Jane et al. 2013. A new genome-driven integrated classification of breast cancer and its implications. The EMBO Journal 32: 617-628.
Fagan, Melinda B. (2017), Pathways to the clinic: cancer stem cells and challenges for translational research. In Boniolo and Nathan (eds), 165-191.
Geyer, Felipe C. et al. 2010. Molecular analysis reveals a genetic basis for the phenotypic diversity of metaplastic breast carcinomas. J Pathol 220: 562–573.
Ghasemi, Mojtaba, Iraj Nabipour, Abdolmajid Omrani, Zeinab Alipour and Majid Assadi. 2016. Precision medicine and molecular imaging: new targeted approaches toward cancer therapeutic and diagnosis. American Journal of Nuclear Medicine and Molecular Imaging, 6(6): 310-327.
Golubnitschaja, Olga et al. 2016. Medicine in the early twenty-first century: paradigm and anticipation - EPMA position paper 2016. EMPA Journal 7: 23. DOI 10.1186/s13167-016-0072-4

Goossens, Nicolas, Nakagawa, Shigeki, Sun, Xiaochen., and Yujin Hoshida 2015. Cancer biomarker discovery and validation. Transl Cancer Res 4: 256-269.
Guinney, Justin et al. 2015. The consensus molecular subtypes of colorectal cancer. Nat Med. 21:1350-6.
Hastie, Trevor, Robert Tibshirani and Jerome Friedman Jerome. 2008. The elements of statistical learning. data mining, inference, and prediction, 502-503. New York: Springer.
Hennig, Christian. 2015. What are true clusters? Pattern Recognition Letters 64: 53-62.
Hennig, Christian 2017. Cluster validation by measurement of clustering characteristics relevant to the user, arXiv:1703.09282v1[stat.ME]
Hooker, Cliff (ed.) 2011. Philosophy of complex systems, Elsevier: Amsterdam.
Kaiser, Jocelyn. 2015. Obama gives East room rollout to Precision Medicine Initiative, Science, doi: 10.1126/science.aaa6436 (Jan. 30 2015).
Kaiser, Marie. 2013. Complexity. In Encyclopedia of Systems Biology, eds Werner Dubitzky, Olaf Wolkenhauer, Kwang-Hyun Cho, Hiroki Yokota, 456-460. New York: Springer.
Kitchin, Rob. 2014. Big data: new epistemologies and paradigm shifts. Big Data & Society 1(1): 1-12.
Kohane, Isaac S. 2015. Ten things we have to do to achieve precision medicine. Science 349: 37-38.
Koychev, Ivan, Emma Barkus, Ulrich Ettinger, Simon Killcross, Jonathan P. Roiser, Lawrence Wilkinson, and Bill Deakin. 2011. Evaluation of state and trait biomarkers in healthy volunteers for the development of novel drug treatments in schizophrenia. J Psychopharmacol. 25: 1207-1225.
Ladyman, James, James Lambert, and Karoline Wiesner 2013. What is a complex system? European Journal for the Philosophy of Science 3: 33-67.
Ladyman, James, and Karoline Wiesner. Forthcoming, 2018. What is a complex system? Princeton: Princeton University Press.
Lee, J. Jack 2003. Statistical methods for biomarker analysis for head and neck carcinogenesis and prevention. In Head and neck cancer, eds John F. Ensley, J. Silvio Gutkind, John R. Jacobs, and Scott M. Lippman, 287-304. San Diego: Academic Press.
Lemoine, Maël. 2017. Molecular complexity: Why has psychiatry not been revolutionised by genomics (yet)? In Philosophy of molecular medicine, eds Giovanni Boniolo and Marco J. Nathan, ch. 4, London: Taylor and Francis.
Leonelli, Sabina. 2008. Bio-ontologies as tools for integration in biology. Biological Theory 3: 8-11.

Leonelli S (2014) What difference does quantity make? On the epistemology of Big Data in biology. Big Data Soc 1: 1–11
Leonelli, Sabina. 2016. Data-centric biology: A philosophical study. Chicago: University of Chicago Press.
Martelotto, Luciano G. et al. 2014. Breast cancer intra-tumour heterogeneity. Breast Cancer Research 16: 210, doi:10.1186/bcr3658.
Mitchell, Sandra. 2003. Biological complexity and integrative pluralism. Cambridge: Cambridge University Press.
Mitchell, Sandra. 2009. Unsimple truths. Science, complexity and policy. Chicago and London: The University of Chicago Press.
Mordente, Alvano, Elisabetta Meucci, Giuseppe Ettore Martorana, and Andrea Silvestrini. 2015. Cancer biomarkers discovery and validation: State of the art, problems and future perspectives. Adv Exp Med Biol 867: 9-26.
Morganella, Sandro et al. 2016. The topography of mutational processes in breast cancer genomes. Nature Communications 7: 11383.
Mehmood, Muhammad Aamer, Sehar, Ujala, and Niaz Ahmad. 2014. Use of bioinformatics in different spheres of life sciences. Data Mining in Genomics and Proteomics 5(2) 1000158.
Nabipour, Iraj, and Majid Assadi. 2016. Precision medicine, an approach for development of the future medicine technologies. ISMJ 19: 167-184.
Negm, Robert S., Mukesh Verma, and Sudhir Srivastava. 2002. The promise of biomarkers in cancer screening and detection. Trends Mol Med 8: 288-293.
Nik-Zainal, Serena et al. 2012. The life history of 21 breast cancers. Cell 149: 994-1007.
Nik-Zainal, Serena et al. 2016. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 534: 47-54.
O'Malley, Maureen A., and Okrun Soyer. 2012. The roles of integration in molecular systems biology. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 43: 58–68.
Ouzounis, Christos. 2012. Rise and Demise of Bioinformatics? Promise and Progress. PLOS Computational Biology 8(4) e1002487.
Pereira, Bernard et al. 2016. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. Nature Communications 7: 11479, doi:10.1038/ncomms11479.
Perez-Iratxeta, Carolina, Andrade-Navarro, Miguel, and Wren, Jonathan. 2007. Evolving research trends in bioinformatics. Briefings in Bioinformatics 8(2): 88-95.
Perou, Charles M, Sørlie, Therese et al. 2000. Molecular portraits of human breast tumours.

Nature 406: 747-752.
Plutynski, Anya. 2013. Cancer and the goals of integration. Studies in History and Philosophy of Biological and Biomedical Sciences 44: 266-276.
Plutynski, Anya. Forthcoming, 2018. Explaining cancer: Finding order in disorder. Oxford: Oxford University Press.
Pu, Fan, Xue, Jingjuan Qiao, Anvi Patel, and Jenny J. 2016a. Towards the molecular imaging of prostate cancer biomarkers using protein-based MRI contrast agents. Curr Protein Pept Sci. 17(6): 519-33.
Pu Fan, Shenghui Xue, and Jenny J. Yang 2016b. ProCA1.GRPR: a new imaging agent in cancer detection. Biomark Med. 10(5): 449-52.
Ratti, Emanuele 2016. The end of 'small biology'? Some thoughts about biomedicine and big science. Big Data & Society. https://doi.org/10.1177%2F2053951716678430.
Robertson, A. Gordon et al. 2017. Comprehensive molecular characterization of muscle-invasive bladder cancer. Cell 171: 540-556.e25.
Ross-Adams, Helen. 2015. Integration of copy number and transcriptomics provides risk stratification in prostate cancer: A discovery and validation cohort study. EBioMedicine 2: 1133–1144.
Russnes, Hege G et al. 2017. Breast cancer molecular stratification: From intrinsic subtypes to integrative clusters. Am J. Pathol. 187: 2152-2162.
Sabatello, Maya and Paul S. Appelbaum. 2017. The precision medicine nation. Hastings Center Report 47, 4: 19-29.
Scatena, Roberto (ed.) 2015. Advances in cancer biomarkers. From biochemistry to clinic for a critical revision. Heidelberg: Springer.
Shyr, Derek, and Qi Liu. 2013. Next generation sequencing in cancer research and clinical application. Biol Procedures Online 15:4, doi: 10.1186/1480-9222-15-4
Song, Qingxuan., Sofia D. Merajver, and Jun Z. Li 2015. Cancer classification in the genomic era: five contemporary problems. Human Genomics 9: 27, doi: 10.1186/s40246-015-0049-8
Strasser, Bruno J. 2017. The "Data-Deluge": Turning private data into public archives. In Science in the archives. Pasts, presents, futures, ed. Lorraine Daston, 185-202. Chicago: Chicago University Press.
Tonelli Mark R, and Brian H Shirts. 2017. Knowledge for precision medicine. Mechanistic reasoning and methodological pluralism. Jama 18: 1649-1650.
Torres, Lurdes et al. 2006. Intratumour genomic heterogeneity in breast cancer with clonal divergence between primary carcinomas and lymph node metastases. Breast Cancer Res Treat

102: 143–155.
Van Smeden, Maarten, Frank Harrell, and Darren Dahly. 2018. Novel diabetes subtypes. The Lancet 6: 439-440.
Vasan, Ramachandran S. 2006. Biomarkers of cardiovascular disease molecular basis and practical considerations. Circulation 113: 2335–2362.
VV. AA. 2013. Integration in biology: Philosophical perspectives on the dynamics of interdisciplinarity. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences 44: 461-562.
Von Luxburg, Ulrike, Robert C. Williamson and Isabelle Guyon. 2012. Clustering: science or art?, JMLR, Workshop and Conference Proceedings 27: 65-79.
Weddell Nicola et al. 2015. Whole genomes redefine the mutational landscape of pancreatic cancer. Nature 518, 26: 495–501.
Weinberg Robert A. 2014. Coming full circle—From endless complexity to simplicity and back again. Cell 57: 267 – 271.
Wimsatt, William. 2007. Re-engineering philosophy for limited beings. Piecewise approximation to reality. Cambridge: Harvard University Press.
Xue Shenghui, Jingjuan Qiao, Fan Pu, Matthew Cameron, and Jenny Yang. 2013. Design of a novel class of protein-based magnetic resonance imaging contrast agents for the molecular imaging of cancer biomarkers. Wiley interdisciplinary reviews Nanomedicine and nanobiotechnology 5(2): 163-179. doi:10.1002/wnan.1205.
Zhang, Xiaohua Douglas D. 2015. Precision medicine, personalized medicine, omics and big data: concepts and relationships. Journal of Pharmacogenomics and Pharmacoproteomics 6:2.