ABSTRACT

Both clinical research and basic science rely on the epistemic practice of extrapolation from surrogate models, to the point that explanatory accounts presented in review papers and biology textbooks are in fact composite pictures reconstituted from data gathered in a variety of distinct experimental setups. This raises two new challenges to previously proposed mechanistic-similarity solutions to the problem of extrapolation: one pertaining to the absence of mechanistic knowledge in the early stages of research and the second to the large number of extrapolations underpinning explanatory accounts. An analysis of the strategies deployed in experimental research supports the conclusion that while results from validated surrogate models are treated as a legitimate line of evidence supporting claims about target systems, the overall structure of research projects also demonstrates that extrapolative inferences are not considered definitive or sufficient evidence, but only partially justified hypotheses subjected to further testing.

1 Introduction
2 Surrogate Models
   2.1 What exactly is a surrogate model?
   2.2 Why use surrogate models?
3 Prior Validation of Surrogate Models
   3.1 The validation and ranking of surrogate models in the early stages of basic research
   3.2 The validation of surrogate models in later stages of basic research and in clinical research
4 ‘Big Picture’ Accounts and the Extrapolations Underpinning Them
   4.1 The mosaic nature of mechanistic descriptions in basic science
   4.2 Challenges for mechanistic-similarity-based validation protocols
5 Retrospective Testing of Extrapolated Knowledge
   5.1 Holistic confirmation
It is generally acknowledged that extrapolation plays an important role in applying the results of basic science for medical purposes. Basic science can demonstrate that an antiviral treatment reduces the viral load in simian immunodeficiency virus (SIV) infected rhesus monkeys, that an anti-cancer drug kills human hepatocytes \textit{in vitro}, and that some commonly used food additive is carcinogenic in mice. Given this information, medical practitioners, pharmaceutical companies, and healthcare policy makers must make a decision whether to proceed to human clinical trials for an AIDS treatment, issue a toxicity warning for the anti-cancer drug, and prohibit the use of the additive in the food industry. The most informative data for making these decisions would be to repeat the experiments of basic science in humans. However, for ethical reasons, these experiments should not be performed if there is no sufficient evidence to demonstrate that the benefits they may bestow the tested subjects outweigh possible harm. The alternative is to extrapolate findings from basic science. In other words, use SIV-infected rhesus monkeys as a surrogate model for studying HIV infection in humans, \textit{in vitro} grown hepatocytes as a surrogate model for testing the effect of anti-cancer drugs on human livers, and mice as a surrogate model for testing the effects of food additives on humans.

The use of surrogate models raises an obvious difficulty, known as the problem of extrapolation: given the differences between the surrogate systems in which the findings are actually documented and the target to which the findings are extrapolated, there is no guarantee that what is true of the surrogate must also be true of the target within a tolerable degree of approximation (LaFollette and Shanks [1993], [1996]; Howick \textit{et al.} [2013]). The fact that many signalling, developmental, and metabolic mechanisms and pathways are conserved provides general insights about what may hold true, with some variation, of a large number of phylogenetically related organisms (Schaffner [2001]; Weber [2005]; Ankeny and Leonelli [2011]). Nevertheless, while such insights play an important role in guiding the discovery process, clinical researchers seek more precise answers. The goal is to figure out whether a particular result documented in the surrogate also holds true, ideally down to minute qualitative and quantitative detail, of the target system. When aiming at this degree of precision, the suitability of a surrogate model needs to be evaluated on a case by case basis (Burian [1993]; Ankeny [2001];
Leonelli [2007]), where the evaluation procedure relies on an assessment of the relevant similarities between surrogate and target (Bolker [2009]). It has been often argued that the most relevant similarities are those concerning mechanisms causally productive of the phenomenon of interest (Wimsatt [1976]; Cartwright [1989]; Schaffner [2001]; Weber [2005]; Craver [2007]; Steel [2007]). Surrogate models that share mechanistic features with their targets are more likely to generate the phenomena of interest via the same causal pathways and respond in similar ways when these pathways are disturbed. By contrast, surrogates that do not share causal features might generate similar phenomena by means of different mechanisms, and thus behave differently when subjected to similar experimental interventions.

In this article, I want to draw attention to the largely overlooked fact that extrapolation from surrogate models is by no means limited to clinical research. Basic science is equally indebted to the epistemic practice of extrapolation from surrogate models, with all the benefits and risks this practice entails. The discovery process in basic science faces a multitude of experimental hurdles, and it is seldom the case that a single experimental setup succeeds in addressing all difficulties. The usual way around this is to systematically trade one experimental difficulty for another by changing experimental setups, in the hope that a more complete knowledge of the phenomenon and its underlying mechanisms can ultimately be attained by conducting studies in a multitude of setups. One of the most striking consequences of this practice is that descriptions of phenomena and explanatory accounts—such as the diagrams of mechanisms in cell and molecular biology, including their extensions in other fields of investigation, basic or applied—are in fact mosaic jigsaws reconstituted from bits and pieces of data gathered in distinct experimental setups. How adequately this knowledge reflects reality depends in no negligible part on the extent to which results can be safely extrapolated from one experimental setup to another.

The widespread use of extrapolations in basic science raises new challenges to anyone interested in tackling the problem of extrapolation. One immediate difficulty stems from the fact that current solutions work on the premise that we know a sufficient amount about mechanistic similarities and differences between surrogate and target. While such knowledge is available in later stages of research, it cannot be assumed in the initial stages, when nothing is known about the mechanistic basis of a phenomenon. Thus, if the standard solution to the problem of extrapolation relies on prior mechanistic knowledge, then the new challenge is how to address the problem in the absence of mechanistic knowledge. A second challenge stems from the fact that solutions to the problem of extrapolation are framed in relation to the justification of isolated extrapolations given an already available background knowledge, the origin of which is not put into question. This is the typical scenario in clinical
research, where the task at hand is figuring out how to justify a particular extrapolation, usually one of immediate practical import, given substantial mechanistic knowledge already provided by basic science. The situation changes radically when it comes to justifying the very background knowledge from basic science, which, as it turns out, combines data from tens or even hundreds of distinct experimental setups. If there is a probability of error associated with any given extrapolation, then as knowledge increases, the number of extrapolations increases as well, and with it the probability of error. Thus, what may constitute a reasonable strategy for addressing the problem of extrapolation in the context of clinical research is unlikely to provide an equally workable way around the problem in basic research.

An analysis of scientific practice reveals that scientists treat extrapolation as a matter of taking an epistemic risk. I argue thus that it is misleading to think that there is such a thing as a definitive, universally applicable solution to the problem of extrapolation. The relevant question is rather how to plan scientific research in such a way as to keep the possibility of error under control. As I will illustrate in what follows, researchers in the life sciences deploy a surprisingly varied array of tactics, from model validation protocols aimed at minimizing the possibility of error for prospective extrapolations, to holistic confirmation strategies aimed at the retrospective testing of previously made extrapolations and fallback positions aimed at providing a stable epistemic basis for troubleshooting anomalies, all of which are carefully orchestrated in such a way as to ensure the overall viability of a research project. This vindicates a more nuanced view about the epistemic status of extrapolations. Extrapolations are used in the context of an overall research strategy that combines both a bottom-up process of inferring mechanistic accounts based on experimental data and a subsequent top-down testing of predictions made by these accounts. On the one hand, defenders of extrapolative practices are right in claiming that the use of surrogate models is a reliable experimental practice, and the tens of thousands of monthly articles publishing results gathered in surrogate models certainly demonstrate that extrapolations from validated surrogate models are treated as legitimate evidence supporting the bottom-up construction of explanatory mechanistic accounts. On the other hand, however, the equally well-documented efforts deployed to retrospectively test extrapolative inferences also demonstrate that findings from any given surrogate model are not considered definitive or sufficient evidence, but simply one line of legitimate evidence,1 subsequently strengthened and corroborated by means of top-down testing of predictions from mechanistic models.

1 Steel ([2013]) reaches a similar conclusion using a case study involving socio-economical mechanism.
The article is organized as follows. Section 2 provides a general discussion of surrogate models and their use in science. In Section 3, I discuss the model validation strategies in virtue of which surrogate models are evaluated and ranked relative to their suitability for supporting extrapolations to their intended targets. In Section 4, I argue for the mosaic nature of the descriptions of mechanisms available in the scientific literature and show how the systematic use of extrapolations in basic science raises new challenges for current attempts to provide a solution to the problem of extrapolation. In Section 5, I discuss retrospective confirmation strategies and fallback positions used to test and troubleshoot mechanistic accounts. Finally, in Section 6, I summarize my findings and arguments.

2 Surrogate Models

2.1 What exactly is a surrogate model?

An experimental surrogate model can be defined as a more manageable experimental setup for studying a phenomenon, where this experimental setup serves as a substitute for another, experimentally less manageable, but physiologically more relevant setup. It is hoped that the findings generated by the investigation of the surrogate model can be safely extrapolated to the target experimental setup despite the differences between the two setups (Steel [2007]; Bolker [2009]; Baetu [2014]).

According to the above definition, a surrogate model is an experimental setup in which the phenomenon of interest can be consistently documented or replicated. An experimental setup consists of an organism, tissue, cell, molecular components, or other structure derived from living organisms on which researchers conduct experimental interventions, usually in order to gain knowledge of the causes and mechanisms of biological phenomena. An important feature of an experimental setup is that it is not just a physical object, but a labelled object. An experimental setup also includes an operationalized protocol detailing how and from which source the system in question has been obtained, how it has been grown, maintained, and prepared prior to experimentation, and what criteria and experimental controls have been used to ensure and demonstrate that the system has been indeed obtained and prepared as described in the protocol (Rheinberger [1997]; Weber [2005, 2008]). Labelling and standardization play a crucial role in ensuring that the

2 In addition to experimental surrogate models, there are also mathematical models of physical systems explicitly aimed at providing in silico surrogates for investigating quantitative-dynamic aspects of the systems they model (Bechtel and Abrahamsen [2010]; Winsberg [2010]; Brigandt [2013]; Baetu [2015], [forthcoming]; Gross [2014]). Theoretical surrogate models raise additional challenges related to abstraction, idealization, and causal interpretation, which will not be discussed here.
setup can be reproduced and identified by different research teams (Clarke and Fujimura [1992]; Bowker and Star [1999]; Ankeny [2001]; Müller-Wille [2007]), which in turn ensure that findings generated in the context of the same setup are more likely to be about ‘the same thing’, and thus can be directly compared and synthesized into more comprehensive bodies of knowledge.3

In as much as an experimental setup is used as a surrogate for extrapolating results to different experimental setups, a risk of error is expected due to the fact that similar phenomena may be produced by different mechanisms in different experimental setups. Unlike physics and chemistry, where one can often talk of natural kinds such as ‘electron’ or ‘benzene ring’, in the life sciences, there is no such thing as a ‘generic mammal’, ‘generic animal cell’, or ‘life in general’, only a multitude of individual biological systems. Incontestably, there exist mechanisms conserved among all living organisms, as well as mechanisms shared within more or less inclusive taxa. However, ‘conserved’ does not mean ‘identical’ or even ‘similar enough for all intents and purposes’. Subtle and not so subtle differences are omnipresent, from variations in the highly conserved mechanisms responsible for biological activities shared by all living things (for example, slight variations of the genetic code, or the more significant differences between eukaryotes and prokaryotes making possible the use of antibiotics), to differences in the mechanisms underpinning similar biological functions (for example, acquired immunity is a shared characteristic of all jawed vertebrates, yet the mechanisms underpinning it vary in different species), to differences between individuals of the same species (for example, AIDS pathogenesis in humans is variable due, in part, to resistance mediated by truncated receptors in human cells and the presence/absence of specific mechanisms of defence). This epistemic risk is anticipated by researchers and met with a variety of strategies aimed at justifying the prospective use of surrogate models for the purposes of making extrapolations (Section 3), as well as the retrospective testing of resulting explanatory accounts and the extrapolations on the basis of which they were developed (Section 5).

3 Note that any given experimental setup may or may not be used as a surrogate model. It is worth keeping in mind that many experimental setups are used to investigate phenomena directly in the biological systems of interest. Examples abound in the life sciences—there is a human model of protective immunity to HIV, a human blister model for studying early inflammatory responses in humans, a shark model for studying adaptive immunity in sharks, and so on. In such cases, an experimental setup is said to be a good model for studying a phenomenon because it allows researchers to reproduce the phenomenon of interest with a high rate of success, as well as to detect and gain experimental control over causal factors relevant to the production of the phenomenon (Baetu [2013], [2014]).
2.2 Why use surrogate models?

Given the expectation of an epistemic risk associated with extrapolation, one may wonder why researchers rely on surrogate models in the first place. The most common answer, in order to increase experimental control over causally relevant factors, simply means that surrogate models are used in order to conclusively disentangle causally relevant factors from a variety of confounding variables that turned out to be extremely hard or plain impossible to disentangle by conducting experimental interventions directly on the target system.

In order to better understand why surrogate models are used in basic science, it is worth considering an actual example. The early stages of an immune response are described at a cytological level as follows: following exposure to certain pathogens, T-cells in mice, humans, and other mammals are activated, and then multiply for a brief period of time, after which cells return to their usual number and inactive state. Since it is difficult to work with whole organisms, and even more difficult to figure out which among the many changes in whole organisms are causally linked to T-cell activation, immunologists quickly switched to simpler experimental setups consisting of primary T-cells (cells extracted from healthy organisms and grown in vitro), which to this day serve as a common surrogate model for studying T-cell activation in vivo (that is, in the living animal). Researchers were able to replicate the results in a primary T-cell model and figure out that not only do T-cells divide in response to stimulation, they also produce a variety of chemicals required for mounting inflammatory and other immune responses, after which they seem to die via apoptosis (or programmed cell death, to be contrasted with necrosis, or damage-induced cell death).

These results were, however, mitigated by the fact that primary T-cells do not survive for long in culture and, sooner or later, they die whether they are stimulated or not. This experimental hurdle prompted yet another model switch, namely, to an immortalized T-cell model, which is by far the most commonly used surrogate model for studying T-cell activation in vivo. Among the many T-cell lines used in immunology, the most widely used are the Jurkat T-cells, originally derived from a human leukemia patient. Since immortalized cell lines are cancerous derivatives of primary cells, the two kinds of cells

---

4 In order to maintain the focus on epistemic issues of interest to the philosopher of science, the technical details of the case are kept to a minimum. Further information can be found in any recent immunology textbook under the index entries ‘T-cell activation’ (for example, Rich et al. [2008], Chapter 13).

5 A subset of lymphocytes (white blood cells) responsible for handling intracellular parasites and orchestrating adaptive and innate immune responses in vertebrates.

6 As assessed by an increase in cytoplasmic mass, which is indicative of an increased metabolism.

7 For a general assessment of the role of the Jurkat cell line model in the elucidation of the mechanisms of T-cell receptor signalling, see (Abraham and Weiss [2004]).
are identical in many relevant respects: both express the same range of cell surface receptors, both respond to antigens by producing the same kind of chemicals (cytokines), both can be infected by the same viruses. The differences are thought to lie in one or more mutations that allow immortalized cells to replicate indefinitely; by contrast, primary cells undergo a limited number of divisions after which they die. There are many advantages of using cell lines, the immediately obvious one being that these cells can be cultured indefinitely in vitro, thus making it possible to conduct experiments and manipulations requiring longer periods of time and large quantities of cells, such as genetic manipulations. Cell lines are also highly standardized, since they constitute a population of stable clones with well-characterized morphological features that can be obtained from cell culture banks, such as the American Type Culture Collection. Genetic and phenotypic homogeneity reduces experimental variation, thus making it possible to amplify, identify, and purify molecular components, and to generate control markers for a variety of experimental designs.

Using immortalized T-cell lines, immunologists were able to demonstrate that, following stimulation, these cells do indeed undergo apoptosis and, as a consequence, the immune response shuts down. With this additional knowledge in hand, researchers turned back to animal models and were able to provide some evidence that what is true of primary and immortalized T-cells stimulated in vitro reflects what goes on in vivo. The combined results gathered in these systems made possible a detailed description of the phenomenon to be explained—in this case, a sequence of events beginning with a peak of activation of T-cells following stimulation with certain inducers, during which cells divide and secrete specific chemicals required for mounting an immune response, and ending with cell death and the shutdown of the immune response.

3 Prior Validation of Surrogate Models

3.1 The validation and ranking of surrogate models in the early stages of basic research

As illustrated in the above example, even descriptions of phenomena are usually gathered across diverse experimental setups. Special care is taken in order to ensure that the most important features of the phenomenon—in our example, cell division, synthesis of chemicals, and cell death—are consistently cross-referenced in all the experimental setups used. This cross-referencing, demonstrating that a qualitatively and quantitatively identical or highly similar phenomenon is documented and can be experimentally reproduced in multiple experimental setups, provides a crucial criterion for validating the use of these setups as models for studying that phenomenon. Conversely, the failure
to reproduce key aspects of the phenomenon of interest disqualifies the use of the setup as a surrogate.

If an experimental setup succeeds in overcoming a particular experimental hurdle, it seldom does so without creating a host of collateral problems, usually related to a loss in physiological relevance. As exemplified previously, the use of cell models is motivated to a large extent by the need to eliminate the background noise of biological activity not specifically linked to the phenomenon of interest. However, cell models may fail to adequately replicate histological and other systemic aspects of the physiological context of the whole organism, thus providing an incomplete or even distorted picture of the causal mechanisms underpinning the phenomenon. Likewise, genetic engineering has its own advantages and drawbacks. On the one hand, it allows researchers to intervene on specific molecular components in order to demonstrate their causal relevance to the phenomenon of interest. On the other hand, it also subjects cells and organisms to unusual stresses, hence there is always a risk that the genetically engineered cell or organism may be different from its unmodified counterpart in respects others than those intended.

Furthermore, testing the causal contribution of the many putative components of a molecular mechanism requires a multitude of slightly different experimental setups, each involving specific genetic changes required for the testing of a particular component; with this multiplication of setups comes an inevitable increase in the risk of error. Alternatively, the shortcomings associated with genetically engineered cell and animal models can be avoided by developing pharmaceutical compounds targeting molecular components in non-engineered cells or organisms. But this approach creates a different set of problems, since it involves a significant loss in the ability to specifically manipulate only those mechanistic components being tested, thus rendering the interpretation of experimental interventions notoriously difficult.

It is thus important to keep in mind that different surrogate setups rank differently in terms of physiological relevance. Since the phenomenon of interest is usually something that occurs naturally in whole organisms, data from animal models provide the most adequate information about the phenomenon and the mechanisms responsible for causing it. In our example, animal experimental setups are the most relevant systems to study the phenomenon of interest, and so the immediate target of investigation. Relative to this

---

8 For a more detailed philosophical analysis of the experimental interventions used to demonstrate causal and constitutive mechanistic relevance, see (Craver [2007]; Baetu [2012]).
9 For a critique of extrapolations of lab-based results to physiological conditions, see (Howick et al. [2013]).
10 The further question of which animal experimental setup provides the best basis for extrapolation to humans or mammals in general (clinical or evolutionary relevance) is a problem to worry about later, after some knowledge of the mechanisms responsible for the phenomenon is acquired. Then, an animal model may itself become a surrogate relative to other targets.
target, which occupies the highest rank of physiological relevance, other experimental setups are ranked based on the presence of relevant similarity with the target. Higher ranked experimental setups make better surrogate models of the target than lower ranked ones, where extrapolations from better models are expected to involve a lower risk of error.

As stated earlier, the most relevant similarities for the purpose of extrapolation are thought to be those concerning mechanisms known or likely to be relevant to the phenomenon of interest (Wimsatt [1976]; Cartwright [1989]; Schaffner [2001]; Weber [2005]; Craver [2007]; Steel [2007]). This approach is indeed documented in the biomedical practice of model validation, as well as later stages of research in basic science (Section 3.2). Nevertheless, in the initial stages of research, relying on similarities at the level of the causal structures is of little use, since it is precisely these structures that researchers aim to elucidate. In as much as researchers cannot count on prior knowledge of mechanisms, it would seem that the only option would be to justify extrapolations based on known developmental or phylogenetic relatedness between surrogate and target, which are thought to increase the likelihood of causal-mechanistic similarities between surrogate and target (Schaffner [2001]; Weber [2005]; Bolker [2009]; Ankeny and Leonelli [2011]).

While such considerations play a role in evaluating the suitability of surrogates, an analysis of scientific practice shows that researchers tend to adopt a more sophisticated strategy, whereby a comparative degree of relevant similarity is indirectly inferred based on the origin of the experimental setup and the number of modifications required to obtain it. The key concept here is that of keeping track of the differences introduced as a given system is gradually tweaked in order to respond to the demands of experimental research. It is possible to circumvent altogether the problem of developmental/phylogenetic differences if the origin of a surrogate model is the target itself. In our example, the phenomenon of interest can be successfully reproduced in the context of parts of the target system, such as human blood extracts, purified lymphocytes, or T-cell lines. Thus, cell-based experimental setups can be used as experimentally more manageable surrogate models for studying in vivo immune responses, mainly by minimizing the ‘background noise’ of phenomena unrelated to immunity. Parts that have been subjected to fewer modifications are likely to be more similar to the target than parts that have been subjected to more extensive modifications. For example, relevant parts of organisms that are maintained in in vitro conditions that mimic in vivo (physiological) conditions are more similar to whole organism setups than parts that are maintained under more artificial conditions. Likewise, primary cell models are more similar to the whole organism target than immortalized cell lines because the latter are thought to contain further mutations not present in healthy organisms. Similar ranking criteria are applied to surrogate models
derived from systems other than the target, with the notable difference that one has to factor in additional phylogenetic and developmental differences. For example, parts of organisms phylogenetically related to the human target, such as mouse T-cells, can also be used as surrogates in which immunity phenomena can be studied, with the obvious caveat that they score lower in terms of physiological relevance due to the additional risk of introducing species-specific differences.

The above strategy for evaluating and ranking of surrogate models does not eliminate the risk associated with extrapolations. Nevertheless, it keeps it under tight control. By keeping track of each difference introduced, as well as where along the process of generating a given surrogate model differences are introduced, it becomes possible (1) to specify families of complementary models—that is, families of models that do not share the same set of modifications, and thus are not likely to be subjected to the same sources of error; and (2) rank models within each family as better or worse surrogates, according to their relative degree of physiological relevance.

An example of (1) involves complementing studies in human T-cell models with studies involving rodent models. The former do not face the problem of species differences, but introduce the risk that some mechanistic component of immunity present in the whole organism is left out of the surrogate cell model. Conversely, the later family of models avoid this problem, but introduce the risk of species differences. If a finding can be successfully cross-references in both models, it can be inferred that, in as much as that finding is concerned, neither systemic effects nor differences between humans and rodents are likely to be a significant source of error. In more general terms, since each experimental setup has its particular advantages and drawbacks, it seems reasonable to conclude that it is possible to trade one experimental difficulty for another by switching from one experimental setup to another, such that a more complete and accurate knowledge can be attained by studying a phenomenon in a multitude of complementary surrogate models. This is an immensely fruitful research strategy on which biologists rely routinely, to the point that, as I will show in Section 4.1, the diagrams of signalling pathways and mechanisms found in molecular biology textbooks are in fact composite jigsaws in which ‘local snapshots’ gathered in different models are juxtaposed together in order to reveal the ‘big picture’ of the whole pathway or mechanism.

An example of (2) is trusting more extrapolations from primary cells than extrapolations from cell lines. Primary cells are healthy cells extracted from living animals, but differ from in vivo cells because they have been subjected to a process of extraction and short-term cultivation in an in vitro culture medium that matches some, but not necessarily all, aspects of the physiological environment of the living animal. Cell lines are likely to be even more dissimilar from the target because cells are not only grown in vitro,
but, as discussed above, are also pre-cancerous cells that may not always behave in the same way as their primary counterparts. For this reason, extrapolations from cell lines to *in vivo* conditions is considered to be less trustworthy than extrapolations from cell lines to primary cells or from primary cells to *in vivo* conditions.\(^{11}\)

### 3.2 The validation of surrogate models in later stages of basic research and in clinical research

The validation of surrogate models according to a criterion of double similarity—symptom similarity at the level of the phenomenon under investigation and structural similarity at the level of the physical system in which the phenomenon is documented—is preserved in later phases of research, with the notable difference that as knowledge of underlying mechanisms becomes available, the conditions a good surrogate model should satisfy become more stringent. For example, in order to extrapolate from an animal model of human disease, empirical evidence must be provided that the model accurately matches specific attributes of human disease (*Cardiff et al.* [2004]; Bolker [2009]; Piotrowska [2012]). These attributes include descriptive features of the phenomenon (for example, symptoms of AIDS) and, if available, known or potential causes, causally relevant factors, or partially elucidated mechanisms (for example, immunodeficiency symptoms associated with HIV infection). Any given surrogate animal model is judged to be more or less suitable for drawing certain extrapolations, depending on which, how many, and to what degree attributes are matched.

In later stages of research, model validation strategies may also exploit the fact that many biological mechanisms are decomposable in causally modular steps that can be investigated on an independent basis (Woodward [2002]; Steel [2007]). In turn, modularity makes it possible to adopt a ‘divide and conquer’ strategy, whereby different stages in pathogenesis are studied in different surrogate models, each fine-tuned to accurately match a particular stage of pathogenesis. For example, AIDS pathogenesis is decomposable in a consistently recognizable series of stages: transmission, acute infection, latent stage, relapse, complications associated with secondary infections, and

---

\(^{11}\) It is perhaps worth pointing out that physiological relevance has nothing to do with the quality of data generated by an experimental setup. If anything, the quality of data is inversely related to physiological relevance. Researchers can achieve much tighter experimental control in a cell line model than in primary cells or whole organisms, and for this reason data from cell lines are ‘cleaner’ (less background noise) and more precise (leave less space for alternate interpretations), thus providing more compelling evidence for whatever conclusions that data are supposed to support (Baetu [2012], [2013]). The problem is that cell line models generate highly trustworthy findings about what goes on in cell lines, while researchers would like to know if these findings reflect with any reasonable degree of accuracy what goes on in physiologically more relevant models.
death. Extensive knowledge about HIV infection and AIDS in humans was generated by studying SIV infection in rhesus macaques and HIV infection in humanized rodents\textsuperscript{12}; these models are also used to test treatments such as vaccines and antiviral drugs. Each model is validated in respect to the aspects and degree of similarity with HIV infection/AIDS in humans. Macaques are phylogenetically, genetically, and anatomically closer to humans, thus making them better models for studying the mechanisms of disease transmission. On the other hand, SIV is different from HIV (for example, genes specific to HIV; simian AIDS develops within a year of SIV infection, while human AIDS develops five to ten years after HIV infection). Furthermore, due to their long developmental cycle and maintenance costs, macaques are not ideal for genetic experiments; researchers must cope with the potential confounding effects of genetic diversity and the absence of transgenic/knockout organisms. Humanized rodent models overcome these shortcomings. Rodent models are genetically more uniform, their T-cells can be infected by HIV, antiretroviral drugs affect HIV replication in the same way as in human clinical trials, and the same patterns of drug resistance develop over time so that these models are especially useful for predicting clinical antiviral efficacy in humans. At the same time, rodent models are less suitable for studying the mechanisms of viral transmission and AIDS progression due to genetic, anatomical, and lifespan differences between rodents and humans.

A more elaborate strategy that exploits the modularity of biological mechanisms, dubbed ‘comparative process tracing’, is discussed in detail by Daniel Steel (\cite{Steel2007}). Steel’s strategy bears striking similarities with the origin and number of modifications method used in basic science in the sense that both rely on keeping track of the points of difference introduced along the steps of a process, albeit Steel is interested in comparing mechanistic processes while model ranking in basic science tracks points of difference introduced in the process of producing surrogate models.

4 ‘Big Picture’ Accounts and the Extrapolations Underpinning Them

4.1 The mosaic nature of mechanistic descriptions in basic science

Once a phenomenon is characterized, the next step is explaining it. In the life sciences, explaining a phenomenon usually means discovering the mechanism responsible for the phenomenon (\textcite{Weber2005, Bechtel2006, Craver2007});

\textsuperscript{12} Transgenic rats expressing the human cell-surface receptors used by HIV to infect T-cells or, more recently, immunodeficient mice whose immune cells were reconstituted from human stem cell precursors. For an assessment of the use of humanized animal models, see (\textcite{Piotrowska2012}).
In our example, the explanation turned out to revolve around a negative feedback loop regulatory mechanism (Figure 1). The details of the mechanism are not particularly important. What is significant, relative to the problem of extrapolation, is that the description of the mechanism depicted in Figure 1 is in fact a highly condensed summary of findings published in more than 250 original research papers. Given the huge number of studies involved, the kind of information accessible in review papers and textbooks does not include detailed descriptions of the experiments by means of which this information was gathered. However, if we were to pay attention to the experimental details behind each little bit of information, we would quickly realize that Figure 1 is in fact a mosaic in which many ‘local snapshots’ are juxtaposed together in order to reveal a ‘bigger picture’, where each ‘local snapshot’ amounts to a bit of information.

13 This number, taken from my dissertation (Baetu [2001]), includes only papers investigating the role of NF-κB regulatory mechanism in T-cell activation, not the initial papers in which the inflammation responses are characterized.
reconstituted from data from several experiments, conducted in the context of distinct experimental setups, each designed to overcome a particular experimental difficulty. In this example, findings were gathered from humans, human primary cells models, as well as a variety of genetically engineered human cell lines and whole organism transgenic/knockout murine models. The mosaic nature of the big picture account stems from the fact that not all the findings entering its composition were systematically cross-referenced in all the experimental setups used to elucidate the mechanistic basis of the phenomenon, partly due to technological roadblocks, partly to non-negligible pragmatic constraints such as time and cost. Crucially important here, the mosaic nature of the mechanistic account described in Figure 1 is in no way exceptional. Any mechanism elucidated by means of the molecular techniques that have driven advances in biology since the 1970s—which is to say, any molecular mechanism depicted in a recent textbook of molecular biology, cell biology, developmental biology, immunology, microbiology, or any other branch of biology—can be shown to be a mosaic combining results gathered in many distinct experimental models.

4.2 Challenges for mechanistic-similarity-based validation protocols

The notion that models can be ‘validated’ as surrogate objects of investigation from which something can be learned in respect to different target objects has been vigorously criticized (LaFollette and Shanks [1993], [1995], [1996]; Howick et al. [2013]). The main argument is that extrapolations are justified only if there are ‘no causally relevant disanalogies between the model and the thing being modeled’ (LaFollette and Shanks [1995], p. 147). Given that the absence of relevant disanalogies can be ascertained only if one already has access to a relatively advanced knowledge of the mechanisms at work in the two systems, extrapolators are trapped in a vicious circle whereby establishing the suitability of a system as a surrogate model would require already possessing knowledge of the mechanisms at work in the target system, in which case the extrapolation would be unnecessary. The conclusion reached by LaFollette and Shanks ([1996], p. 199) is that the use of surrogate models belongs to the context of discovery and not to that of justification. In other words, surrogate models are excellent tools for generating good candidate hypotheses, but this should not be taken to mean that these hypotheses are in any way justified. In response to this critique, Daniel Steel ([2007], pp. 88–92) shows that, in the context of clinical research, the circle can be broken and new knowledge can be gained given partial knowledge of mechanistic similarities and differences between surrogate and target. With this counter-example in hand, Steel ([2007], pp. 96–9) proceeds to argue that there is no sharp divide between the contexts of discovery and justification,
and thus there is continuity between the use of surrogate models as a means to generate new hypotheses and their use to justify new knowledge.

While Steel makes a good case for the validity of extrapolations, his solution depends crucially on the premise that we know a sufficient amount about mechanistic similarities and differences between surrogate and target. But one has to keep in mind that the partial knowledge of mechanisms needed to strengthen extrapolative inferences in toxicology and other areas of clinical research has to come from somewhere, and this somewhere happens to be basic research. This raises the following conundrum: in order to trust extrapolations based on mechanistic similarities, one must first trust available mechanistic accounts; but these accounts are themselves heavily dependent on extrapolations that must be made in the absence of any previous knowledge of mechanisms.

A second problem to be factored in is that the mechanistic accounts of basic science combine data from tens or even hundreds of distinct experimental setups. There is direct empirical evidence showing that current validation protocols, mechanistic-similarity protocols included, fail to completely eliminate all risks of error associated with extrapolations. It can happen that mechanisms at work in the surrogate model turn out to lack substantive physiological relevance in the target system,\(^14\) or that treatments successful in a surrogate fail in the target,\(^15\) despite the fact that the surrogate model has been thoroughly validated. Thus, even if we have good reasons to assume that any given extrapolation has a high probability of being correct, the probability that knowledge inferred on the basis of multiple extrapolations is correct cannot but decrease as the number of extrapolations involved increases.\(^16\) In turn, this would entail that even if it is justifiable to trust extrapolations on an isolated basis, it would be unjustifiable to accept them systematically. Again, a weak link conundrum arises: since clinical science relies on the results from basic science, it would seem that one has first to accept the high risk of error resulting from compounding many extrapolated results in order to

\(^{14}\) For example, significant differences between molecular pathways at work in Jurkat cell lines and primary T-cells have been eventually discovered (Abraham and Weiss [2004]).

\(^{15}\) For example, the promises and failures of gene therapy (Sheridan [2011]).

\(^{16}\) The problem can be more or less severe depending on how extensively findings are cross-referenced in multiple experimental setups. As a general rule, experimental difficulties render impossible a complete cross-referencing, leaving a considerable margin of error due to multiple extrapolations. The problem can be further aggravated by interpretation issues. Unsuspected but relevant differences between experimental models can make it such that the resulting big picture account is in fact a chimera, an aggregate of findings that does not describe a mechanism that actually exists in any cell or organism. Difficulties surface when the total knowledge struggles to accommodate seemingly incompatible findings, as there is no unique recipe for interpreting such findings. For instance, incompatible findings might be treated as evidence that a balancing regulatory mechanism is at work, that the phenomenon of interest is mediated by alternate mechanisms, or that the proposed mechanism is wrong; or again, these incompatibilities may simply be the unfortunate result of differences between experimental setups.
obtain the mechanistic knowledge necessary to then make some exceptionally trustworthy extrapolations.

The emerging concern is that mechanistic-similarity solutions cannot account for the overall practice of extrapolating results from surrogate models in science. Not only there is a substantial segment of scientific knowledge that cannot benefit from mechanistic-similarity solutions, but as it turns out this knowledge seems to act as a weak link undermining the application of mechanistic-similarity solutions in those contexts in which they could be applied, such as clinical research.

5 Retrospective Testing of Extrapolated Knowledge

5.1 Holistic confirmation

If mechanistic-similarity approaches cannot provide a general solution, then how is the problem of extrapolation dealt with in basic science? The answer may disappoint those seeking in extrapolation a modest means of making inferences that can be trusted in the absence of any further experimental testing. An analysis of research practices supports a mitigated view, according to which validated surrogate models are useful tools for generating reliable, but fallible, starting point assumptions for guiding subsequent research in the target systems (Schaffner [2001]; Weber [2005]). In this sense, defenders of extrapolations are right in claiming that the use of surrogate models is a reliable experimental practice, and an impressive number of monthly articles publishing results gathered in surrogate models clearly demonstrate that extrapolations from validated surrogate models are treated as legitimate justificatory evidence supporting claims about what goes on in the target systems.

Nevertheless, equally well-documented efforts deployed to further test extrapolated knowledge also demonstrate that findings from any given surrogate model are not considered definitive or sufficient evidence, but simply one line of legitimate evidence, to be strengthened and corroborated by other methods of investigation.

To many, this would probably not come as a shock. According to the newly established canons of evidence-based medicine (Glasziou and Del Mar [2003]), the gold standard of clinical research is evidence from randomized controlled trials in humans. Meanwhile, mechanistic evidence, especially if extrapolated from animal models of disease, is relegated to the bottom of the list, as a considerably less reliable type of evidence. It is not my intention to debate the pros and cons of evidence-based medicine. It suffices to observe that, contrary to the views espoused by LaFollette and Shanks, extrapolations from surrogate models are explicitly considered to be evidence and not just sources of new ideas and hypotheses; yet, at the same time, this evidence is
deemed ‘inconclusive’ or ‘insufficient’ for the purposes of justifying the success of a treatment.

The situation in basic research is not that much different. As it turns out, basic research does not end at, and cannot be reduced to, compilations of extrapolated data into big picture accounts of the sort depicted in Figure 1. Irrespective of the quantity and quality of similarities between surrogate models and their physiologically relevant targets, there inevitably comes a point in the life cycle of a research project where extrapolated findings are treated as a partially justified hypothesis about what happens in the target system. Unlike an extrapolation—which is simply accepted as probable, but fallible, knowledge—a hypothesis needs to be subjected to further experimental testing.

One common strategy is holistic testing, that is, the top-down testing of the combined effects of several big picture mechanisms, or of the efficacy of treatments and technological applications designed on the basis of mechanistic knowledge. Again, it is easier to understand how this kind of confirmation works by means of an example. Following on from the earlier case study, it has been hypothesized that HIV infected T-cells die during the final stages of AIDS because the virus activates the negative feedback loop mechanism described in Figure 1, which in turn leads to an up-regulation of an apoptosis-inducing (or programmed-cell-death-inducing) cell surface protein. This prediction was inferred from a variety of findings, each documented in distinct experimental setups:

Finding 1: *In vivo* studies in human AIDS patients and SIV primate models revealed that immunodeficiency is the result of massive death of HIV/SIV infected T-cells, coupled with evidence that cell death is due in large measure to apoptosis and not to damage incurred by the cell during the infection.

Finding 2: *In vitro* studies with immortalized human T-cell lines showed that HIV infection leads to the activation of the same gene expression regulation mechanism involved in the regulation of T-cell activation (Figure 1) via a causal pathway that is only partially understood.

Finding 3: A whole array of studies in genetically modified human cell lines and transgenic/knockout animal models demonstrated that this very same mechanism is also involved in the regulation of the expression of several genes coding for apoptosis-inducing cell surface proteins shown to trigger the death of T-cells.

Thus, it seemed reasonable to hypothesize that if HIV infection somehow leads to the activation of the negative feedback loop mechanism described in Figure 1 (Finding 2), and if this mechanism can trigger programmed cell
death (Finding 3), then this would explain the massive death of HIV-infected T-cells during the late stages of AIDS (Finding 1), which in turn would explain why patients suffer and eventually die from immunodeficiency (they lack T-cells, which are crucial for mounting an immune response).

This hypothesis was tested in a humanized mouse model and, as predicted, the AIDS-like stage of the HIV infection turned out to correlate with a significant increase in the expression of apoptosis-inducing cell surface proteins, and the inactivation of the negative feedback loop mechanism resulted in a decrease in T-cell death (Miura et al. [2001]). The fact that it was possible to make a correct prediction about what would happen in one experimental setup based on inferences from findings in a variety of different experimental setups provided additional evidence that the extrapolations from one model to another are likely to be correct, at least in regard to the specific mechanistic details involved in the inferential reasoning. In turn, this renders this piece of mechanistic knowledge a particularly trustworthy basis for supporting further extrapolations based on mechanistic similarities between various biological systems.

A second wave of results from treatments designed in light of the above findings is expected. The eventual success of such treatments will count as further evidence in favour of the correctness of the big picture mechanisms on the basis which these treatments were designed, as well as the extrapolative assumptions underlying the piecing together of each mechanism from a variety of distinct surrogate models. Conversely, the failure of such treatments may show that somewhere the chain of extrapolations is broken—for instance, there might be an essential difference between stimulation in vivo and stimulation in vitro, or it may turn out that primary cells are, in some relevant aspects, very different from their immortalized cousins, and so on.

5.2 Fallback strategies

As pointed out earlier, it is known that mechanisms at work in the surrogate model may turn out to lack substantive physiological relevance in the target system, and that treatments successful in a surrogate can fail in the target. Nevertheless, if extrapolations fail, this does not mean that everything is lost. This only shows that what is true about experimentally modified cells and of the lab models of a disease is probably not true of the cells and of the disease as they behave in physiologically relevant contexts. The fallback strategy is to contextualize findings to the particular experimental setup in which these

17 Treatments involving a combination of the traditional antiviral treatment plus inhibitors of the negative feedback loop mechanism (anti-inflammatory drugs) and inhibitors of apoptosis-inducing cell surface proteins (such as masking antibodies).
findings were obtained (Baetu [2014]). In the event that extrapolations from surrogate models lead to a dead end—which may range anywhere from a collection of contradictory findings to the catastrophic failure of treatments designed on the basis some big picture account from basic science—researchers can always return to an experimental setup of which a set of findings have been demonstrated to hold true, deploy troubleshooting strategies in order to determine which extrapolations are valid, and then reconstruct the mechanistic accounts accordingly. Common troubleshooting strategies include more careful model validation, attempts to elucidate in more detail a mechanism by relying on a single or a small number of highly similar experimental models before extrapolating findings to more dissimilar models, and attempts to reproduce findings in physiologically more relevant models.

6 Conclusions

Almost two decades ago, LaFollette and Shanks ([1996]) mounted a strong case against the use of animal surrogate models in clinical research on the grounds that extrapolative inferences are useful only as means to generate new hypotheses. In response, Steel ([2007]) pointed out that there is no sharp divide between the contexts of discovery and justification, arguing that when properly regimented by validation protocols, extrapolations are a common and reliable method of generating new knowledge. An analysis of the experimental practice of basic science suggests that the truth is somewhere in the middle. Steel is right in claiming that the use of surrogate models is both a common and a reliable scientific method, and that extrapolations from surrogate to target can be justified by means of a variety of theoretical and experimental considerations. Yet, as he later acknowledges, ‘[i]n real-life cases, the question can only be whether and to what extent the extrapolation strengthens the overall body of evidence’ (Steel [2013], p. 186).

I argue that, given the systematic use of extrapolative inferences and the overall absence of strong indicators of relevant similarity in basic science, the

---

18 In as much as the truth of big picture accounts is contingent on the truth of the fallible inductive extrapolations from one experimental model to another, breadth of knowledge is gained at the expense of certainty. The trade-off between breadth and certainty is illustrated by the differences between textbooks and review articles, on one hand, and original research papers, on the other. Textbooks and review articles aim to synthesize current knowledge about the molecular basis of biological activity in big picture mechanisms and systems of mechanisms, but are often fuzzy about the precise details of the experimental setups in which the various bits and pieces of the mechanisms have been elucidated. In contrast, experimental results published in original research papers focus on individual findings, carefully contextualized to particular experimental setups, which are either operationally characterized in the ‘Materials and Methods’ section of the paper presenting the results, or for which a characterization can be referenced in the scientific literature.
justificatory evidence supporting extrapolations is not deemed definitive, as demonstrated by the subsequent efforts deployed to further test knowledge gained by juxtaposing results from different experimental setups. Extrapolations are merely one epistemic tool to be used in conjunction with other methods of investigation, ranging from cross-referencing findings in complementary surrogate models to clinical trials of treatments. Understanding why extrapolation is an acceptable epistemic practice requires thinking beyond the reliability of individual extrapolations and understanding how extrapolations are used in the context of a much more comprehensive research strategy that combines both a bottom-up process of inferring mechanistic accounts based on experimental data—a process that relies heavily on extrapolations across different experimental setups—and a subsequent top-down testing of predictions made using these accounts. More specifically, I propose that extrapolations are an acceptable epistemic practice not only in light of model validation attempts, but also because they are part of an overall research strategy ensuring that relatively poorly justified extrapolations in the initial stages of research are tested in later stages of research, and that fallback positions make possible the troubleshooting of faulty extrapolations.

References


