1 CRISES IN BIOMEDICAL RESEARCH

> It has been said, indeed, that experiments performed on a dog or a frog may be conclusive in their application to dogs and frogs, but never to man, because man has a physiological and pathological nature proper to himself and different from all other animals. It has been further stated that to be really conclusive for man, experiments would have to be made on man or animals as near to him as possible. It was surely with this idea that Galen chose a monkey for his experiments, and Vesalius a pig, as subjects more closely resembling man in his omnivorous capacity. Even today, many people choose dogs for experiments, not only because it is easier to procure this animal, but also because they think that experiments performed on dogs can more properly be applied to man than those performed on frogs. How well founded are these opinions?
>
> —BERNARD (1927/1957), PP. 122–123

Which animals, if any, should biologists study if they want to learn about human biology? As Claude Bernard, who is often regarded as a founding father of experimental physiology, discussed in the chapter's opening quotation, different people answer this question differently. Bernard himself performed experiments on a variety of different animals—including rabbits, frogs, and dogs—and made some important discoveries, notably the principle of physiological homeostasis (Jørgensen, 2001). However, those same experiments, in conjunction with Bernard's apparently rather cavalier attitude toward animal suffering, also contributed to the development of an organized antivivisection movement in Victorian England, which in turn led to the first (and overdue) legislation to regulate animal experiments (French, 1975; Guerrini, 2003).

Ever since those early days of experimental biology, societies have sought some sort of compromise between the desire to minimize animal suffering and the quest for knowledge and specific benefits to human health. The balance of these conflicting motivations has shifted over the years and varied across societies (e.g., Francione,

1996; Caruana, 2020). However, a fairly broad consensus has formed around the notion that scientists should minimize the suffering of animals and use "lower" animals as much as possible. As Russell and Burch (1959) discussed in their influential book on the "3Rs," scientists should *refine* experiments, *reduce* animal numbers, and *replace* sentient animals (i.e., animals presumed to have feelings) whenever this is possible without preventing scientists from answering important questions. According to Russell and Burch, the third of these recommendations—replacement—means that experimentalists should work on non-mammals whenever possible or avoid animals entirely by studying cultured cells or computer models. Similarly, the *Guide for the Care and Use of Laboratory Animals* states that researchers should avoid using animals or replace vertebrate animals with "animals that are lower on the phylogenetic scale" whenever possible (National Research Council, 2011b, p. 5).

These guidelines are well-intentioned but leave many questions unanswered. Given that the notion of a "phylogenetic scale" has long been disavowed by evolutionary biologists (Hodos & Campbell, 1969), which animals are lower than others? Is it justifiable to kill numerous lower animals when the findings do not extrapolate to humans? When is it sufficient to work with cultured cells or computer simulations rather than animals? Is it better to work on intact lower animals or cultured human cells? Should the choice of animal species depend on the study's purpose? Does it matter whether scientists are using the animals to test a novel cosmetic or a life-saving COVID-19 vaccine? And is it okay to kill any animals for research that does not promise direct, immediate benefits to human health? Again, different people tend to answer these questions differently. Among biologists, individuals must find their own comfort zone, balancing the perceived importance of their research with their own animal welfare concerns as well as the relevant regulations (Kwon et al., 2010; Franco et al., 2018). Nonscientists as well balance these factors in diverse ways and, therefore, vary in their attitudes toward animal research (Joffe et al., 2016; Bradley et al., 2020).

Importantly, the compromises involved in selecting research subjects have not received as much attention as they deserve, at least among biologists (Orzack, 2012). Anti-vivisectionists do discuss these issues, but their positions tend to be so fervently against all animal research that biologists find them unworkable. Historians and philosophers of science have scrutinized how biologists select their models and use them (e.g., Burian, 1993; Bolker, 2019; Ankeny & Leonelli, 2020; Dietrich et al., 2020), but most biologists pay little heed. Biologists do occasionally pen thoughtful commentaries or reviews on the pros and cons of various model systems (see chapter 2), but the vast majority of these papers advocate primarily for the model that the authors themselves are working on. Few are willing to critique the choices of other biologists,

perhaps because they worry about being associated with animal rights activists or exposing their colleagues to extremist attacks (Endersby, 2007, p. 408).

Meanwhile, the balance of which models receive the lion's share of research attention shifts slowly over time (see chapters 3 and 4). One hears occasional outcries from those whose favored models are being phased out, but otherwise the changes receive only fleeting attention. This situation would be acceptable if the research consistently achieved its stated aims. However, biologists over the last two decades have realized that much of the supposedly "translational" research on model systems fails to generalize to humans, at least as judged by the alarmingly high failure rate of clinical trials.

1.1 THE TRANSLATABILITY CRISIS

Louis Pasteur, who developed some of the earliest vaccines (see chapter 5), once wrote that "there are no such things as applied sciences, only applications of science . . . which are related to one another as the fruit is related to the tree that has borne it" (see Wellems, 2010). This is true, I think, but distinctions between "basic" and "applied" science continue to be popular. One reason for this persistence is that, over the last twenty years or so, societies around the globe have increasingly pushed scientists to demonstrate that their research has tangible benefits (Zerhouni, 2003; Maienschein et al., 2008; van der Laan & Boenink, 2015). Within biology, those benefits include the development of novel biofuels, more productive crops, and bacteria that can digest plastic (Ru et al., 2020). However, the vast majority of applied research in biology focuses on the development of novel diagnostics, drugs, and other therapies. Such efforts are often called *translational research* because their aim is to "translate" laboratory discoveries into tools that physicians can use. Although one can distinguish various types or phases of translational research (Sung et al., 2003; Fort et al., 2017), the research covered in this book involves mainly efforts to apply knowledge obtained from animals or in vitro preparations (i.e., preclinical research) to the enhancement of human health (i.e., clinical research).

Before new drugs or vaccines can be used in humans, they must be approved by regulatory agencies such as the Food and Drug Administration (FDA) in the United States. This regulatory approval usually requires a series of human clinical trials to show that the compounds are safe and effective, unless such trials would be unethical or simply not feasible (e.g., if they involve exposure to chemical weapons). The clinical trials, in turn, are usually based on preclinical research with cultured cells and studies on one or more animal species. Sadly, each step along this process tends to be more expensive than the previous, with late-phase clinical trials often costing hundreds of

millions of dollars, and the entire process typically taking a decade or more. It is profoundly distressing, therefore, to learn that the vast majority of promising compounds that have emerged from animal research have failed in clinical trials.

For example, Kola and Landis (2004) examined the fate of all the drugs developed by the 10 largest pharmaceutical companies between 1991 and 2000. The average rate of success for all these drugs going from first-in-human testing to regulatory approval in the United States and Europe was just 11%. Cardiovascular therapies were approved at a rate of 20%, but drugs for cancer and neurological disorders succeeded only 5% and 8% of the time, respectively. Most of the attrition occurred at relatively late stages of clinical development, when major costs had already been incurred; the main reasons for the failures were a lack of efficacy, toxic side effects, or other safety concerns. A larger analysis of 1,738 drugs developed by the top 50 pharmaceutical companies between 1993 and 2004 revealed a slightly higher mean success rate of 16% (DiMasi et al., 2010). Promising as this increase appeared, an even larger study of 4,451 drugs developed by 738 biotech and pharmaceutical companies between 2003 and 2011 once again yielded an overall success rate of merely 10%, with cancer, cardiovascular, and neurological drugs exhibiting the lowest approval rates (Hay et al., 2014).

The picture is even bleaker when you focus on specific disorders. Most depressing is the failure rate for drugs targeting Alzheimer's disease, which was 99.6% for the period between 2002 and 2012 (Cummings et al., 2014). Many additional Alzheimer's drug trials have been conducted since then, but all of them have failed. (The drug aducanumab received accelerated approval by the US Food and Drug Administration [FDA] while this book was in press, but the FDA's own panel of experts had previously recommended against its approval [Hoffman, 2020; FDA Commissioner, 2021].) In light of this high failure rate, several large pharmaceutical companies have closed their neurological divisions. Indeed, across all therapeutic areas, the number of new drugs approved per dollar spent on research and development has declined substantially, and rather steadily, since 1950 (figure 1.1A). This does not bode well for the future of drug companies or, more importantly, patients. There are some indications that approval rates have recently increased in a few therapeutic areas (e.g., rheumatoid arthritis and cancer immunotherapy), but the overall success rate for novel therapies remains distressingly anemic.

Multiple reasons likely account for this persistent crisis of biomedical translation. For one thing, novel drugs must generally be more effective (or have fewer side effects) than older drugs that are already available, which makes it harder for the newer drugs to get approved. In addition, the approval process has generally become stricter, although the FDA has occasionally loosened a few rules. Other likely explanations are that some of the clinical trials may have been flawed in some way (e.g., in the composition of their subject pool), or that some of the preclinical models were not, in fact, good

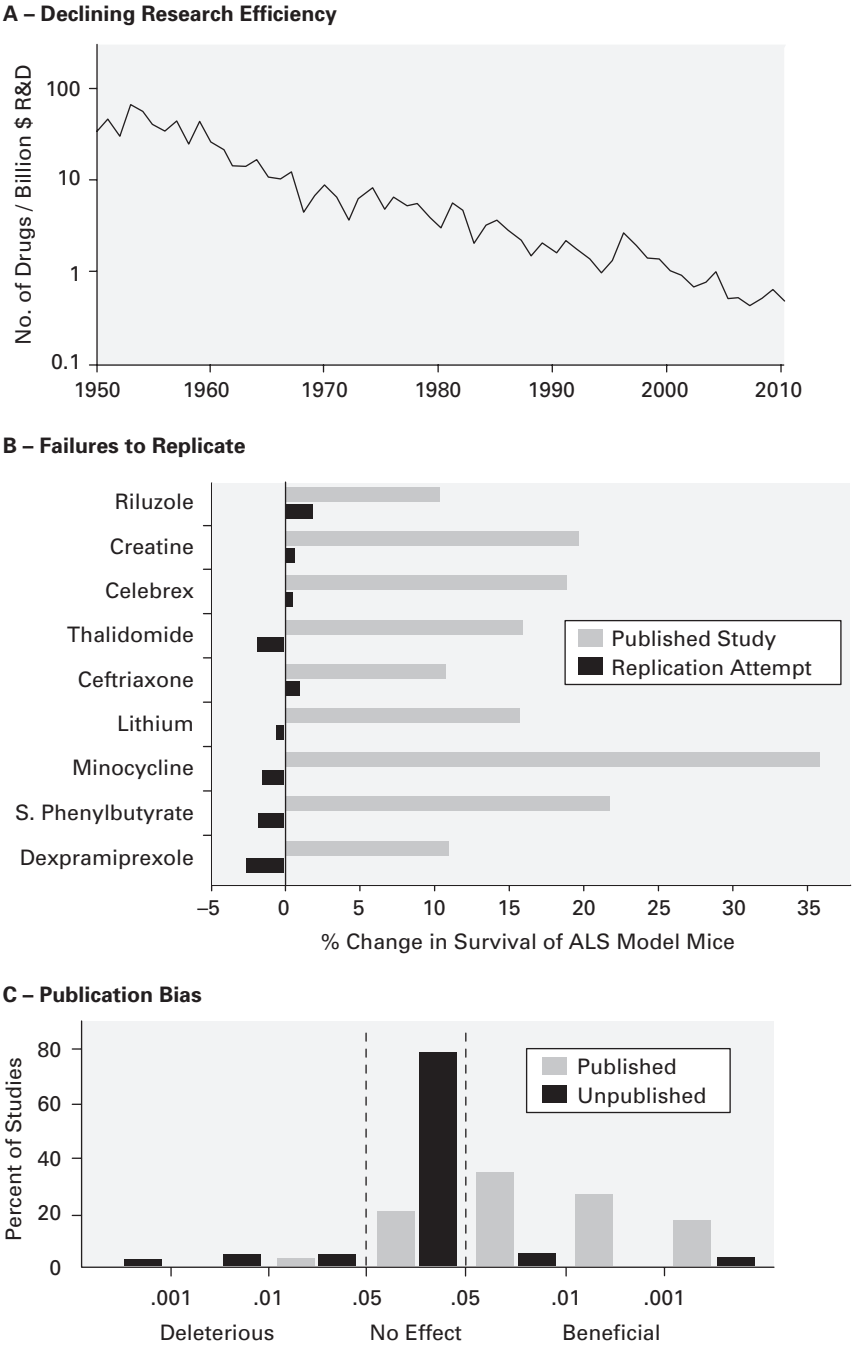


Figure 1.1

Crises in biomedical research. (*A*) Declining research efficiency. The number of new FDA-approved drugs has decreased substantially over the last 60 years, relative to the inflation-adjusted investment in research and development (R&D). (*B*) Failures to replicate. The ALS Therapy Development Institute tested nine compounds in amyotrophic lateral sclerosis (ALS) model mice that had previously been reported to be beneficial in mice. The graph shows how they could not replicate the previously published findings. Eight of the illustrated compounds also failed in human clinical trials; only riluzole is approved for ALS. (*C*) Publication bias. A comparison between 42 published studies on motor (rotarod) performance in R6/2 mouse models of Huntington's disease and analogous unpublished studies, performed by PsychoGenics, Inc., shows that the published studies are biased toward more positive results (*x*-axis shows *p* values). Adapted from (A) Scannell et al. (2012); (B) Perrin (2014); (C) Brunner et al. (2012).

models for the human condition. We will explore this topic thoroughly in later chapters. Finally, it is quite possible that in many cases the preclinical research was simply not strong enough to support the clinical trials; this hypothesis is supported by the dawning realization that many findings in biology (and other fields) do not replicate reliably.

1.2 THE REPLICABILITY CRISIS

Since pharmaceutical companies stand to lose large sums of money if their clinical trials do not succeed, it is not surprising that they usually attempt to validate the preclinical research before proceeding to trials. When scientists from Bayer Health Care examined their own efforts in this regard for a period of four years and including 67 separate projects (focused on cancer, women's health, and cardiovascular therapies), they found that the published preclinical data were "completely in line" with their in-house preclinical results in only 20% to 25% of their projects (Prinz et al., 2011). More than half of the cases produced significant inconsistencies. According to the authors of that study, these findings are consistent with the conventional wisdom among venture capitalists that "at least 50% of published studies, even those in top-tier academic journals, can't be repeated with the same conclusions by an industrial lab" (Prinz et al., 2011, para. 7). An analogous effort at another large biotechnology company revealed that only 6 of 53 "landmark" preclinical studies on blood diseases and cancer could be replicated in-house (Begley & Ellis, 2012).

Even more frustrating results were reported by a nonprofit biotechnology company that focuses exclusively on the development of therapies for amyotrophic lateral sclerosis (ALS). They screened more than 70 drugs in 18,000 mice and found "no statistically significant positive (or negative) effects for any of the 70 compounds tested, including several previously reported as efficacious" (Scott et al., 2008, p. 5). The authors blamed the irreproducibility mainly on experimental design flaws in the earlier studies, which generally lacked the statistical power to overcome the variability of the studied animal models. A later study from the same company reported more specifically on eight compounds that had emerged as especially promising in the preclinical research but ultimately failed in clinical trials (figure 1.1B). When the company tried to replicate those original studies, the failure rate was 100% (Perrin, 2014). Similarly, a large-scale effort funded by the National Institutes of Health (NIH) to replicate preclinical studies on spinal cord injury revealed that the majority could not be replicated by other investigators (Steward et al., 2012).

All these failures to replicate previous studies are part of a general reproducibility crisis in science (Kafkafi et al., 2018). In psychology, for example, several large-scale efforts to replicate major findings yielded success rates of only 39% to 54% (Open

Science Collaboration, 2015; Klein et al., 2018). In the social sciences, the replication rate for 21 studies published in *Nature* and *Science* between 2010 and 2015 was 63% (Camerer et al. 2018). Even highly cited clinical studies are not immune to this problem, although it does appear that the replicability of clinical studies is greater than that of the preclinical research (Ioannidis, 2005b; Collins & Tabak, 2014). In aggregate, these findings prompted the leadership of the NIH to declare that "the irreproducibility of significant numbers of biomedical-research publications demands immediate and substantive action" (Collins & Tabak, 2014, p. 613). Fighting this problem is not an easy matter, however, as it stems from multiple causes.

Some of the blame falls on poor experimental design, such as experimenters assigning subjects to treatment groups nonrandomly, not being "blinded" to those assignments when they perform their data analyses, or performing interim statistical analyses to determine whether the sample size should be increased (Landis et al., 2012; Steward, 2016). Moreover, the majority of preclinical studies, as well as many clinical trials, are underpowered statistically, meaning that they do not include enough experimental subjects to reach robust conclusions on the hypothesized effects (Ioannidis, 2005a; O'Collins et al., 2017). These insufficient sample sizes do not just make it difficult to demonstrate true positive effects but also increase the likelihood of reporting false-positive results (Button et al., 2013a, 2013b).

A related problem is that experimenters usually perform many different experiments, as well as multiple analyses, but they tend to publish only their positive results (figure 1.1C) (Turner et al., 2008; Brunner et al., 2012; Drucker, 2016). Those results are typically reported as having a less than 5% chance of being false, but the statistical calculations tend not to include the failed efforts (Simmons et al., 2011; Tsilidis et al., 2013). A similar problem plagues high-throughput studies that test thousands of hypotheses at the same time (e.g., transcriptome analyses). Most of these studies do titrate the rate of likely false discoveries to some generally acceptable level (Benjamini & Hochberg, 1995), but the statistical procedures for combining the results of such studies across multiple iterations (i.e., replications) are nontrivial (Amar et al., 2017).

One way to fight this crisis of replication is to train young scientists more thoroughly in proper experimental design and statistics (Howells et al., 2014), which should help them realize that some time-honored methods for obtaining good results are in the long run misleading. In addition, scientists should be encouraged to preregister their experiments, describing what will be done and how, and then report even negative results. This approach extends the methodology of clinical trials to preclinical research and is becoming more popular in a few disciplines (Nosek et al., 2018).

However, even with preregistration, a non-negligible number of the statistically positive results would likely be due to chance (Colquhoun, 2014). Moreover, a strict

requirement for preregistration may stifle exploratory inquiry by reducing the frequency of unexpected but ultimately important discoveries. To mitigate the latter dilemma, some authors have proposed that extreme experimental rigor should be required only for preclinical research that is about to become the basis for a clinical trial (Mogil & Macleod, 2017). If such preclinical trials fail, then canceling the clinical trials would save a great deal of effort and money and, importantly, minimize false hope.

1.3 RECKONING WITH BIOLOGICAL VARIATION

Although it may be tempting to attribute the current translatability crisis solely to replicability issues, this is probably wishful thinking. A pernicious additional problem is that many scientists are insufficiently cautious about extrapolating findings from their particular model systems to a much broader target population. To combat this bad habit in psychology, some authors have suggested that each published research paper should be accompanied by an explicit "constraints on generality" statement (Simons et al., 2017).

In the biomedical context, the analogous problem is overly optimistic generalization from the preclinical models to human patients. For example, researchers (and press releases!) often fail to highlight that the manipulations used to generate a disease-like state in animal or cell culture models tend to be poor or merely partial imitations of what causes the human disease. More specifically, a transgenic mouse carrying a human "disease gene" may not accurately model the human disease, if that disease has other or additional causes (Garner, 2014).

Unfortunately, the idea that species differences may be relevant to the preclinical modeling of human diseases is often neglected. As a former director of the NIH once observed, it is not unusual for biological psychiatrists to "assume that a mouse is a small rat, a rat is a small monkey, a monkey is a small human, and that all of these are 'models' for studying abnormal behavior or abnormal brain function in humans" (Insel, 2007, p. 1337). More generally, many biologists seem to think that "the fish is a frog . . . is a chicken . . . is a mouse" (Kimmel, 1989) and that, looking back on the last few decades of biological research, "it did not matter which animal you chose—fundamental processes were fundamentally conserved" (Grunwald & Eisen, 2002, p. 722).

This strong belief in our ability to generalize from specific findings in a few species to biology in general was boosted enormously by the success of genetics and molecular biology in revealing a number of broadly conserved principles of life. This optimism began to weaken, however, when biologists started to realize that, even within the human species, unrestrained generalization can cause real harm.

1.3.1 Taking Sex Differences Seriously

For many years, women were rarely included in clinical trials, largely because scientists wanted to protect them and their potential fetuses from the possibly detrimental effects of novel drugs. It eventually became apparent, however, that this approach shortchanged women by neglecting the possibility of biological sex differences and simply assuming that the results obtained in men would generalize (Clayton & Collins, 2014). As we now know, this assumption is quite often false: women occasionally exhibit different drug side effects than men, require different drug dosages, experience different symptoms for the same health issues (e.g., a heart attack), and benefit maximally from different therapies (Neigh & Mitzelfelt, 2016; Westergaard et al., 2019). Furthermore, some diseases that affect predominantly women were relatively neglected in the male-dominated clinical research. In recognition of these inequities, the NIH Revitalization Act of 1993 mandated the inclusion of women in clinical research (Mazure & Jones, 2015). Pursuant to this legislation, just over half of NIH-funded clinical research participants were women by 2014 (Clayton & Collins, 2014).

The inclusion of female subjects in preclinical research has proceeded at a much slower pace, especially in pharmacology and neuroscience. It was only in 2016 that the NIH officially mandated that all preclinical research must include sex as a biological variable, unless strongly justified otherwise. This edict received some pushback at the time, in part because scientists believed that cyclical variations in hormone levels made the results obtained from females more variable than those obtained from males. This assumption has now been falsified for several research domains (Beery, 2018).

However, researchers also complained that analyzing their results for males and females separately would force them to increase their animal numbers, which would increase experimental costs and duration (as well as animal suffering). Indeed, testing for sex differences can require more animals when differences exist, but being unaware of those differences can cause important, sex-specific findings to be missed (Cahill & Hall, 2017).

Fortunately, it has now become increasingly common for preclinical studies to include both males and females, and to report sex differences when they exist (Arnegard et al., 2020). In fact, it seems to me that the existence of sex differences is currently more broadly accepted for nonhuman animals than for humans, especially in domains other than reproduction (Cahill, 2014).

1.3.2 Personalized Medicine

The NIH Revitalization Act of 1993 also specified that NIH-sponsored clinical trials should include "members of minority groups and their subpopulations." This clause was included, at least in part, because minority groups were, like women,

underrepresented in many earlier clinical trials (Nazha et al., 2019). This lack of proportional representation probably caused minorities to miss out on some therapies that might otherwise have been developed for them. Moreover, the underrepresentation surely led to a relative lack of information about drug dosages and side effects for some of those minorities.

Indeed, significant health disparities for some racial or ethnic groups have been reported frequently. For example, self-described Blacks/African Americans in the United States die of cancer far more frequently than non-Hispanic Whites, and their infant mortality is more than twice as high (Office of Minority Health, 2020). These health disparities arise in large measure from socioeconomic and environmental inequalities, but more inclusive clinical trials would surely provide more data relevant to those inequities.

A case in point is the discovery that a heart failure treatment, called BiDil, improves survival rates and time-to-hospitalization in self-identified Black patients (Temple & Stockbridge, 2007). This drug was first tested in a large clinical trial that revealed no statistically significant benefits for the patient population as a whole. However, a post hoc analysis indicated a significant reduction in heart attack mortality for the Black/African American subpopulation. Based largely on this finding, the FDA suggested that the company should do a follow-up trial specifically on Black patients (because a more inclusive trial would have been prohibitively expensive). This subsequent trial showed that BiDil reduced mortality by 43%, which then prompted the FDA to approve the drug specifically for African Americans. Critics complained that this selective approval risked stereotyping African Americans and was premature in any case, given that it remains unclear why the treatment has different effects in different subpopulations. The FDA responded that the mechanisms of action remain unknown for many drugs and that this ignorance is not enough to withhold approval, especially when such a withholding might well be deemed unethical.

A more complicated issue is that self-reported race or ethnicity is, in the words of NIH director Francis Collins, an "imperfect proxy" for biological or environmental differences that might be medically relevant (Collins, 2004). Indeed, contemporary biologists have pointed out that genetic variation between racial or ethnic groups is significantly smaller than the variation within these groups (Christensen, 2004), suggesting that the biological concept of race in humans is blurry at best. In fact, some "race-specific" diseases, such as sickle cell anemia, are less specific than commonly assumed and, therefore, often fail to be diagnosed in people of other races (Yudell et al., 2016). Because of those blurred boundaries between human races, and genetic heterogeneity among humans in general, biologists have increasingly focused not on race but on specific molecular features that correlate with disease risks and therapies.

This search for biomarkers that can guide medical research and treatments has been called precision medicine (National Research Council, 2011a). However, its original name—personalized medicine—seems more appropriate because it emphasizes that humans differ from one another in a wide variety of ways and may, therefore, benefit from personalized medical treatment.

Personalized medicine has been most successful in cancer research and therapy, mainly because the DNA of tumor biopsy samples can nowadays be sequenced rapidly, revealing mutations that are likely to have transformed the cells (i.e., made them cancerous). Scientists can then examine which of the many available cancer cell lines have similar mutations and how they had responded to diverse treatments in cell culture (Keshava et al., 2019). Such in vitro information can then be translated back to the human patient, giving them the treatment that worked best in the cultured cells and, hopefully, will be ideal for the patient as well. This approach is not without problems, but it works reasonably well for some types of cancer (see chapter 5).

Whether personalized medicine will be equally successful for other diseases remains unclear because the associations between most diseases and specific genes are relatively weak and do not readily suggest ideal treatments. In addition, extensive subdivision of the subject pools in clinical trials reduces the statistical power of the individual analyses, which weakens the evidence underlying highly personalized therapies (Djulbegovic & Ioannidis, 2018; Kimmelman & Tannock, 2018). Still, the number of drugs approved for personalized therapies has increased over the last decade, suggesting that the general approach is promising (Personalized Medicine Coalition, 2019).

1.3.3 The Differences That Broke a Clinical Trial

If sex and population differences deserve consideration in clinical trials, then the possibility of differences between humans and the animals used in preclinical research should also be considered carefully. Much of this book is about such differences, but a powerful introduction to this topic is provided by a famous clinical trial that ended disastrously, namely the trial of drug TGN1412. This trial is sometimes cited as evidence for the general claim that animal research cannot predict how humans will respond to test compounds, but the case is more nuanced and interesting than that.

TGN1412 was a genetically engineered monoclonal antibody against a signaling molecule (called CD28) that is located on the surface of specialized white blood cells, namely T cells. Initial experiments in laboratory rats showed that injections of TGN1412 caused a subset of these T cells (regulatory T cells) to proliferate and release molecules that reduce inflammation. This finding suggested that TGN1412 might be an effective treatment against autoinflammatory diseases such as rheumatoid arthritis. This hypothesis was supported by experiments on cultured T cells from humans. The

investigators then injected various doses of TGN1412 into macaque monkeys and found that the animals tolerated at least 50 mg/kg of the drug. On the basis of these preclinical data, the company developing TGN1412 received approval to conduct a small clinical trial. On March 13, 2006, six men were injected with 0.1 mg/kg of the drug.

Sadly, in the words of a participating researcher, "on this day, scientific excitement turned into a nightmare" (Hünig, 2016, p. 3325). Within an hour of receiving the drug, all six patients developed headaches, nausea, and strong back pain. In the next few hours they developed high fevers and multiple organ failure. All were transferred to intensive care units; two of them required hospitalization for more than a week, and some lost fingertips and toes. Further analysis revealed that the drug had triggered a cytokine storm, defined as a massive release of inflammatory molecules (cytokines). Thus, the drug's effect was functionally the opposite of what the researchers had expected on the basis of their animal and in vitro studies. The company soon went bankrupt, but the question remained: what went wrong?

Follow-up studies revealed that the response of cultured human white blood cells to TGN1412 depends on how the drug is presented to the cells. In the original experiments, both the drug and the cells had floated freely in the culture medium, and under those conditions the cells do not release inflammatory cytokines. However, when the drug was attached to the inner surface of the culture dish before adding the cells, presumably mimicking the natural condition more closely, the cells released large amounts of the inflammatory cytokines (Stebbings et al., 2007). A separate set of experiments revealed that TGN1412 can also activate white blood cells when they are precultured at high density (Hünig, 2012). Either way, we can conclude that, had the original in vitro experiments been designed just slightly differently, they would have raised safety concerns much earlier, preventing the calamity.

Even more interesting for present purposes is that white blood cells from macaques do not, in contrast to their human counterparts, release inflammatory cytokines in response to TGN1412, even when the drug is bonded to the culture dish (Stebbings et al., 2007). The explanation for this species difference lies in a specific subset of the white blood cells, namely the memory effector T cells. In humans these cells express CD28 and release inflammatory cytokines upon activation by TGN1412. Macaques also have memory effector T cells, but they do not express CD28, so they cannot be activated by TGN1412 (Eastwood et al., 2010). Putting it all together, we can conclude that the safety testing of TGN1412 in macaques failed to trigger the cytokine storm observed in the human volunteers because the drug activates the cytokine-releasing T cells only in humans, not macaques. Because the in vivo testing of TGN1412 in rats had generated the desired anti-inflammatory response without triggering a cytokine

storm, it appears that rats and monkeys have similar T cells, making humans the oddballs here. This hypothesis has not been tested directly, but later studies showed that human immune cells tend to be hyperreactive even in comparison with those of chimpanzees (Soto et al., 2010).

One lesson scientists took away from the TGN1412 trial was that experimental drugs should not be given to multiple subjects at the same time. The drugs should also be injected more incrementally so that the trial can be aborted before unanticipated negative effects become severe. In addition, regulators began to recommend that dosages for first-in-humans trials should be based on minimum effective doses, rather than tolerability (Stebbings et al. 2009).

Less widely recognized is what the TGN1412 trial taught us about the dangers of extrapolating from animal and in vitro models to humans. Clearly, subtle differences in culture conditions can have profound effects on experimental results and so can species differences, even when the species are as closely related as humans and macaques. One may ask, therefore, how we can improve our ability to predict which preclinical findings will generalize to human patients, and which will not. More generally, how can we get better at predicting which species or in vitro preparations will be good models for any given human condition?

1.4 WHICH MODEL IS "BEST"?

One response to failed clinical trials is to declare preclinical research worthless and recommend that novel therapies be developed solely in humans (Horrobin, 2003). Although good arguments for more extensive clinical research can certainly be made, abandoning safety and efficacy testing in animals is unlikely to increase the rate of drug discovery, except perhaps in cases of an ongoing pandemic (these words were written in the midst of the COVID-19 scourge). The suffering of laboratory animals would certainly decrease if animal research were substantially curtailed, but veterinary medicine would be held back and human suffering would surely escalate. Nor is working with cultured human cells the panacea that some claim it to be. As TGN1412 exemplified, in vitro experiments often fail to mimic the in vivo conditions. So, if we accept that animal experiments are indispensable, then one must ask, as Claude Bernard did in the chapter's opening quotation, which species and laboratory strains should scientists select for their research?

The most common answer to this question is that it depends on the research question, but what exactly does that mean? Often it simply means that a researcher wants to do a particular type of experiment that can only be done, or is most conveniently

performed, in a particular species, strain, or in vitro system. This is a reasonable rationale, but if the results from that experiment cannot be generalized as broadly as the researcher had hoped, is the rationale still good? If the larger aim of the research is to improve human health but the experimental findings do not translate to humans, is the research still justified? Many scientists would argue that the answer is yes, because such findings would still contribute to a larger store of biological knowledge that may yield unexpected dividends at some point down the road. This, too, seems fair enough, because the benefits of pure, basic research are widely recognized (Comroe & Dripps, 1974; Fricker, 2016; Flexner, 2017; Spector et al., 2018). Still, all experimental biologists, especially those just starting out in their careers (Yartsev, 2017), would likely benefit from thinking more comprehensively about how and why they selected their preferred model systems. This book will hopefully assist them in this task.

1.5 THIS BOOK'S APPROACH AND ORGANIZATION

My aim in this book is to review the question of model selection in biomedical research from a variety of perspectives, ranging from philosophy and history to the perspective of practicing biologists. Balancing the sometimes competing perspectives on biological model systems is difficult, and most readers are likely to object to some of the stated viewpoints. However, my overarching goal in this book is not to convince you of any specific position, but to provide you with a broad array of information, putative diagnoses, and food for thought that you can use to reach conclusions of your own.

As a first step in this effort, chapter 2 reviews some philosophical concepts relating to model systems. Many biologists have little patience for philosophy, and the famous physicist Richard Feynman supposedly once said that "philosophy of science is about as useful to scientists as ornithology is to birds" (quoted in Trubody, 2016). However, dismissing philosophy is not the same as having no philosophy, and Feynman actually knew the subject well. Personally, I have often found philosophical discussions of biology to be enlightening (e.g., LaFollette & Shanks, 1993; Bolker, 1995, 2017; Ankeny & Leonelli, 2011; Noble, 2011; Parkkinen 2017), and some insights from this work are distributed throughout this book. At a minimum, being aware of one's own assumptions about the scientific process can certainly not hurt (Burian, 1993).

To clarify what working with a model means, I distinguish between abstract models and material models. Both tend to be simpler than their target (i.e., the system being modeled), but the former are explicitly stripped down to some core elements, whereas the latter (which include all cell culture and animal models) come with their own complexities, which are partly unknown and may be undesirable. The chapter also explores some of the main assumptions biologists tend to make about how animals resemble

one another and how those similarities vary with phylogenetic distance and biological level. These factors sometimes come into conflict with ethical concerns about animal suffering, which leads to complex compromises and cognitive dissonances that are often highly personal and rarely expressed.

Chapters 3 and 4 trace the history of the most widely used animal and in vitro models, respectively. Instead of focusing on the role of individual scientists in developing these models, the chapters foreground the models themselves: how their popularity has waxed or waned, and how the animals and cultured cells were sometimes modified to suit the research purposes. Special emphasis is placed on the kinds of research questions that the model species helped to answer, how those questions changed over the years, and why the species were selected initially. These chapters also address historical changes in societal attitudes toward animal welfare and the legislation that was passed to regulate some types of animal research. Overall, the various animal and in vitro models compete with one another in a sort of ecosystem of models. In such a system, the rise and dominance of specific models is usually adaptive, but it can, at times, hinder medical progress.

Chapters 5 and 6 review how animal and cell culture models have been used to develop therapies for bacterial and viral infections, cancer, cardiovascular diseases, and (in chapter 6) a variety of neurological disorders. None of these disease groups is discussed in detail; instead, emphasis is placed on how research on these types of diseases has taken advantage of, or been hampered by, model system differences. In the fight against infectious diseases, for example, the development of novel in vitro models greatly reduced the need for research animals, though treatments are still tested for safety and efficacy in animals, including nonhuman primates. By contrast, the development of surgical treatments for heart disease relied more heavily on animal research, especially large animal models (e.g., dogs). Cancer research has employed a variety of in vitro and animal models, including hybrid models that combine the two. Its special challenge is that cancer consists of many different cancer types; nonetheless, some good treatments for select cancer types are now available. The neurological disorders have been very difficult to study in model systems, and therapy development has been frustratingly slow. It is often said that the various models are partial or incomplete, but calling them "imperfect" may be more accurate.

The book's final chapter attempts to diagnose the basic problems that make biomedical research on model systems so challenging. This diagnosis is presented in the form of four different perspectives on the challenges and their potential solutions. Although these four perspectives are distinct, they are largely compatible with one another. In fact, the field as a whole can accommodate them all. I finish the book with some specific recommendations: (1) know your animals and cells, (2) standardize, but not

too much, (3) learn from clinical trial failures, (4) embrace diversity, and (5) reckon with complexity.

Overall, the book offers no facile solution to the translatability crisis in biology, but it explores the problem and its underlying causes in considerable depth and from a variety of perspectives. Hopefully some of the presented information and analyses will stimulate you, the reader, to engage in further thought. Perhaps some of those thoughts will make a difference.

This is a section of doi:10.7551/mitpress/14366.001.0001

Model Systems in Biology

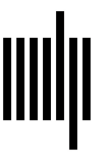
History, Philosophy, and Practical Concerns

By: Georg Striedter

Citation:

Model Systems in Biology: History, Philosophy, and Practical Concerns
By: Georg Striedter
DOI: 10.7551/mitpress/14366.001.0001
ISBN (electronic): 9780262370028
Publisher: The MIT Press
Published: 2022

The open access edition of this book was made possible by generous funding and support from MIT Press Direct to Open



The MIT Press

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Adobe Garamond Pro and Berthold Akzidenz Grotesk by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Striedter, Georg F., 1962– author.
Title: Model systems in biology : history, philosophy, and practical concerns / Georg Striedter.
Description: Cambridge, Massachusetts : The MIT Press, [2022] | Includes bibliographical references and index.
Identifiers: LCCN 2021033979 | ISBN 9780262046947 (hardcover)
Subjects: LCSH: Animal models in research. | Animal experimentation.
Classification: LCC R853.A53 S77 2022 | DDC 616.02/7—dc23
LC record available at https://lccn.loc.gov/2021033979