Microsoft Corporation
One Microsoft Way
Redmond, WA 98052-6399

Tel 425 882 8080
Fax 425 706 7329
www.microsoft.com

Microsoft

Learning from other Domains to Advance AI Evaluation and Testing

We are grateful to the authors of the enclosed expert report, which forms part of a broader series commissioned by Microsoft.

These reports were commissioned as part of Microsoft's effort to draw lessons from other domains to strengthen testing and evaluation as a cornerstone of AI governance.

The insights contained in this report reflect the authors' independent analysis and expertise. The views expressed are those of the authors alone.

We thank the authors for their intellectual generosity and thoughtful engagement throughout this project.

Full Series

Civil aviation: *Testing in Aircraft Design and Manufacturing* by Paul Alp.

Cybersecurity: *Cybersecurity Standards and Testing—Lessons for AI Safety and Security*, by Stewart Baker.

Financial services (bank stress testing): *The Evolving Use of Bank Stress Tests*, by Kathryn Judge.

Genome editing: *Governance of Genome Editing in Human Therapeutics and Agricultural Applications*, by Alta Charo and Andy Greenfield.

Medical devices: *Medical Device Testing: Regulatory Requirements, Evolution and Lessons for AI Governance*, by Mateo Aboy and Timo Minssen.

Nanoscience: *The regulatory landscape of nanoscience and nanotechnology, and applications to future AI regulation*, by Jennifer Dionne.

Nuclear energy: *Testing in the Nuclear Industry*, by Pablo Cantero and Gerónimo Poletto Antonacci.

Pharmaceuticals: *The History and Evolution of Testing in Pharmaceutical Regulation*, by Daniel Benamouzig and Daniel Carpenter.

January 2025

The History and Evolution of Testing in Pharmaceutical Regulation

Daniel Benamouzig

Daniel Carpenter[1]

Modern pharmaceuticals are inherently dual and uncertain in nature. Like any other technology, pharmaceuticals carry treatment value and the threat of harm – both of which are unknown – and are regulated accordingly. While they promise relief and even curing of serious diseases, they also carry risks even when they are effective. Pharmaceutical regulation is specifically grounded in this duality and uncertainty, aiming to learn about their benefits and risks and then weigh the benefit-risk balance. Modern pharmaceutical treatment co-evolved with systematic testing regimes, and the rise of the modern biotech industry is inseparable from that of the testing and regulatory regimes that have been established to govern its products. Modern pharmaceutical regulation is best understood as a system of *approval regulation*, that is, experimental minima[2] combined with potential state veto of research and development (where veto is premised upon subset of

[1] Benamouzig: Sciences Po Paris, Paris France (daniel.benamouzig@sciencespo.fr). Carpenter: Harvard University, Cambridge, Massachusetts, United States of America (dcarpenter@gov.harvard.edu). The authors conceived and wrote this essay independently of Microsoft Corporation and the Corporation had no control over the research leading to this chapter, the drafting and editing of this chapter, or the inclusion or exclusion of any idea or conclusions herein.

[2] By which we mean a set of experiments that any drug must go through, with all of the variation in experimental burden being above this floor.

experimental results).[3] The link between experimentation and state gatekeeping power is as pronounced in the pharmaceutical realm as in any other domain of economic and social activity. The fact that a new or modified molecule must receive regulatory authorization in order to be sold (and hence generate revenue) underpins a vast system of global experiment. Approval regulation as a regulatory-experimental system evolved due to scientific innovation, public and economic demand for information about products, and political pressure forged through critical historical episodes.

Introduction: Benefits and Risks in the Design of Testing

The mass development, sale and utilization of pharmaceutical treatments stems from the intersection of scientific innovation, population and economic growth and increased life expectancy. Increased life expectancy generates greater demand for treatments that will improve the quality of life and its expected duration. Nineteenth-century humans certainly cared about cancer, but it did not subtract decades from life expectancy as it did on the twentieth century and after. An extended life expectancy also means humans spend more of their lives being sick, and prolonged time in a state of illness. An empirically disproportionate amount of medical spending occurs in the last decade of most patients' lives (French et al 2017).

Regulation and testing in this domain co-evolved with a burgeoning market (Daemmrich 2017). In many cases, regulation did not follow market expansion but preceded it (Marks 1997, Carpenter 2010). Testing-based regulation has arisen for several intertwined reasons. First, the nature of health insurance and of medical risk generate strong demands for information and for the reduction of uncertainty. This is

[3] For a historical treatment, consult Carpenter (2010). For theoretical discussion and mathematical models, consult Carpenter (2004), Carpenter and Ting (2007), Carpenter, Grimmer and Lomazoff (2010), Henry and Ottaviani (2019) and Otttaviani and Wickelgren (2024).

true not only for the people who ultimately "consume" pharmaceuticals, but for those who prescribe them (physicians and other licensed professionals) and those who pay for them (insurers both private and public) (Carpenter, Grimmer and Lomazoff 2010). Second, pharmaceutical companies themselves have invested heavily in experimentation, in part because the financial return on their investment depends massively upon the efficacy and safety of their products (Daemmrich 2017). Third, political crises in the pharmaceutical industry – especially that associated with national or global public health tragedies (e.g., sulfanilamide in 1938, thalidomide in the early 1960s, Vioxx in 2004) has played a critical role in crystallizing public pressure for greater testing requirements. These economic and political forces have combined with the organizational embedment of testing requirements, so that testing is embedded in pharmaceutical companies, government agencies and research hospitals (Benamouzig and Borraz 2016).

Regimes of testing and experimentation have come to define the modern pharmaceutical industry.[4] As a general matter, each new molecule proceeds through defined sequential phases of testing such that each level of testing can mark an "up or out" juncture. Before the 1950s, the phases were different tests of acute and chronic toxicity with added tests for efficacy based primarily upon observation (Marks 1997; Carpenter 2010, Chapter 3). Beginning with cancer medications in the 1950s and then FDA regulations, the global pharmaceutical industry converged to phased testing in the 1960s and 1970s, in which Phase I trials entail generally non-controlled tests for safety and toxicity, Phase II trials involve observational and controlled tests for safety and efficacy, and Phase III trials involve randomized controlled trials in which dosage varies

[4] There is indeed an argument that testing and experimentation define modern pharmaceuticals too much, limiting the rise of potentially valuable treatments. The normative trade-offs are difficult to evaluate with precision and are beyond the scope of this chapter.

more commonly and larger-sample efficacy and safety are examined. There has been considerable modification of this basic three-phase framework, and R&D now retrofits two additional phases onto this structure: "Phase 0" studies that involve animal models and "Phase 4" studies that examine safety and efficacy after regulatory approval.[5]

First implemented in the United-States in the 1950s, testing procedures and the successive phases of research and development were gradually extended to various parts of the world (Vogel, 1998). They eventually achieved global reach in the early 2000s. The model was first extended to countries that were within the direct sphere of American influence in the 1960s, such as the Federal Republic of Germany and Japan, before being more broadly adopted by some European countries having a pharmaceutical industry, such as the United Kingdom and France in the 1970s, and finally by the entirety of Europe in the early 1990s, in the context of the creation of a single European economic market (Abraham, 2002 ; Hauray, 2006). At the same time, these procedures were also adopted in various Commonwealth countries, under the influence of the British model, with evaluations conducted in London often serving as an entry point for the entire Commonwealth. In the following decade, other countries adopted the same standards in a context of globalization. While clinical trials were increasingly implemented across the world by the pharmaceutical industry, the International Conference on Harmonization (ICH) played a key role to disseminate common standards of registration across the world (Timmermans, 2004). From 1990, the ICH convened in a formalized way representants of regulatory authorities and pharmaceutical companies to define principles and standards collectively. After a series of working group, the initial drafts produced by the ICH are open to public consultation and modified accordingly, before

[5] In recent decades important modifications have been made to this system, including single-trial designs for Phase III and arguments about whether Phase II and II should be merged. Another important modification exists in accelerated approval pathways and their intersection with surrogate endpoint measurement.

being adopted and widely disseminated, including to non-members States. China, for instance, centralized its process of drug registration and adopted the international standards of clinical testing in 1999.

The sequential model of testing has also been gradually complemented by a set of additional control and evaluation criteria, such as the quality of pharmaceutical production ("*Good manufacturing practice"*) and distribution ("*Good distribution practices"*), or drug trials guidelines. These criteria, in the same way, were first defined in the United States from the post-war decades, then adopted during the latter part of the twentieth century by most countries subjected to the same process of normative expansion. These principles now constitute a complex international set of norms that closely associates medical communities, industry players, and regulatory authorities responsible for drug evaluation worldwide. A key aspect of pharmaceutical domain is that testing has become organizationally embedded and expressed (Benamouzig and Borraz 2017; Carpenter 2010). Entire new forms of economic and health organization, including new professional societies, have arisen to assess the safety and efficacy of medical treatments (Benamouzig 2005). The organizational and procedural complexity of pharmaceutical testing is not merely a matter of government regulation. It appears as a complex set of organizations and procedures convening a vast range of players according to methods, management rules, data and committees of all sorts. They are also inseparable from the structuring of globally significant pharmaceutical markets, in which a limited number of global, capital-intensive companies concentrate most activities. This complex collective organization is one reason why testing regimes are difficult to change.

An Overview of the Testing Landscape

Pharmaceutical testing in Europe and the United States (as well as other countries) examines new molecules (or already approved molecules for new uses). Much testing is disease-specific, examining the hypothesis that the treatment is effective for a given

medical indication (the disease that the drig is intended to treat). The medical indication shapes a range of key parameters for biopharmaceutical products:

- *the measures of benefit*, since these are defined in relation to the benchmark medical condition (which may amount to observed mortality/morbidity or which may be more speculative, as in surrogate endpoints).[6]
- *the efficacy-safety tradeoffs* at issue; if the disease has high mortality, subtracts quality-adjusted life years, carries important social or emotional weight (many cancers) or has few available treatments, the benefits of a medically effective treatment are considered higher.[7]
- *the sample size* of observational and experimental studies, since some diseases affect larger populations.

Which risks and/or benefits are tested for? In general, pharmaceutical testing examines two main variables: the safety of a drug and the effectiveness of a drug. Both are understood in relationship to the drug's indication (the disease that it targets), but efficacy in particular is more heavily shaped by indication than safety is. Safety tests examine both acute and chronic safety, as the risk associated with some pharmaceuticals emerges only months or years after their continual use. Such risks may involve hepatotoxicity (the active molecule for many pharmaceuticals is in fact the

[6] The concept of endpoint was initially applied to mortality, as in the duration of life potentially extended by a cancer treatment. Surrogate endpoints are considered observable markers short of mortality that are correlated with reduced mortality (or another "hard endpoint" that we might care about). For solid tumors in oncology, tumor size reduction is often used. In stufies of diabetes medications regulators and developers use HbA1c (or "a1c"), a biomarker that indicates mean plasma glucose levels over an interval of three months.

[7] This is one reason why rare or orphan drug markets have become a primary pathway for many new molecules. Once approved for the orphan or rare disease designation, they are often tested for their potential application to other diseases.

metabolite produced by action of the human liver) or cardiovascular risk (as with COX-2 inhibitors).

Measurement. The effectiveness of a drug is measured in direct or surrogate fashion.[8] Here the measurement will depend upon the nature of the disease. In fields such as cardiology and oncology, direct measurements may involve mortality (time to death, conditioned upon diagnosis or primary metastasis) or morbidity (metastasis for cancer, myocardial infarction or stroke for heart disease). In the past three decades, a range of both surrogate endpoints and testing regimens premised upon them has emerged, such that oncologic drugs are now often approved on the basis of changes in tumor size or biomarkers (Chen, Haslam and Prasad 2020), or cardiovascular drugs are based upon changes in measured lipid levels or observed ventricular function.

Finally, important properties that shape both safety and efficacy are also subject to testing. The most important is *dosage*. Pharmaceutical products are not approved as one-size-fits-all entities, and the characteristics of dose-response mapping constitute some of the most important rationales for testing.[9] A second is administration, packaging and bioequivalence. The mere color of a pill may have important implications for the way it is metabolized, such that two pharmacologically identical compounds may not be bioequivalent. Bioequivalence – the identity of one version of the drug to another, including its metabolization in the human body – is important because if pre-market testing is to be informative as to post-market utilization, the pharmaceutical product marketed should, at least in theory, be identical to the one tested in phased trials. Bioequivalence (or for more complex proteins, biosimilarity)

[8] We move back and forth between efficacy and effectiveness but the two are subtly different in law and regulatory science.

[9] It is well known that cardiovascular morbidity and mortality associated with oral contraceptives used by women in the 1960s and 1970s stemmed from an initial dosage that was, in retrospect, too high. This fact was revealed only with considerable post-market testing.

also undergirds the modern generic drugs industry, which provides an important source of price competition for patients, prescribers and insurers.

Thresholds. The aim of most testing is to derive larger-sample estimates of measures of efficacy and risk. The phased system of clinical experiment usually relies upon increasing sample size across the phases, such that Phase III trial are larger than Phase II trials, which are larger than Phase I trials. The thresholds for efficacy and safety are defined only generally in statutory text. More precise definitions rely upon rulemaking and guidance documents issued by the FDA, which often differ across disease category (oncology, central nervous system, etc.). These disease-specific guidelines often embed some of the tradeoffs between severity and benefit that vary so heavily across disease types. They also reflect the presence of existing treatments for a drug or whether there is "unmet medical need" (Carpenter et al 2010). In general, pharmaceutical sponsors will wish to have clinical trials that show a statistically and substantively significant positive difference in an efficacy measurement between their treatment and one or more controls, where the controls can include (1) no treatment, (2) a placebo or (3) an existing and better-known treatment alternative.

In recent decades, a system that relied heavily upon regulatory discretion has come to involve greater back-and-forth dialogue between regulators, patients, scientific professionals and sponsors. For what are considered "pivotal" clinical trials in Phase III, the FDA and sponsors hold "end-of-Phase-II conferences" at which negotiations over the proper efficacy and safety criteria are discussed. In recent years, non-government scientists and patient advocates have also weighed in to the process, most visibly at advisory committee meetings but now frequently in earlier stages of molecular development.

The Organization and Timing of Testing. It is pharmaceutical companies, largely publicly traded, that conduct the vast majority of the research geared toward the refinement and eventual authorization of a new drug. The central role of experimentation and has transformed the pharmaceutical industry. When Peter Drucker told Pfizer's executive in the 1950s that "What you are making and selling is knowledge, and manufacturing is incidental," he was speaking to a transformation that was already underway. As testing standards strengthened and regulation became more widespread and intensive, entire companies reorganized their structures so that experimentation and regulatory submission – and management of relations with the regulator – defined the organization charts and strategies of the modern biopharmaceutical enterprise.

In the pharmaceutical domain, testing happens across the life-cycle from tests in animal models to randomized, controlled clinical trials to post-market studies of safety and efficacy. That said, *testing in the pharmaceutical domain is front-loaded*, meaning that most of it occurs before regulatory authorization. Once a drug is authorized (licensed) and marketed, the amount of experimentation goes down. Pharmaceutical sponsors have the greatest incentive to conduct costly experiments before the drug is authorized because it is the regulator, and not the company, who makes the final launch decision. Consistent with these incentives, a range of so-called "Phase IV commitments" by companies to conduct experiments after regulatory approval are carried out more slowly than in the agreement or are not realized at all (Moneer et al, 2022).

The front-loaded character of regulation has been the subject of considerable debate. Critics of global pharmaceutical regulation argue that the process is so expensive that the world never sees important therapies that are abandoned because the cost of testing them would be too high. Responding to some of these critiques, the FDA and EMA and other global pharmaceutical regulators have searched for ways to shorten and cheapen the R&D process. Yet the criticisms still remain. There are also arguments

that pharmaceutical testing provides public goods that would not be generated by an unregulated marketplace, and the information from these experiments is critical not only for safety but also for optimizing efficacy (dosage experiments) but also for physicians, insurers and providers.

Historical Development in Pharmaceutical Testing

In some sense, casual experimentation with putative "cures" has been around for centuries. Yet in the pharmaceutical domain, testing co-evolved with the development of regulation and industry expansion. An entire branch of medicine – pharmacology – emerged to assess the safety and efficacy of pharmaceutical treatments. The Biologics Control Act of 1902 required that establishments that manufactured vaccines, serotoxins and other treatments be subject to annual licensing. The Federal Food, Drug and Cosmetic Act of 1938 established the first system of approval regulation, requiring any new drug sold in interstate commerce to be first subject to safety tests and submitted to the FDA. After the 1938 Act in the United States, there was a broad development of pharmacological testing methods and apparatuses in the pharmaceutical industry, in academic medicine and in government agencies. Modern standards of effectiveness and of phased experiment arose as much in government agencies as they did in academic medicine and drug companies (Marks 1997; Keating and Cambrosio 2019).

Given that pharmaceutical treatments are generally applied to individual human patients and that the risk is viewed as individualized, the methods of testing have focused upon large-sample observational and prospective experimental studies with human subjects. The two main changes to testing regimes have occurred in testing regulations and testing methods.

Changes in Regulations. An important change in policy occurred in the late 1930s when the drug elixir sulfanilamide killed over 100 people in the United States. The episode, combined with rulemaking and statutory drafting at the U.S. Department

of Agriculture (then the umbrella agency housing the FDA) led to the Federal Food, Drug and Cosmetic Act of 1938. The 1938 Act created the basis of modern approval regulation by requiring drug companies to provide tests of safety before marketing. If the FDA was not convinced by these tests, it could delay or reject authorization for the product. This created a time-limited veto, in which the product was authorized if the FDA did not reject it within a specified period of time (often 60 days).

The evolution of requirements for efficacy began in the 1930s and 1940s with broad debates about what pharmaceutical safety meant. Recognizing that no pharmaceutical product had zero risk or toxicity, pharmacologists began to gesture toward an efficacy-safety tradeoff by arguing that a drug had to have some efficacy in order to be safe. As scholars have demonstrated (Marks 1997, Carpenter 2010), a functioning mode of efficacy regulation was well established in the U.S. FDA well before the thalidomide crisis. A second global tragedy – the thalidomide crisis in which thousands of children were born with severe birth defects[10] – led in the United States to congressional delegation of vast powers of approval and experimental regulation to the FDA.[11] After the 1962 Kefauver-Harris Amendments to the 1938 Act, demonstration of a drug's "effectiveness" through "well-controlled trials" was now a prerequisite to marketing authorization. The three-phase system of experiment, now globalized (Carpenter 2010, Chapters 5, 12), was created not in statute but in rulemaking in 1963. The primary innovations in this space have occurred less in the arrangement of phases – three basic phases still structure the process – but rather in what each phase contains. In recent decades, important modifications to regulatory standards have occurred

[10] The number of spontaneously or electively aborted children has never been counter but probably rests in the thousands or tens of thousands.

[11] See Carpenter (2010), however, who shows that much of the system of efficacy regulation, experimental regulation by rulemaking and guidance, and phased experiment was already in place by 1961.

within administrative government, especially for surrogate endpoints. Yet the basic system of phased experiment and efficacy requirements endures and has, moreover, been adopted globally, albeit with regional and national modifications.

Changes in Methods. From an early twentieth century state in which drugs and food additives were tested for safety (first *acute* or immediate, then *chronic* or long-term), a range of methods emerged from academic medicine and pharmacology to more rigorously assess safety and efficacy. Perhaps one of the most important developments came in the large-sample randomized clinical trial (RCT), which emerged in part from statistics and agronomy but since the 1950s has been embedded centrally in medical evaluation of treatment procedures as well as pharmaceuticals. A controlled trial contrasts the aggregated results among individual patients in a treatment arm (those receiving the drug in question) versus those in a control arm (those not receiving the drug). A key concern among those making inferences about the drug's safety and efficacy is whether the patients in the control and treatment arms differ in any meaningful respect other than the administration of the treatment. If the difference is so small as to be "ignorable," then a key condition for proper causal inference has been satisfied. It is now well-established that randomized assignment to treatment and control arms is the best way to achieve this "ignorability." An additional question is what the control arm involves. Should it involve no treatment, a placebo, or the best available standard of treatment? And if the control arm involves a treated that has been previously or elsewhere demonstrated to be safe and effective, should superiority or non-inferiority be used as the standard of comparison?

Explaining the Changes and Continuity. Three broad factors have driven the transformation of experiment. The first is *scientific innovation*, including enhanced assays for toxicity, surrogate endpoints for efficacy, and changing research designs (sequential trials, Bayesian trials, randomized controlled trials). These occurred at the intersection of academic medicine, statistics and the FDA, and in part because of

extensive industry and patient-group criticisms of the pharmaceutical regulation process. The reforms have been incremental and piecemeal rather than simultaneous and comprehensive, reflecting the slow nature of institutional and regulatory change in the United States and its system of separated powers, legalism and a decentralized exeutive branch. A second factor is public and legislative demand for regulation, a force that long led the United States and Europe to regulate pharmaceutical development more stringently and which in recent years has begun to emphasize greater patient access and lower-cost development.[12] Public demand for regulation has been driven in part by crises and public tragedies (in the United States, sulfanilamide in 1938, thalidomide in 1962, Vioxx in 2004), and in part by long-term distrust of the pharmaceutical industry, a development now shared as much or more among conservative electorates in the United States and Europe as among left-leaning voters.

A third factor is path dependence, namely the institutional stability that makes it difficult for stakeholders to change the style of regulation once adopted (Pierson, 2000). In the long run, existing testing criteria are complemented and incrementally redefined, rather than radically changed. Shifts remain rare because they are costly.

A good example of such path dependance is the difficulty of complementing existing procedures based on the clinical evaluation of drug safety and efficacy, with economic standards measuring the cost-utility of drugs. While relevant quantitative methods were invented from the 1980s, particularly in Britain, most countries were reluctant to shift from a pure clinical regime of evaluation, based on well experienced principles, to an economic paradigm incorporating "pharmaco-economics" (Hassenteufel, 2017). Methodologies and expert communities do exist worldwide, and

[12] There is good reason to believe, however, that the Trump Administration in the United States will take a less pro-innovation stance toward pharmaceuticals, perhaps less so than any Democratic (much less Republican) presidential administration since the 1970s. Early indications suggest scrutiny of user-fee programs for agency funding and for accelerated approval mechanism.

their studies are considered by health care providers. But few countries include cost-utility studies in their evaluation processes. The rule-proving exception is Britain, where these methods were invented, and to a lesser extent, a few other countries like Australia and Canada. In most countries (such as the U.S.) economic information is considered separately from the approval regulation process.

The political context of regulatory changes is essential and should be highlighted. The history of testing and regulation in many domains suggests that political context, including trust or distrust in the regulated industry, is a critical force shaping the kinds of tests that are applied, the kinds of things tested for, and the costliness of the testing system. Even as the global biopharmaceutical industry has created large-scale innovations that have delivered massive health benefits – the emergence of mRNA vaccines for COVID-19 is one of the best examples – it is also the case that the industry remains broadly distrusted, and for many different reasons. In some cases, safety crises have amplified doubt in the pharmaceutical industry. In other cases, pricing dynamics and globalized economic inequality have contributed to the distrust. It is certainly plausible that such distrust might be misplaced and that popular politics and over-reaction to crises may have led to counterproductive regimes of regulation. Yet in terms of understanding the history of pharmaceutical testing and regulation, political forces must rest at the center of understanding and analysis.

Discussion: Domain Lessons Learned and Policy Considerations

There is lack of consensus on the way forward. The massive expense required to bring a new molecule to market has left many researchers, drug companies, patients and patient advocates and politicians pushing for a relaxation of clinical trial requirements, especially for pre-market testing. Some reforms have already occurred. Yet pharmaceutical companies themselves resist the most radical reforms (such as

jettisoning the efficacy requirement). This is due less to capture (the costs that barriers impose upon potential and actual competitors) than due to the fact that testing requirements undergird an ecology expectations and confidence that supports a more profitable marketplace.

Evaluation procedures in the pharmaceutical sector often serve as benchmarks for other health-related sectors. However, they have significant particularities that must be emphasized when drawing inspiration from them for application to other fields, such as artificial intelligence.

First, the evaluation of medicines is comprehensive in the sense that all pharmaceutical products require market authorization based on testing of marketed products. This is not necessarily the case for all health-regulated products in other sectors. In the chemical sector, for example, comprehensive risk assessment focuses on the riskiest products, according to criteria that vary across national regulations. In the American context, where regulation under the Toxic Substances Control Act (TSCA) passed in 1976 includes a first "prioritization" phase, only products containing substances identified as concerning by the Environmental Protection Agency are considered and subject to testing. In Europe, under the regulation REACH adopted in 2006, only products with a certain concentration of chemicals are subject to the "registration" process, while further "evaluation" process is carried out to some of them, according to their level of risk (Boullier, 2019, 2022). In both cases, many products are otherwise exempted. Similarly, in the medical devices sector, the level of evaluation depends on a level of risk previously associated with the device, based on its characteristics, and uses. Only devices presenting the highest risks undergo a comprehensive evaluation, known as level III, which requires testing and clinical trials comparable to those demanded for medicines. The evaluation of artificial intelligence could result in adaptations inspired by existing regulations, with pharmaceuticals offering the most comprehensive forms in health-related contexts.

Secondly, ways of regulating distribute risks among different types of actors, notably industrial and administrative, who are involved to varying degrees in the organization of regulation (Gaudillère and Hess, 2013). The evaluation of medicines implements a principle of assessment of technologies by a regulatory authority, which determines whether they can be marketed. This principle is different from that of compliance with pre-defined standards or classification set by a regulator, such as the levels of risks now in place in Europe with the Artificial Intelligence Act. With dedicated public bodies in charge of the assessment, evaluation of medicines involves a shared risk between manufacturers and regulators, whose responsibilities may be engaged in case of errors, particularly when tort-based compensation must be provided to the victims of a poorly evaluated medicine. In a compliance regime, responsibility rests with the manufacturer, who is held accountable before an ordinary court and can, as such, be sued by claimants and public authorities.

Thirdly, evaluation procedures have decisive and transformative effects on market structure. The requirement for pre-market testing places the most powerful companies in a position of strength compared to smaller competitors, whose capacities to conduct clinical studies are limited. Only the largest global biopharmaceutical firms can invest in large-scale evaluation processes, which mobilize hundreds of millions or even billions of dollars and involve significant financial risks in case of rejection or limitations imposed by authorities (DiMasi and al., 2016 ; Sertkaya, 2024). Since the diffusion and globalization of drug evaluation procedures, pharmaceutical market structure has become concentrated in a small number of global companies, the only ones capable of financing the experiments and developmental protocols required by regulators. Testing is often applied on innovations produced by other actors, like university spinoff inventions whose patents are then purchased. The testing can also be conducted by externalized specialized companies, known as contract research organizations (CROs). They oversee collecting and aggregating systematically quantified data produced by multiple clinical

services, respecting quality standards and transparency requirements. In the past half century, experimental and regulatory norms have dramatically transformed the biopharmaceutical industry. The creation of evaluation procedures inspired by drug regulation in the field of artificial intelligence could have similar effects. It could place a limited number of large operators in a position of strength compared to the competition from many small companies, whose innovations might struggle to meet increasingly demanding requirements, particularly financial ones. Conversely, industrial concentration in AI might render regulatory licensing more feasible (Carpenter and Ezell 2024). It remains to be seen whether the firms then in a position of strength would be those already mastering these methods and procedures, stemming from the pharmaceutical industry, or new entrants from the digital industry.

References

Abraham, J. "The political economy of medicines regulation in Britain", in Lawton Smith, H. *The regulation of science and technology*, Basingstoke, Palgrave, 2002.

Benamouzig, Daniel. *La santé au miroir de l'économie. Une histoire de l'économie de la santé en France*. PUF, 2005.

Benamouzig, Daniel, and Julien Besançon. "Les agences, alternatives administratives ou nouvelles bureaucraties techniques ? Le cas des agences sanitaires." *Horizons stratégiques* 1 (2007): 10-24.

Benamouzig, Daniel, and Olivier Borraz. "Bureaucratie, organisations et formalisation des savoirs." *L'Année sociologique* 66, no. 1 (2016): 9-30.

Boullier, Henri *Toxiques légaux. Comment les firmes chimiques ont mis la main sur le contrôle de leurs produits*, Paris, La Découverte, 2019.

Boullier, Henri, and Henry, Emeric. 2022. "Toxic Ignorance. How Regulatory Procedures and Industrial Knowledge Jeopardise the Risk Assessment of Chemicals," *Science as Culture*, 31(4) 480–503.

Carpenter, Daniel P. 2004. "Protection without capture: Product approval by a politically responsive, learning regulator," *American Political Science Review* 98 (4) (2004): 613-631.

Carpenter, Daniel. 2010. *Reputation and Power: Organizational Image and Pharmaceutical Regulation at the FDA* (Princeton: Princeton University Press).

Carpenter, Daniel, Justin Grimmer, and Eric Lomazoff. 2010. "Approval regulation and endogenous consumer confidence: Theory and analogies to licensing, safety, and financial regulation," *Regulation & Governance* 4 (4) (2010): 383-407.

Carpenter, Daniel, Susan I. Moffitt, Colin D. Moore, Ryan T. Rynbrandt, Michael M. Ting, Ian Yohai, and Evan James Zucker. 2010. "Early Entrant Protection in Approval Regulation: Theory and Evidence from FDA Drug Review." *The Journal of Law, Economics, & Organization* 26, no. 3 (2010): 515-545.

Carpenter, Daniel, and Michael M. Ting. 2007. "Regulatory errors with endogenous agendas." *American Journal of Political Science* 51, no. 4 (2007): 835-852.

Carpenter, Daniel and Carson Ezell. 2024. "An FDA for AI? Pitfalls and Plausibility of Approval Regulation for Frontier Artificial Intelligence," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (Vol. 7, pp. 239-254).

Chen, E.Y., Haslam, A. and Prasad, V., 2020. FDA acceptance of surrogate end points for cancer drug approval: 1992-2019. *JAMA internal medicine*, *180*(6), pp.912-914.

Daemmrich, Arthur. 2017. "Pharmaceutical Manufacturing in America: A Brief History," *Pharmacy in History*, 59 (3), 59-72.

DiMasi, Joseph, Grabowski, Henry, Hansen, Ronald. 2016 "Innovation in the pharmaceutical industry. New estimates of the R&D costs", *Journal of Health Economics,* 47(5) 20-33.

Emory Healthcare. 2024. "New Prescription Weight Loss Drugs: What's All the Hype?" https://www.emoryhealthcare.org/stories/wellness/new-prescription-weight-loss-drugs-whats-all-the-hype (accessed January 16, 2025).

French, Eric B., Jeremy McCauley, Maria Aragon, Pieter Bakx, Martin Chalkley, Stacey H. Chen, Bent J. Christensen et al. 2017. "End-of-life medical spending in last

twelve months of life is lower than previously reported," *Health Affairs* 36 (7) (2017): 1211-1217.

Gaudillère, Jean-Paul, Hess, Volker. 2013. *Ways of regulating drugs in the 19th and 20th Centuries*, Basingstoke, Palgrave, 2013.

Hassenteufel, Patrick, Daniel Benamouzig, Jérôme Minonzio, and Magali Robelet. "Policy diffusion and translation: the case of evidence-based health agencies in Europe." *Novos estudos CEBRAP* 36, no. 1 (2017): 77-96.

Hauray, Boris. 2006. *L'Europe du médicament*, Paris, Éditions de l'EHESS, 2006.

Henry, Emeric, and Marco Ottaviani. 2019. "Research and the approval process: The organization of persuasion," *American Economic Review* 109, no. 3 (2019): 911-955.

Marks, Harry M. *The progress of experiment: science and therapeutic reform in the United States, 1900-1990*. Cambridge University Press, 1997.

Moneer, O., Brown, B.L., Avorn, J., Darrow, J.J., Mitra-Majumdar, M., Joyce, K.W., Ross, M., Pham, C. and Kesselheim, A.S., 2022. New drug postmarketing requirements and commitments in the US: a systematic review of the evidence. *Drug Safety*, *45*(4), pp.305-318.

Ottaviani, Marco, and Abraham L. Wickelgren. 2024. "Approval regulation and learning, with application to timing of merger control," *The Journal of Law, Economics, and Organization* 40, no. 3 (2024): 597-624.

Pierson, Paul, 2000 "Increasing returns, path dependance and the study of politics," *American Political Science Review*, 94 (2) 251-267.

Rubino DM, Greenway FL, Khalid U, O'Neil PM, Rosenstock J, Sørrig R, Wadden TA, Wizert A, Garvey WT; STEP 8 Investigators. 2022. « Effect of Weekly Subcutaneous Semaglutide vs Daily Liraglutide on Body Weight in Adults With Overweight or Obesity Without Diabetes: The STEP 8 Randomized Clinical Trial, » *Journal of the American Medical Association*. 2022 Jan 11;327(2):138-150. doi: 10.1001/jama.2021.23619. PMID: 35015037; PMCID: PMC8753508.

Sertkaya, Aylin, Beleche, Trinidad, Jesup, Amber and Sommers, Benjamin 2024. "Costs of Drug Development and Research and Development Intensity in the US, 2000-

2018", JAMA Network Open. 7(6) e2415445. doi:10.1001/jamanetworkopen.2024.15445

Timmermans, Karin, 2004 "Harmonization, regulation, and trade in the pharmaceutical field", *International Journal of Health Services*, 34 (3) 651-661.

Vogel, Jacob 1998 "The globalization of pharmaceutical regulation" *Governance*, 11 (1) 1-135.