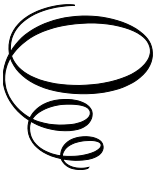
Interdisciplinary Perspectives on the Issues of Proof in Health Science

Interdisciplinary Perspectives on the Issues of Proof in Health Science

Edited by

Léo Coutellec and Amélie Petit

Cambridge Scholars Publishing



Interdisciplinary Perspectives on the Issues of Proof in Health Science

Edited by Léo Coutellec and Amélie Petit

This book first published 2023

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-5275-0707-6
ISBN (13): 978-1-5275-0707-4

TABLE OF CONTENTS

PUTTING WHAT SERVES AS EVIDENCE TO THE TEST: AN INTRODUCTION

LÉO COUTELLEC & AMÉLIE PETIT

In medicine, the question of evidence is a debate as intense as it is continuous, raising numerous ethical and epistemological issues. Patient cure is a deceptive element in the assertion of treatment efficacy. At the time of the founding work by James Lind, Claude Bernard, Louis Pasteur and Pierre-Charles Bongrand, clinical experimentation was conceived as the best way of validating the therapeutic effect of a treatment. In the course of a long historical process, *double-blind randomised clinical trials* became tools for public action in the service of a form of medicine aiming to be more rational (Marks 1999). Clinical trials are based on a particular *style of reasoning* (Hacking 1992), calling on inferential statistics and on the idea according to which a "suffering body is not a reliable witness" of the therapeutic effect attributed to a treatment. The "body is liable to be cured for bad reasons" and become an "accomplice to charlatans" (Stengers 2013, 123-124). Randomisation procedures and blinded treatment protocols were thus conceived as the best way of providing evidence, by excluding from the experimental field aspects liable to bias a causality relationship between a substance and its therapeutic effects[1]. According to this style of reasoning, charlatans are defined as "those who claim their cures as evidence" (Stengers 2013, 121), without having tested them in the setting of a controlled experiment. This is how one of the specificities of modern bio-medical research, particularly with the promotion of randomised clinical

[1] Comparing the effect of an experimental treatment with the effect of a reference treatment (or a placebo) makes it possible to control for *confounding bias*, which can make it impossible to identify the cause of a given effect. Randomisation consists in randomly distributing a treatment between an experimental group and a control group. This randomisation enables an objective allocation of treatments, preserving the process from *selection bias*. Finally, the process of blinding makes it possible to maintain the epistemic effects of randomisation and favours equal treatment in follow-up. Ultimately, the comparison of the effects of the two treatments makes it possible to demonstrate the existence of a difference in efficacy that is *statistically significant*.

trials, was the recognition of a form of presumption of reliability in certain regimes of evidence.

Proof and its administration

This epistemology of experimental evidence first emerged in the 19th century in the setting of the struggle against practices that physicians considered as stemming from para-science, such as mesmerism (Chamayou 2008). Then, shortly after the Second World War, "therapeutic reformers"[2] (Marks 1999) managed, within clinical practice, to import procedures of randomisation originally developed to assess the fertility of seeds and agricultural soils. The objectivity attributed to randomised experiments enabled reformers to see clinical trials as providing "impersonal criteria of scientific integrity" (Marks 1999 p. 18), thus protecting medicine from the commercial influences of a booming pharmaceutical industry (Cheveau 1999).

From the sixties, more than a style of reasoning, clinical trials progressively became a *style governance based on evidence* (Desrosières 2000), enabling the regulation of drug marketing. In the United States and in Europe, under the auspices of the *Food and Drug Administration* (Carpenter 2010) and the European Medicine Agency (Hauray 2006- respectively, the implementation of clinical trials is now an inescapable administrative requirement in the marketing of a treatment. In the nineties, this entanglement of epistemic and regulatory principles gave birth to the paradigm of *Evidence-Based Medicine* (EBM), where one of the restricting aspects of its conception was "to restrict decision-making to what is statistically tested" (Da Silva 2017, 268). Within this movement of normalisation of healthcare practices (Timmermans and Berg 2003), the EBM promoters placed the data from clinical trials at the top of the evidence hierarchy, considering randomised experiments as "a method that claims to "finish with" evidence, in the sense that it would not be possible to produce evidence of better quality" (Jatteau 2020, 28).

In medicine, clinical trials thus constitute a two-fold administration of proof, at once in the sense of an epistemic demonstration of treatment efficacy and in the sense of the administrative management of that same treatment. The production of evidence is thus part of a fully procedural

[2] The phrase "therapeutic reformers" covers a wide range of professionals, including pharmacologists, physiologists, clinicians, statisticians, epidemiologists, political figures and journal editors, convinced by the power of science "to unite researchers and practitioners in medicine *despite* obvious differences in training and practice". (Marks 1999, p.17-

organisation, the complexity of which can sometimes result in neglect of the original purpose of evidence-based medicine (Worall 2017).

The hegemony of this conception of the administration of evidence raises a certain number of issues that are discussed in this book. There is a double objective. First of all, it invites reflection on the motives explaining why, on the basis of data from clinical trials and via the general EBM approach, what stands as evidence for some is not always sufficient to convince others. The stakes here are the value as proof of our attempts to obtain the adherence of a community to a result or a hypothesis. More specifically, there is a need to detail the epistemic and ethical issues relating to regulatory, methodological and statistical evidence following which the results from clinical experiments can be received as evidence and translated into medical recommendations.

Once again, we can see that science, and particularly clinical research, are social and collective activities, plural and conflicting, within which the production of evidence is embedded. In science, as in other fields such as law (Chappe *et al.* 2022), evidence must comply with a set of conventions in order to be accepted. A result therefore cannot claim the status of evidence until it has been collectively experienced as such, within an agreement of the conditions for producing that evidence. The issue of evidence thus partly depends on its communicability: an individual judgement does not really count if it is not supported by a third party, a peer, a journal, an academy, or a health agency, each protagonist referring to assessment criteria that everybody would wish to be common to all. As the sociologist Pierre Bourdieu claimed, "knowledge is based, not on the subjective evidence of an isolated individual, but on collective experience, *regulated by standards of communication and argumentation*" (Bourdieu 2001, 143), what Helen Longino called "social objectivity" (Longino 2002). The sociology of the production of evidence has provided considerable information on the social aspects of the acquisition of scientific credit (Shapin 2014).

Evidence and scientific pluralism

The administration of proof, whose role in Evidence-Based Medicine has taken the form of a hierarchy in the value of proof according to the presumed quality of each level of proof, is however problematic in the absence of "a theory of proof" (Cartwright 2007; Cartwright & Stegenga 2011), following the example of what exists in the legal field (Vergé *et al.* 2015; Leclerc 2013). This raises the question of the existence of criteria - explicit or

implicit – presiding over the selection and hierarchisation of the different pieces of evidence and the different registers of proof in medicine.

Evidence in law is deployed or expressed in a particular context, a trial, and under the constraint of explicit rules that are found in the legislation, i.e. the rules for receiving evidence. These rules manage the administration of proof and in particular enable the principle of freedom of evidence, essential in the legal provisions for proof, according to which any piece of evidence deemed relevant can be presented to the receivers of evidence (a judge or a jury). This set of rules, together with the principle of freedom of proof, makes any *a priori* hierarchy in the value of evidence obsolete. Conclusions during a trial, prefer a management of evidence according to its robustness and relevance rather than a mechanical application of a supposed hierarchy of levels of evidence. This is why a judge can, in theory, grant the narrative of a witness a value of proof as great as that of a biological analysis. The biological analysis can present methodological biases, may have been manipulated, or provide no relevant information for the trial. A witness account, even though it is fragile by nature, can on the contrary provide determining elements. This principle of the "freedom of proof" does not mean that all evidence is of value, nor that "anything can serve as evidence", but that each piece of evidence gathered in a file theoretically has the right to the same epistemic consideration. The issue then is to *compose in situation*. However, it should be underlined that there is in this context persistently strong tension between legal evidence – or evidence administered by a set of rules – and the freedom of proof. This tension can be found in the scientific field, and in particular in medicine, but with the progressively acquired privilege for the processes administered upstream from the production of proof, as illustrated in the hierarchy of evidence of EBM.

The argument usually put forward to justify this *a priori* hierarchy of evidence in medicine focuses on the idea that research methods classified higher up in the hierarchy are less biased than those further down. Yet certain authors defend the idea that there is no epistemic justification for the hierarchical classification of research methods provided for in EBM (Cartwright 2007), and others have demonstrated that, in certain cases, there is no fundamental difference between evidence generated by RCTs and that generated by observational studies (Concato et. al 2000). This has produced new procedural proposals (as in the GRADE system), and epistemological proposals concerning the notion of corroboration of the values of evidence. This corroboration of multiple evidence - "this patchwork of evidential approaches" (Mebius 2014) in clinical reasoning opens the way to pluralist thought in the production of scientific evidence.

Insofar as standards of evidence seek to reduce data production to a single way of questioning reality, this pluralism is still very little recognised as an intrinsic characteristic of the normal functioning of science. This misapprehension of the pluralist character of science can lead to substantial controversies and not to plurality as such, which on the contrary tends to value the complementarity of knowledge.

Evidence and scientific conflict

Science is nevertheless an activity involving conflict which confronts numerous regimes of evidence. Bourdieu showed very well that science is a field of forces inside which individuals are in competition to impose a legitimate acceptation what *stands* as evidence and what *is* evidence (Bourdieu 1976, 1997). Evidence is thus an element in the struggle to possess the monopole of scientific competence. This led Leng to say: "in reality, the mission of all scientists is to convince other scientists of the importance of their own ideas, and they do it by combining reason and rhetoric. They often seek evidence that supports their ideas, and not evidence that might contradict them; they often present evidence in such a way that it seems to give them support; and they often ignore evidence that is bothersome". (Leng 2020).

Despite the fact that the protagonists are trapped in an antagonistic dynamic, not everything in science is debatable. At the crossroads of different regimes of proof, there are always axiological and epistemic principles that escape these struggles, such as the submission of data to critical examination by peers. Adherence to rules for discussion is necessary for the survival of the field. Transgressing the rules would amount to deserting the field or altering it in depth if this transgression is not considered as such. Controversies and disagreements animating the scientific field thus play out in compliance with a series of rules that Bourdieu named in French "illusio", an "immediate adherence to the necessity of the field", a "non-discussed condition of discussion":

> Illusio is not of the order of explicit principles, theses that we propose and defend, it is about action, routine, things we do, and that we do because they are what is done, and we have always done so. All those who are engaged in the field, whether advocating orthodoxy or heterodoxy, have in common a tacit adherence to the same doxa, which makes their competition possible and sets its limits: it in effect forbids the questioning of principles of belief which could threaten the very existence of the field (Bourdieu 1997, 123).

Discussion, intra and inter-disciplinary points of view and any opposition arises, are the constitutive characteristics of scientific practice, and by deduction, of clinical research. The image of a unified and consensual science, a "normal science", is far from suited to reflecting the work on proof, whether in terms of production or in validation and data dissemination. Sciences are collective, plural and divergent. This book allows two pitfalls to be avoided, namely the tendency to unify sciences and the tendency to delocalise proof. By opening the "black box" of the administration of proof in medicine, it offers an inter-disciplinary reflexion on the co-construction of the care ethic and the epistemology of evidence in a context of clinical research and therapeutic urgency.

Presentation of the chapters

The book is organised into eight chapters. In the first chapter, Eli Azria, professor of gynaecology-obstetrics at Paris University, examines in detail the pedagogical and normative logics that have progressively constituted EBM. He raises questions on the evolution of the clinicians' role in their ability to mobilise relevant knowledge for healthcare. In the second chapter, Robin Michalon, a science historian, shows through a historical study of the epistemological composition of therapeutic trials that they respond to two logics, one exploratory and the other regulatory. By doing so, he re-embeds the issue of regimes of proof in a historiographic perspective, which helps to understand that the value of proof is negotiated at the meeting point of several experimental regimes. In the third chapter, Bruno Falissard, a child psychiatrist, Professor of biostatistics and Director of the Centre for epidemiology and population health, focuses on the style of reasoning that is characteristic of statistical inference, detailing the tensions underpinning the conduct of these statistical tests and the reading of their results. These first three chapters give social-historical depth to the notion of evidence and stress its incoherences, paradoxes and contradictions. In the fourth chapter, Tim Daly examines the semantic and clinical difficulties faced by biomedical research in finding a treatment against Alzheimer's disease, and raises questions on researchers' (in)ability to escape therapeutic deadlock in research. In the fifth chapter, Clément Tarantini, an anthropologist and postdoctoral researcher at the Institut Pasteur, focuses on the way healthcare providers had to deal with the uncertainty relating to the efficacy of treatments during the Covid-19 pandemic. The author strives to give an account of the different "investigation regimes" that were deployed to orient therapeutic care. Finally, in the sixth and last chapter, Vincent Israël-Jost, a science philosopher and postdoctoral researcher at CESP, examines the

controversy over hydroxychloroquine and seeks to understand the general public's sensitivity to Pr. Didier Raoult's argumentative strategy, whilst at the same time criticising his failure to comply with a certain number of norms that make science a "democratic" space.

References

Bourdieu, Pierre. 2001. *Science de la science et réflexivité*. Paris: Raisons d'agir.

Bourdieu, Pierre. 1997. *Les méditations pascaliennes.* Paris: Seuil.

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

Chamayou, Grégoire. 2008. *Les corps vils. Expérimenter sur les êtres humains aux XVIIIe et XIXe siècles.* Paris: La Découverte.

Cheveau, Sophie. 1999. *L'invention pharmaceutique, la pharmacie française entre l'État et la société au XXème siècle.* Paris: Institut d'édition Sanofi-Synthélabo.

Cancato, John, Shah, Nimish, and Horwitz, Ralph. 2000. « Randomized controlled trials, observational studies, and the hierarchy of research designs » *New England Journal of Medicine*, 342 :1887–1892.

Cartwright, Nancy. 2007. « Are RCTs the Gold Standard? » *Biosocieties* 2(1):11-20

Chappe, Vincent-Arnaud, Juston Morival Romain, Leclerc Olivier. 2022. « Faire preuve : pour une analyse pragmatique de l'activité probatoire. Présentation du dossier » *Droit et société* 110(1) : 7-20.

Da Silva, Nicolas. 2017. « Quantifier la qualité des soins. Une critique de la rationalisation de la médecine libérale française. » *Revue Française de Socio-Économie* 19(2):111-130

Desrosières, Alain. 1993. *La politique des grands nombres. Histoire de la raison statistique.* Paris: La Découverte.

Hacking, Ian. 1992. « Statistical language, statistical truth and statistical reason : the self-authentification of a style of scientific reasoning. » In *The Social Dimensions of Science*, edited by McMullen E, 130-57. Notre Dame, University of Notre Dame Press.

Hauray, Boris. 2006. *L'Europe du médicament: Politique – Expertise – Intérêts privés*. Paris: Presses de Sciences Po.

Jatteau, Arthur. 2020. *Faire preuve par le chiffre ? Le cas des expérimentations aléatoires en économie.* Paris: IGPDE.

Leng, Gareth, and Leng, Rhodri Ivor. 2020. The Matter of Facts: Skepticism, Persuasion, and Evidence in Science. Massachusetts : MIT Press.

Longino, Helen. 2002. *The Fate of Knowledge.* Princeton : Princeton University Press.

Marks, Harry. 1999. *La médecine des preuves. Histoire et anthropologie des essais cliniques (1900-1990)*. Paris: Synthélabo. Les empêcheurs de penser en rond.

Mebius, Alexander. 2014. « Corroborating evidence-based medicine » J Eval Clin Pract 20(6):915-920.

Timmermans, Stephen, and Marc Berg. 2003. « The Practice of Medical Technology. » *Sociology of Health & Illness* 25: 97-114.

Shapin, Steven. 2014. *Une histoire sociale de la vérité. Science et mondanité dans l'Angleterre du XVIIe siècle*. Paris: La Découverte.

Stengers, Isabelle, and Tobie Nathan. 2012. *Médecins et sorciers.* Paris: La Découverte.

Worrall, John. 2007. « Evidence in Medicine and Evidence-Based Medicine ». Phil Comp 2(6):981-1022.

CHAPTER 1

FROM CLINICAL EVIDENCE TO THE CLINIC OF EVIDENCE, OR THE EVOLVING LOGIC OF EVIDENCE-BASED MEDICINE

ELIE AZRIA

From treatises to scientific journals, an evolution in our relationship with knowledge

For many years, manuals or treatises were sources to which physicians referred. These works, compiled in one or several volumes, provided an overview of knowledge in a speciality, or gathered available knowledge on a disease. These works, which were often bulky, collated knowledge of the time on a particular question, or even a particular discipline. The authors were often powerful hospital professors (known as "*mandarins*" in France) who not only synthesised collective knowledge, but also shed light on the subject from their own observations and experience. If publications of this type of work are becoming scarcer and if some specialist publishers have discontinued these activities or have shifted towards the publication of periodicals or on-line publishing, it is because the evolution of knowledge has become too rapid to enable these imposing documents to survive obsolescence. While this acceleration is still moderate in certain medical fields, in the sectors relating to pharmacological treatments, which are the privileged objects of assessment via clinical trials, the acceleration is far more marked and incompatible with a treatise format. The same goes for fields in which molecular engineering is prominent, such as genetics, biotherapies and imaging. In these fields where knowledge is rapidly evolving, this traditional mode of dissemination is no longer compatible with the pace of scientific production. Periodicals, and on-line periodicals in particular, appear therefore as better-adapted sources of knowledge than treatises. This epistemological evolution involving the progressive disappearance of treatises, which played an intermediary role between

scientific production and practitioners, implies that it is now the practitioners' role to search and select, something that they did not have to do before. This means that they need to be able to navigate in a corpus of knowledge that is extremely large and qualitatively heterogeneous.

Not all the knowledge published in scientific periodicals has the same validity, and publication, even with peer reviewing, does not constitute the guarantee that could be hoped for in terms of quality. Post-publication assessment is thus more than ever necessary. It is therefore fundamental that practitioners should be able to make a critical appraisal of these publications to back up their medical acts with conclusions drawn from reliable research. It is precisely in the light of this observation, and because clinical judgment and the level of up-to-date knowledge is not equally shared among physicians, that Evidence-Based Medicine (EBM) was born. The term is relatively recent and was introduced in the 1990s by a research group in clinical epidemiology from McMaster University, Ontario, which styled itself the EMB Working Group. The term, and the method, rapidly spread after the publication of an article signed by this grouped entitled "Evidence-Based Medicine. A New Approach to Teaching the Practice of Medicine" in the JAMA (1992), which was widely heralded as describing the most suitable way to practice medicine (Evidence-Based Medicine Working Group 1992). The use of the term in scientific publications substantially increased in the following ten years (Claridge and Fabian 2005)

Initially presented as a training programe intended for practitioners, EBM rapidly established itself as a decision-making method and as a method for medical practice. This medicine is based, through a methodology of quantification of judgements, on finding the highest level of evidence. This meant that clinical decisions could be provided with a scientific base (Evidence-Based Medicine Working Group 1992). EBM deploys in a dual approach: an educational approach and an evaluative approach (Fagot-Largeault 2003). This approach has progressively gained a foothold as the best way to practice medicine, becoming the dominating mode in the assessment of the validity of knowledge (Schweitzer and Puig-Verges 2005). In this sense, EBM has become a methodological norm and a meta-methodology producing medical norms.

EBM, a pedagogical approach

Given the observation that scientific production had changed its pace, it was important to break from the habit of considering that medical knowledge was acquired once and for all at the faculty of medicine, and

was thus applicable by practitioners throughout their careers. As the content of knowledge was perpetually evolving, the aim of EBM was to make the practitioner watchful, curious and critical towards this knowledge, constantly updating it (Evidence-Based Medicine Working Group 1992). On this point, the aim is to teach a technique of critical perusal and apply formalised rules to assess the quality of the data provided in the scientific literature (Guyatt *et al.* 2000). In other words, EBM provides a tool to find ways through this literature of very unequal quality. The meta-methodology consisting in classifying study results according to the levels of evidence, which depend on the experimental methodology, and in attributing a score for relevance is part of the original EBM arsenal. The second founding postulate is the inadequacy of individual clinical experience, which is why clinicians are encouraged to resort to reviews of the published literature to search for the best studies dedicated to the question at hand (Paolaggi and Coste 2001). EBM thus proposes to replace the oriented syntheses of treatises by a systematic and "objective" analysis of the literature. This synthesis can then become the basis for medical practice and thus a critical approach would be opposed to dogmatism, doubtful of certainty and prioritise the search for information to gain expert opinion (Durieux 1998). The main principle here is that the search for truth is more relevant if all the evidence is assessed, and not just selected evidence that could favour a particular affirmation. Iain Chalmers, one of the collaborators of the Cochrane Collaboration, very actively campaigned for the recognition of the cumulative nature of scientific activity, demonstrating the lethal, morbid or wasteful consequences of the absence of systematic reviews of evidence in certain domains (I. Chalmers 2007; Iain Chalmers 1993; Djulbegovic and Guyatt 2017).

Furthermore, by developing readers' critical sense in their post-publication assessments and by raising their standards, it was believed that this could in return ascertain the quality of the scientific approach. Indeed, it can be observed that the publishers' requirements were raised, anticipating the critical abilities developed by readers, in particular with the appearance of checklists addressed to authors as a formalised version of the specifications for their submissions for publication. For each methodology used, a specific checklist was instated. Thus, concerning randomised trials, the study and the manuscript have to comply with the CONSORT checklist (Consolidated Standards of Reporting Trials) (Moher, Schulz and Altman 2001), i.e. the STARD (Standards for Reporting Studies in Epidemiology) (Bossuyt *et al.* 2003) and the STROBE (Strengthening the Reporting of Observational studies in Epidemiology) (Vandenbroucke *et al.* 2007) checklists respectively, for assessments of diagnostic tests and

observational studies. For meta-analyses and systematic reviews, formalisations of the PRISMA type (Preferred Reporting Items for Systematic Reviews and Meta-analyses) (Liberati *et al.* 2009) are now imposed on authors. The EQUATOR network (Enhancing the Quality and Transparency of Health Research) gathers initiatives aiming to promote the quality of reporting of scientific studies in the field of health. Thus EBM, aiming to transform practitioners' relationship with knowledge so as to rationalise medical care practice, has produced norms that interact very directly with the modes of production of knowledge.

EBM, an evaluative and normative meta-methodology

Research has diversified considerably, which has led to the exploration of more and more specialised domains entailing increasing technicality. As a consequence, some of its branches have become inaccessible to practitioners. Furthermore, the system of publimetry-based evaluation has contributed to the massive inflation of scientific publications witnessed today. The knowledge databases are plethoric, and paradoxically, a large number of publications only present a minor interest, or even a complete lack of interest, on account of ill-suited methods or patent biases (Ioannidis 2005). Bibliometric assessments of research activity are no strangers to this evolution. There is a need to sort studies of interest from others, as finding a way through this very heterogeneous literature is a genuine issue for healthcare. The assessment of publications is therefore central in the practice of evidence-base medicine. Besides the time at the practitioners' disposal, their limited skills in terms of research methodology restrict their ability to synthesise knowledge on a particular point. This is where, over and above the classifications enabling studies to be graded, two types of tools come into play: powerful search engines with sophisticated sorting systems, such as the interface *PubMed*, provided by the *National Centre for Biotechnology Information* (NCBI), managed by the *National Institute of Health* and the *US National Library of Medicine*, and the syntheses carried out by experts. These syntheses come in different forms, but all have in common expert encounters for their elaboration and the gradation of evidence levels according to the original method proposed by the McMaster clinical epidemiologists or to derived methods. Jeanne Daly (Daly 2005) reported a story that has been told over and over again about David Sackett: how he managed to solve a conflict among experts with the help of evidence. It was during a consensus conference where it was difficult to reach an agreement, as authoritative experts considered their positions as final. As the McMaster University team present at the

conference were unable to get them to see sense, they called on David Sackett. He suggested that the experts should be encouraged to make all the recommendations they wished, but that they should also establish a scoring scale to rate the quality of these recommendations. If a recommendation was based on evidence from a randomized clinical trial with sufficient power, it would thus become top of the list. If a recommendation was based on a case report, it would be accepted but listed as lower on the scale. The hierarchy of levels of evidence was thus born[1]. The format, the fields covered and the conditions of elaboration of these studies differ and evolve, but the objective is to provide clinicians with a frame of reference for their medical practice in the form of recommendations. This agenda thus plays the role of a synthetic link between knowledge and healthcare practice. In order to minimise risks of bias, the elaboration of Recommendations for Clinical Practice (RCP) advocates the deployment of strict and extremely formalised methods. This work, as Anne Fagot-Largeault says, symbolises the collective effort of a profession to synthesize research results, and taking these results into consideration, to discard out-dated practices (Fagot-Largeault 2003). Since its formalisation, EBM, through the meta-methodology presented, has contributed to a multiplication, or even an explosion in the number of recommendations to clinicians.

Medical decisions at the crossroads of scientific data and patient preferences

EBM emerged to meet a need to rationalise healthcare, recognising clinicians' difficulties in finding their way in medical knowledge. Beyond this rationalisation, EBM was also designed to organise the relationship between the singularity of a patient and the intrinsically general character of scientific knowledge. "A rigorous, conscientious and judicious way of using the most recent and highest-level evidence for decisions concerning an individual's healthcare", as the members of the EBM Working Group wrote in 1992 (Evidence-Base Medicine Working Group 1992). Individuals in their singularity have their rightful place in the model, and in addition to their medical singularity, there are also their preferences, references and values. In the EBM model, breaking away from a model that was very prevalent throughout the 20th century, decisions are shared

[1] Daly J, Evidence-based Medicine and the search for a science of clinical care, University of California Press, 2005, p. 76-77

and are based on the following tripod: scientific data, practitioners' experience and patients' preferences and values.

The three main epistemological principles of EBM

1. Evidence is unequal in value, and medical practice needs to be based on the best possible evidence – Hierarchy in levels of evidence.
2. The search for truth is more relevant when the evidence as a whole is assessed, and not by selecting evidence that favours a particular stance –Need for systematic reviews and syntheses
3. Clinical decision-making requires the patients' values and preferences to be taken into consideration (evidence never determines decisions; it is always evidence in the context of particular values and preferences).

In adequacies and limits of EBM

EBM thus proposes a meta-methodology that enables knowledge to be assessed, screened and finally synthesized according to an established hierarchy. This can appear as the way to resolve the problems linked to the application of knowledge to practice, and even more so when we read the manifestos of this school which aim to base healthcare on the best possible knowledge, on individual clinical experience and on patients' preferences (Sackett *et al.* 1996). However, the reality of practices is there to remind us that this principle is something of an ideal which is difficult to approach, and that EBM in its original structure presents substantial failings, as noted by a great number of criticisms (Djulbegovic and Guyatt 2017).

It is important first of all to note that EBM *per se* is not a new paradigm, as has sometimes been claimed. While the term itself is relatively recent, attempts to rationalise medicine have succeeded one another since Pierre Charles Alexandre Louis, an obvious filiation exists between this physician's numerical methods and the principles emanating from McMaster. The development of IT, both in data processing and in communication networks, contributed to the development of this approach. What is new in the EBM approach is the formalisation of its meta-methodology, which is supposed to constitute a tool to be used by all physicians. If there is a change in paradigm, it is not much that it seeks to rationalise healthcare on the basis of knowledge, but rather that EBM

seeks to standardise healthcare, with the transfer of authority held by the physician towards statistical, or even procedural criteria.

The widespread criticism of EBM led the model to evolve. The EBM system is no longer piloted by an Evidence-Based Medicine working group; as it became generalised, it escaped from its creators, and EBM became the generic name of a system of rationalisation and standardisation of medical practice. This system now evolves with the different changes brought about by a globalised community of researchers, clinicians and national and international organisations. In what follows, some of the criticisms addressed to the initial EBM model will be discussed, as well as the evolutions they triggered in the trajectory of the system.

An a priori qualification of the level of evidence

In its first formalisation, EBM was based on an *a priori* qualification of the level of evidence based on the methodological design of the study, placing randomized controlled trials (RCTs) at the top of the hierarchy of levels of evidence. Very rapidly, voices were raised against this notion, outlining quite rightly that the implementation in the field of a study with an ideal methodological design could lead to the emergence of bias, and that RCTs were not exceptions to this and therefore could not be considered as automatically providing a high level of evidence (Worral 2002). This simplistic approach to the levels of evidence very rapidly evolved and, moving from a hierarchy initially exclusively centred on the methodological design of the studies, more sophisticated systems were developed, in particular the GRADE system (Grades of Recommendation Assessment, Development and Evaluation), which has become the most accomplished elaboration (D. Atkins *et al.* 2004). This system has been recognised and implemented by many organisations, including the Cochrane Collaboration, the National Institute for Health and Care Excellence, the World Health Organisation and many other academic societies. GRADE provides a hierarchy of evidence that goes beyong methodological design, and deals with other aspects that are liable to alter, negatively or positively, the credibility of the evidence: risk of bias, precision, coherence compared to other studies, applicability, potential publication bias, effect size, the presence of dose-response gradients or possible confounding factors.

Non-integration of physio-pathological knowledge

If physiology, as Claude Bernard would have wished, cannot reasonably constitute the sole basis for medical practice, to preclude it for this motive would not be reasonable either. It is nonetheless what happened with EBM in its initial formalisation, which placed physio-pathological studies at the bottom of the evidence hierarchy. If the physio-pathological argument can be contended, the conclusions should be entirely determined by numerical results from statistical analyses (Claridge and Fabian 2005). In a situation of non-congruence between the physio-pathological model and the conclusions, it is generally this model that is invalidated. However, as has been previously demonstrated in different fields, physio-pathological knowledge enriches knowledge. When examining the modes of mechanical ventilation for patients with Acute Respiratory Distress Syndrome, Deyfuss *et al.* showed that progress achieved in this domain owed more to better knowledge of the lesions induced by artificial ventilation and to data from cohort studies than to randomised trials (Dreyfuss 2005). Examples where physio-pathological knowledge proved to be determinant are numerous. For Steven Goodman, these phenomena of non-consideration of non-epidemiological knowledge are natural consequences of the statistical methods that have been promoted and implemented by EBM (Goodman 1999). These methods, according to him, have completely undermined our ability to distinguish a statistical result from a scientific conclusion.

However, the evolution towards the GRADE method has tempered this trend, enabling the level of evidence to be modulated according to the methodological design of the study on the basis of external factors. Congruence with physio-pathological knowledge is one of these factors, even if physio-pathological studies remain in the category “very low-grade evidence” (D. Atkins et al. 2004).

Duplicity in the frequentist statistical approach, or the illusion of an experimental science without a theory

While for some, the very use of the term “evidence” is problematic, particularly in its French translation, Jean-François Foncin showed that hierarchy in different levels of evidence was not without its problems either (Foncin 2007):

> “A proposition for which we provide evidence has a unit probability (it is true); a proposition for which we provide the opposite evidence has a null

> probability (it is false). Between the two, we cannot talk of levels of evidence but of probability, between 0 and 1, for the proposition to be true: this is the general case with experimental sciences, in which scientific medicine aspires to take part."

Thus, the very idea of a "level of evidence" is a misuse of language from a mathematical viewpoint. We could take the criticism further by mentioning the rhetoric that instrumentalises evidence in an attempt to do away with the scientific uncertainty, that is consubstantial with it, whether or not this evidence is defined on different levels. If the notion of levels of evidence can potentially play a heuristic role that makes this uncertainty apparent, it seems that it could contribute to reducing the uncertainty linked to the results produced by methodologies placed at the top of the EBM hierarchy.

EBM accepts the recognition of uncertainty in its principles and even attempts to some extent to promote it. In managing knowledge that is conjectural by essence, the acceptation of the "uncertainty" dimension in clinical medicine is a prerequisite to rational and coherent practice. Therefore, one of the main strengths of the EBM formalisation as conceived by the reformers that developed it, is that it introduces statistical uncertainty into medical reasoning. There are however two limitations to which practice confronts us on a daily basis. The first, notwithstanding the difficulties in apprehending the epistemological significance of the notion of evidence, consists in the extremely widespread conceptual error that leads statistical significance to be assimilated to certainty of difference, or a statistical correlation with the cause. The second limitation concerns what is outside statistical significance, i.e. what is not validated but is not necessarily false. We can observe that, even without openly promoting it, EBM is responsible for the negation of presumption of non-provability by assimilating absence of evidence and negative results. Indeed, "absence of evidence is not proof for absence" and not evidencing a difference between two treatments, despite statistical power deemed adequate, is often assimilated to proof of the absence of a difference. This means ignoring the risk of error that we accept each time a statistical test is carried out.

The pedagogical approach contained in EBM should have enabled this erroneous interpretation of statistical tests to be contested and should have made the users of knowledge aware of the consubstantial uncertainty of processed knowledge.

The failings previously mentioned that lead to Statistic-Based Medicine, which we seek to distinguish from Science-Based Medicine, find their explanation in the generalised use of frequentist statistical

methods. These methods entail calculations of frequency, standard deviation, and statistical tests expressed by a "*p*" value, and they have enjoyed a dominant status since they were developed by Fisher, Neyman and Pearson in the 1920s and 1930s. They have indeed been used in a quasi-ubiquitous manner in scientific medical publications, leaving very little space for alternative approaches. Frequentist reasoning relies on the null hypothesis test summarised by Meyer *et al.* as follows: when a trial is carried out comparing two treatments,

> We hypothesise that the two treatments do not differ (null hypothesis), which is the same as saying that, because of the effect of randomisation, there is little likelihood that a substantial difference will appear; we quantify the difference between the two treatments; if the difference is large and has less than a 5% likelihood of being observed, we assume that the initial null hypothesis was false and therefore that the two treatments differ. We say that p<5% (or p<0.05) and that the test is significant at a threshold of 5% (…) It is important to stress the correct interpretation of p<0.05. It means that if the hypothesis of treatment equality is true, there is less than 5% likelihood that a difference as great as that observed in the trial will be observed by chance. However, p does not in any way tell us what the probability is for the hypothesis of treatment equality to be true (or false). (Meyer, Vinzio and Goichot 2009; Browner and Newman 1987)

However, the misunderstandings or the false interpretations of this *p* value by users, whether clinicians or researchers, are widespread (Goodman 1999; Browner and Newman 1987; Freeman 1993; Diamond and Forester 1983; Andreu *et al.* 2021). There are many who think that *p* is a direct measure of the probability that a null hypothesis is false. This is a false belief in a measure where p is calculated considering that the null hypothesis holds true. This error of logic reinforces the dominant conception that the data alone can provide the probability that a hypothesis is true. Steven Goodman compares the use of hypothesis testing in medicine to a justice system that is not concerned with the guilt or innocence of defendants, but focuses on controlling the number of errors in judgements made (Goodman 1999). This demonstrates the logical impossibility of expressing the power of evidence under the null hypothesis with only one figure and the frequency of type-1 error under this same hypothesis. Pearson and Neyman recognised this impossibility by underlining that the search for objectivity implied that a price should be paid, that of abandoning the idea of passing judgement on evidence, and settling for statistical significance (Neyman and Pearson 1933). Given this impossibility, a combinatory logic was established between the hypothesis test and the calculation of *p* by fixing type-1 error in advance (5% in

general) as well as the power (80% or more) before the experiment, and then by calculating *p* from the data and rejecting the null hypothesis if *p* is below the already fixed type-1 error. This combinatory logic leads to the error consisting in assimilating *p* to the so-called type-1 error. If for statisticians the limitations of this combinatory logic are the subject of great debate, these debates do not appear in the scientific literature, where statistics are applied and appear as mathematical abstractions arousing no controversy whatsoever. The objective nature of the method explains its dissemination beyond its boundaries. The extension of these methods, as shown by Marks, has triggered a shift of power from those who possess physiological knowledge to those who master quantitative methods (Marks 2000). Furthermore, this statistical procedure exerts a form of tyranny over modes of thought (Skellam 1969), and only tests of statistical significance give data a value. This is how evidence-based medical reasoning, which calls itself scientific, is in reality statistic-based and makes no attempt to reposition data in the wider context of scientific knowledge and physiological and physio-pathological knowledge. The method has a lot to do with this, leaving us to believe that each study is able on its own to generate conclusions for which the truth is tempered by a certain probability, instead of considering that it is only a building block in the construction of knowledge within a theoretical framework. The heart of scientific practice, as underlined by Goodman, is this space for discussion and criticism where data from real life, laboratory work and previous research is collated; the combination of the hypothesis test and *p* does not enable this crucial task to be achieved (Goodman 1999).

One of the solutions suggested to escape the *p dictatorship* would be a return to the Bayesian approach, which was contested precisely because it introduced subjectivity. Yet data will only gain in validity by the re-introduction of this subjectivity via expert opinion, previous data and physio-pathological data.

Bayes' theorem can be set out in the following way:

$$a_1 = \frac{p_1 q_1}{p_1 q_1 + p_2 q_2}$$

In this formula, a^1 designates the probability we are seeking, i.e. the *a posteriori* probability as it emerges from the experiment for hypothesis 1 to be true. To calculate a^1, we need to have the following *a priori* probabilities: p^1 (probability before the experiment that hypothesis 1 is true), q^1 (probability that the experiment will be successful if hypothesis 1

is true), q^2 (probability that the experiment will be successful if hypothesis 2 is true, i.e. hypothesis 1 is false, as q^1 and q^2 are two independent probabilities). In a clinical trial comparing molecules A and B in a given indication, a "significant" difference at the 0.05 threshold means that the conditional probability that their efficacy will differ is 0.95. We have previously observed that one common error consists in confusing statistical significance and probability, and that the $p^1 = p^2$ assumption is purely arbitrary. The use of outside information via an a priori distribution (p^1) enables conclusions obtained in a therapeutic trial to be modulated. By introducing conclusions from previous studies in the Bayesian analysis of results from a trial (in the *a priori* distribution) there is a cumulative effect on the data. This cumulative effect is characteristic of scientific functioning, and it is all the observations obtained in all the studies conducted that have contributed to the conclusion (Meyer, Vinzio and Goichot 2009). Using a Bayesian approach to re-analyse data initially processed following a frequentist approach, Brophy et al. demonstrated the discrepancy that can occur between the results (Bromphy & Joseph, 1995). Meyer's explanations on the usefulness of this approach are eloquent:

> The *a priori* law also enables different viewpoints to be confronted. In a specific domain, several experts may have divergent opinions on the same issue. Bayesian analysis makes it easy to confront these opinions by carrying out calculations with each of these laws defined *a priori* by experts. If the conclusions are the same or very close, whatever the *a priori* distribution, the divergence then wanes in the face of the data and the conclusion is robust. If on the contrary the analysis provides very different results, it suggests that knowledge needs to be expanded by increasing the number of observations before a decision can be made. There again, a frequentist approach does not enable different expert opinions to be readily confronted. Moreover, if the data is sufficiently substantial, the result will be stable whatever the *a priori* distribution used. All this goes to show that the *a priori* law does not play an exclusive role and that it is not possible to use it to assert just anything, as the data should always have the last say when it is available. However, when experimental subjects, and therefore data, are sparse (costly examinations, orphan diseases, the very small numbers of transgenic animals), Bayesian methods enable the conclusions of an experiment to be modulated and the researchers' reflection to be enriched, by making use of this *a priori* law. But here again, these few lines should not lead to the conclusion that Bayesian methods systematically enable results to be obtained, but they at least enable knowledge to be enriched through more refined reflection on the results. (Meyer, Vinzio and Goichot 2009).

Calculation of this *a priori* probability can only result from a theorisation of the problem. This necessity, fairly apparent the in Bayesian approach, is far removed from the original EBM methods. The fact that results, however relevant they are, cancel previously acquired knowledge and the little consideration given to physio-pathological data are among the approaches that tend to dissociate experimental science from theory, establishing knowledge on the sole calculation of a number, the statistically significant result. However, there is no experimental science without theory (Foncin 2007), this is true for all informal sciences, including medical science. This shift away from theory places EBM closer to Comte's positivism. In his positivist philosophy, Auguste Comte indeed considered the notion of cause to be a "metaphysical" notion, which therefore needed to be banished. Our relationship with cause today is complex, and while we do not banish it completely as such, we banish the physio-pathological explanation from the process of elaboration of evidence: it will at most serve to bolster results once the facts have been proven. In the same way as past knowledge, physio-pathological data participates in this theorisation of the problems on which EBM, as it is practised today, too often turns its back.

If some ardently campaign in favour of a more frequent resort to Bayesian approaches (Meyer, Vinzio and Goichot 2009; Goodman 1999), this evolution, even though it is timid, does exist. From 365 medical publications referenced on the PubMed database in 2000 involving Bayesian approaches, the number increased in 2021 to more than 8000. A sure sign that an evolution is on-going.

The issue of external validity

Archibald Cochrane, one of the founding fathers of EBM, stated that between the measure carried out via a randomised clinical trial and the benefit to the population, there was an often-underestimated gulf (Cochrane 1972). It is precisely this gap that leads us to discuss what is called the external validity of a study. Notwithstanding its internal validity, which guarantees limited bias for a study, its external validity is of considerable importance, as on it will depend the usefulness of the study in clinical practice and the applicability of the results to a group of patients with given characteristics. With its hierarchy of levels of evidence, in the conflict between internal and external validity, EBM has very clearly chosen to opt in favour of the first. By favouring internal validity, a flagrant lack of consideration for the determining factor of the applicability of the results can be observed. This lack of consideration for

the external validity of randomised controlled trials in the initial EBM model, resulting in an over-determination of the internal validity, has led to underuse in clinical practice of treatments despite the fact that they had proved their efficacy in trials with good internal validity and had even sometimes been recommended by experts on the basis of these results (Mant 1999; Davis and Taylor-Vaisey 1997; Cabana *et al.* 1999). Peter Rothwell showed this imbalance between internal and external validity, on the basis of the literature itself. He thus showed that methodological research to improve the internal validity of trials and systematic reviews was more substantial in volume than research focusing on the way results should be applied to clinical practice (Rothwell 2005). For a long time, organisations in charge of assessing new drugs with a view to their marketing authorisation, such as the Food and Drug Administration in the USA, focused only very little during their decision-making process on the usefulness of the treatment in a particular population, and finally only retained the results for the sample (Wermeling 1999). Research agencies such as the *UK Medical Research Council* has produced documents serving as a referential for the conduct of randomised controlled trials that heavily stress requisites for internal validity, with a number of recommendations to ensure it. It is surprising to observe that in terms of external validity, nothing was then explicitly formulated on the importance of guaranteeing the clinical applicability of the results.

For instance, The RALES trial published in 1999 concluded to the improvement in prognosis for congestive heart failure under the effect of spironolactone when it was combined with an angiotensin-converting enzyme (ACE) inhibitor, a molecule usually prescribed in this context (Pitt *el al.* 1999). The publication of this trial in the *New England Journal of Medicine* was followed by a change in attitudes among practitioners, who from then on combined ACE inhibitors and spironolactone, thus following the trial results. This change in prescription habits was assessed by Juurlink *et al*., who showed that the numbers of patients hospitalised for congestive heart failure under ACE inhibitors and spironolactone, which were 34 out of 1000 in 1994, had increased to 149 out of 1000 just a few months after the publication of the RALES trial ($p<0.001$) (Juurlink *et al.* 2004). As shown by Dreyfuss, using the morbidity data collected by Juurlink, the adjunction of spironolactone to ACE inhibitors coincided "in real life" with a marked increase in the number of hospitalisations for hypokalaemia (rising from 2.4 for 1000 in 1994 to 11.0 for 1000 in 2001, $p<0.001$), and in deaths linked to this ionic disorder (from 0.3 out of 1000 in 1994 to 2.0 out of 1000 in 2001, $p<0.001$), but with no decrease in the number of hospitalisations for heart failure. In this case, the internal

validity of the study, i.e. its methodological quality, came as a direct contradiction to its external validity (Dreyfuss 2005). It is because methodological "purity" was required, capable of guaranteeing the best internal validity, that the investigators drew away from the population of patients to treat on the basis of inclusion and exclusion criteria that drastically reduced the heterogeneity of the sample. In the end, with the extreme reduction of the sample to obtain a homogeneous sub-population, samples lose their representativeness and studies lose clinical credibility. It is this balance between robustness (internal validity) and relevance (external validity) that forms the concept of reliability (Coutellec 2019).

By complying with an EBM diktat, which favours internal validity to the detriment of the applicability of the results to clinical practice, we tend to forget too soon that clinicians are at the service of patients and methodologists are at the service of clinicians, and not the reverse. The finality of research remains the improvement of the care of patients, and with the reversal of power, where methodologists have the upper hand on clinicians, EBM has lost sight of this objective in favour of methodological purity. We say "lost sight" for a purpose, as we tend to believe that it is a transformation of the initial project, which has somehow escaped from its creators. Whether David Sackett, Gordon Guyatt or their predecessors and inspirers, Bradford Hill, Archibald Cochrane, or even Alvan Feinstein, all in their own time were aware of the limitations of EBM and never lost sight of the fact that this method was above all intended to serve clinical practice. Feinstein, on the subject of the *University Group Diabetes Programme* study, precisely because of his expertise in statistics and methodology, warned against the dangers of sacrificing *clinical wisdom and scientific judgment to rigid doctrines inspired by statistics* (Marks 2000). As the trial developed and the methodology took precedence, he could see a decrease in its potential interest for clinicians who had the task of offering treatments to singular individuals.

A critical approach thus means having the ability to question both the internal validity of a trial and its external validity, in order to potentially not integrate a result into clinical decisions, even if the methodology that produced it places it at the top of the EBM hierarchy. Nevertheless, it is clear that EBM, in the face of criticism, has been able to make its model evolve, and the GRADE system is a perfect illustration of this, integrating applicability among the factors moderating the effects of the methodological design (D. Atkins *et al.* 2004). Another illustration of this evolution is the observable broadening of characteristics of populations that can be included in trials.

The time parameter

A British doctor ironically complained about being so busy with exploiting databases and conducting critical reviews of article abstracts that he did not have time to see his patients (Grahame-Smith 1995). A database such as Medline references more than 30 million biomedical publications, published in nearly 5600 indexed periodicals. More than 12,000 randomised controlled trials are published each year and integrated into the Medline database, and the number of systematic reviews published each year is more than 30,000, against 1,400 a year in 2000. Richard Grol calculated in 2001 that it would take two and a half days a week for physicians to read all the published procedures in their fields of practice (Grol 2001) - it would certainly take much longer today. To imagine an entirely *Evidence-Based* practice without taking this time parameter into consideration is like dreaming of a brand of medicine where all participants would be machines and where all decisions would be produced by algorithms - medicine where the human factor would be put aside. Thankfully of course, this is not the case, but despite the filters and syntheses EBM offers, time remains a central problem that comes up against the will to rationalise. It is not through negligence but more often because of a lack of time that it is not possible to implement everything that could lead to the most suitable decisions on the basis of published data. Ensuring good medical practice by constantly updating knowledge and basing decisions on more reliable data takes time, and to think that the application of EBM principles can enable that time to be reduced is probably an illusion.

Practitioners' dual dependence

EBM is seen as a way of practicing medicine suited to the greatest number of situations and to the greatest number of practitioners. If the time parameter already seems to get between this claim and reality, the way EBM is attuned to practitioners' training is also an obstacle to the implementation of EBM practice respectful of its principles.

The prerequisites to understanding the results of a clinical trial are to understand the epidemiological vocabulary, to have minimum basic knowledge of the methods used, and to have basic notions of statistics and clinical epidemiology. Surveys on practitioners have shown very clearly that this prerequisite is far from being met in all contexts (Perneger *et al.* 2004; Estellat *et al.* 2006; Heller *et al.* 2004; Godwin and Seguin 2003;