Learning from Principles of Evidence-Based Medicine to Optimize Nonclinical Research Practices

Isabel A. Lefevre and Rita J. Balice-Gordon

Contents

Abstract

Thousands of pharmacology experiments are performed each day, generating hundreds of drug discovery programs, scientific publications, grant submissions, and other efforts. Discussions of the low reproducibility and robustness of some of this research have led to myriad efforts to increase data quality and thus

I. A. Lefevre (✉)
Rare and Neurologic Diseases Research, Sanofi, Chilly-Mazarin, France
e-mail: Isabel.Lefevre@sanofi.com

R. J. Balice-Gordon
Rare and Neurologic Diseases Research, Sanofi, Framingham, MA, USA

A. Bespalov et al. (eds.), *Good Research Practice in Non-Clinical Pharmacology and Biomedicine*, Handbook of Experimental Pharmacology 257,
https://doi.org/10.1007/164_2019_276

reliability. Across the scientific ecosystem, regardless of the extent of concerns, debate about solutions, and differences among goals and practices, scientists strive to provide reliable data to advance frontiers of knowledge. Here we share our experience of current practices in nonclinical neuroscience research across biopharma and academia, examining context-related factors and behaviors that influence ways of working and decision-making. Drawing parallels with the principles of evidence-based medicine, we discuss ways of improving transparency and consider how to better implement best research practices. We anticipate that a shared framework of scientific rigor, facilitated by training, enabling tools, and enhanced data sharing, will draw the conversation away from data unreliability or lack of reproducibility toward the more important discussion of how to generate data that advances knowledge and propels innovation.

Keywords
Data reliability · Decision-making · Evidence-based medicine · Nonclinical pharmacology · Research methodology

1 Introduction

Over the last 10 years, debate has raged about the quality of scientific evidence, expanding from a conversation among experts, amplified by systematic reviews and meta-analyses published in peer-reviewed journals, into a heated discussion splashed across mainstream press and social media. What is widely perceived as a "reproducibility crisis" is the subject of countless, and sometimes inaccurate, statements on the poor "reproducibility," "replicability," insufficient "rigor," "robustness," or "validity" of data and conclusions. In the context of nonclinical pharmacological data, these are cited as foundational for later clinical trial failure. The decision to advance a compound to human testing is based on a substantial body of evidence supporting the efficacy and safety of a therapeutic concept. Nonclinical studies that support, for example, an investigational new drug (IND) filing or a clinical trial application (CTA), which gate studies in humans, are reviewed under quality control procedures; most safety studies must comply with regulations laid out by health authorities, whereas nonclinical efficacy studies are usually performed in a nonregulated environment (see chapter "Quality in Non-GxP Research Environment"). If clinical trial results support both efficacy and safety of the intervention, health authorities review *all* of the evidence, to determine whether or not to approve a new therapeutic.

Once a new therapeutic is made available to patients and their physicians, clinical trial findings and real-world observations contribute to forming a larger body of evidence that can be used for decision-making by a physician considering which treatment option would best benefit a patient. In many countries, medical students are taught to critically appraise all the accessible information in order to choose the "best possible option," based upon the "best possible evidence"; this process is part of evidence-based medicine (EBM), also known as "medicine built on proof." In EBM, clinical evidence is ranked according to the risk of underlying bias, using the available sources of evidence, from case studies through randomized, controlled clinical trials (RCTs) to clinical trial meta-analyses. Well-designed randomized trial

results are generally viewed to be of higher reliability, or at least less influenced by internal bias, than observational studies or case reports. Since meta-analysis aims to provide a more trustworthy estimate of the effect and its magnitude (effect size), meta-analyses of RCTs are regarded as the most reliable source for recommending a given treatment, although this can be confounded if the individual RCTs themselves are of low quality.

A well-established framework for rating quality of evidence is the Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) system (http://www.gradeworkinggroup.org). GRADE takes the EBM process a step further, rating a body of evidence, and considering internal risk of bias, imprecision, inconsistency, indirectness, and publication bias of individual studies as reasons for rating the quality of evidence down, whereas a large effect, or a dose-response relationship, can justify rating it up (Balshem et al. 2011). The Cochrane Collaboration, which produces systematic reviews of health interventions, now requires authors to use GRADE (https://training.cochrane.org/grade-approach). The British Medical Journal has developed a suite of online tools (https://bestpractice.bmj.com/info/us/toolkit) with a section on how to use GRADE, and various electronic databases and journals that summarize evidence are also available to clinicians. In a recent development, the GRADE Working Group has begun to explore how to rate evidence from nonclinical animal studies, and the first attempt to implement GRADE in the nonclinical space has successfully been performed on a sample of systematic reviews and examples, with further efforts planned (Hooijmans et al. 2018). In contrast, with the exception of those who also have medical training or clinical research experience, most scientists are unaware of the guiding principles of EBM and are unfamiliar with formal decision-enabling algorithms. At least in part due to the diversity of nonclinical experiments, systematic reviews and meta-analyses are far less common in nonclinical phases than in clinical ones, and there are very few broadly accepted tools with which to assess nonclinical data quality (Hooijmans et al. 2018; Sena et al. 2014). Pioneering work in this area came from the stroke field, with nonclinical research guidelines and an assessment tool elaborated by STAIR, the Stroke Therapy Academic Industry Roundtable (Hooijmans et al. 2014) (https://www.thestair.org). The CAMARADES collaboration (originally the "Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental Stroke") has now extended its scope to support groups wishing to undertake systematic reviews and meta-analyses of animal studies in research on neurological diseases (http://www.dcn.ed.ac.uk/camarades). The Systematic Review Centre for Laboratory Animal Experimentation (SYRCLE) has designed a comprehensive method to systematically review evidence from animal studies (Hooijmans et al. 2014), based on the Cochrane risk of bias tool. SYRCLE's tool covers different forms of bias and several domains of study design, many of which are common to both clinical and nonclinical research (Table 2 in Hooijmans et al. 2014). As a consequence, measures known to reduce bias in clinical settings, such as randomization and blinding, are recommended for implementation in nonclinical research. Although the tool was primarily developed to guide systematic reviewers, it can also

be used to assess the quality of any in vivo experimental pharmacology study. However, these structured approaches have had limited uptake in other fields.

The attention to sources of bias that can influence study conduct, outcomes, and interpretation is an essential element of EBM. A catalog of bias is being collaboratively constructed, to map all the biases that affect health evidence (https://catalogofbias.org). In the nonclinical space, despite a number of publications and material from training courses and webinars (e.g., http://neuronline.sfn.org/Collections/Promoting-Awareness-and-Knowledge-to-Enhance-Scientific-Rigor-in-Neuroscience), an equivalent, generalizable framework, or a common standard for rating evidence quality is lacking, as is a unified concept of what constitutes the best possible material for decision-making. Discussions are also limited by confusing and varied terminology, although attempts have been made to clarify and harmonize terms and definitions (Goodman et al. 2016). Here we will use the word "reliability," in its generally accepted sense of accuracy and dependability, of something one expects to be able to rely on to make a decision. As a consequence, reliability also reflects the extent to which something can consistently be repeated. Both meanings apply to experimental pharmacology studies in all parts of the biomedical ecosystem.

The EBM framework is used to find reliable answers to medical questions. Here we will describe the purposes, current practices, and factors that contribute to bias in addressing scientific questions. We will consider which EBM principles can apply to nonclinical pharmacology work and how to strengthen our ability to implement best research practices, without limiting innovation that is urgently needed.

2 Current Context of Nonclinical, Nonregulated Experimental Pharmacology Study Conduct: Purposes and Processes Across Sectors

2.1 Outcomes and Deliverables of Nonclinical Pharmacology Studies in Industry and Academia

Experimental pharmacology studies in biopharma companies and nonclinical contract research organizations (CROs) can have various purposes, such as furthering the understanding of a disease mechanism, developing a model or assay, or characterizing the effects of a novel compound. Such studies can also document a patent application and/or generate data on the efficacy or safety of a compound that is to enter clinical development, in which case the study report may ultimately be part of a regulatory submission to a health authority. In academia, the primary goal is to provide experimental evidence to answer scientific questions and disseminate new knowledge by publishing the findings; academic scientists also occasionally file patents and, in collaboration with biopharma companies, perform studies that may in turn become part of regulatory submission dossiers. Academic drug discovery platforms, which have sprouted in recent years, mainly aim to provide nonclinical

data that will be further leveraged in biopharma drug discovery programs, although it is increasingly common that these data are used to advance clinical studies as well.

Different business models and end goals across academia and industry, and different outcomes of nonclinical research, imply different processes and deliverables, which can be associated with a step or feature in EBM, as described in Table 1.

Investigating a scientific hypothesis is often done in a stepwise manner; from an initial idea, several questions can be asked in parallel, and answers are generated in both an incremental and iterative manner, by performing additional experiments and repeating cycles. Choices and decisions are made at each step, based on data; if these data are under- or overestimated, their interpretation will be biased, affecting subsequent steps. For example, in a drug discovery project, inaccurate estimates of in vitro potency or in vivo efficacy can skew the doses tested in nonclinical safety experiments, and bias the estimate of the dosage range in which *only* the desired response is observed in nonclinical species, and, most importantly, affect the subsequent determination of the corresponding dosage range to be tested in humans. As sponsors of clinical trials, among other responsibilities, biopharma companies have an ethical duty to conduct human trials *only* if there is a solid foundation for a potential clinical benefit with limited safety risks. In academic research, individuals and institutions are accountable to funders and to the community for contributing to the body of scientific knowledge. In all fields and sectors, biased interpretations of experimental data can result in wasted experiments; scientists are therefore responsible for the quality of the evidence generated.

2.2 Scientific Integrity: Responsible Conduct of Research and Awareness of Cognitive Bias

Over the last two decades, many governments and agencies involved in funding and conducting research have taken a strong stance on scientific integrity, issuing policies and charters at international, national, and institutional levels. Compliance with these policies is mandatory for employees and scientists applying for funding (examples: MRC, https://mrc.ukri.org/publications/browse/good-research-practice-principles-and-guidelines; NIH, https://grants.nih.gov/policy/research_integrity/what-is.htm; CNRS, http://www.cnrs.fr/comets/IMG/pdf/guide_2017-en.pdf). Scientific integrity means absolute honesty, transparency, and accountability in the conduct and reporting of research. Responsible research practices encompass the adherence to these principles and the systematic use of measures aiming to reduce cognitive and experimental bias.

Training on responsible scientific conduct is now mandatory at masters or PhD level in many universities; at any stage of their career, scientists can access training resources on scientific integrity and responsible research practices (see list made by EMBO, http://www.embo.org/science-policy/research-integrity/resources-on-research-integrity; NIH, Responsible Conduct of Research Training, https://oir.nih.gov/sourcebook/ethical-conduct/responsible-conduct-research-training; Mooc, https://www.fun-mooc.fr/courses/course-v1:Ubordeaux+28007EN+session01/about#). The US Department

Table 1 Parallel between EBM and nonclinical research purposes and processes across organizations

	In private sector nonclinical research	In academic nonclinical research	In EBM
Outcomes and deliverables	Patents (intellectual property) Decision to move a compound to clinical development Nonregulated study reports CRO study reports, data for customers Additions to catalog	Publications Patents (intellectual property) Study reports and data provided to public or private funders	Recommendations Guidelines Treatment decisions
Process	**Purpose in biopharma companies and CROs**	**Purpose in academia**	**Relevant EBM feature**
Initiating a research project	Driven by company strategy Triggered by prior data, exploratory studies, literature CROs: mainly triggered by requests from customers and market opportunities	Driven by science and funding opportunities Triggered by prior data, exploratory studies, literature, serendipity	Framing a question, collecting all available data, and ranking quality of evidence
Existence and use of guidelines	Company nonclinical quality and compliance rules, best research practice guidelines, patent department rules	Variable; rules of institutions, funding agencies, grant applications, journals, built into collaborations	Guidelines on use of EBM and EBM guidelines
Use of experimental bias reduction measures in study design and execution	Variable; field-dependent Detailed study plans usually mandatory for compounds selected to enter clinical development (less so for early test compounds) and systematically used by CROs	Variable; field-dependent; funding or grant-dependent; increasing due to pressure from funders, journals, peers; awareness that credibility is suffering	Core feature of EBM: studies with lowest risk of bias assumed to be most reliable
Biostatistics: access and use	Company biostatisticians and software (mostly proprietary); mandatory review of statistical analyses for compounds entering clinical development CROs: variable	Variable, somewhat "do-it-yourself": depending on statistical literacy or access to relevant expertise, widespread use of commercially available suites, free online tools	Adequate study power Meta-analyses
Data: integrity, access, and sharing	Electronic lab notebooks, electronic data storage, dedicated budgets Mandatory archive of all data and metadata for clinical stage compounds Restricted company-only access	Variable, depending on institution and resources, in particular to fund long-term safekeeping of data Ability to access data highly variable	Access to all data in real time

of Health and Human Services Office of Research Integrity has developed responsible conduct of research training courses that incorporate case studies from an academic research context (https://ori.hhs.gov/rcr-casebook-stories-about-researchers-worth-discussing). Several companies have adopted a similar case-based approach from a biopharma context.

Inaccuracy and biased interpretations are not necessarily due to purposeful scientific misconduct; in fact, most of the time, they are inadvertent, as the consequence of poor decision-making, training, or other circumstances. Mistakes can be made and can remain undetected when there is no formal process to critically review study design in advance of execution, an essential step when study outcomes gate decisions with long-term consequences, in particular for human subjects and patients. One aspect of review is to examine the multiple forms of bias that compromise data reliability, confounding evidence, and its analysis and interpretation. Experimental protocols can be biased, as can be experimenters, based on individual perceptions and behaviors: this is known as cognitive bias, i.e., the human tendency to make systematic errors, sometimes without even realizing it. Particularly problematic is confirmation bias, the tendency to seek and find confirmatory evidence for one's beliefs, and to ignore contradictory findings. Scientists can work to develop evidence to support a hypothesis, rather than evidence to contradict one. Beyond designing and performing experiments to support a hypothesis, confirmation bias can extend to reporting only those experiments that support a particular expectation or conclusion. While confirmation bias is generally subconscious, competition – for resources, publications, and other recognitions – can obscure good scientific practice. Confirmation bias can be both a cause and a consequence of publication or reporting bias, i.e., omissions and errors in the way results are described in the literature or in reports; it includes "positive" results bias, selective outcome reporting bias, "Hot stuff" bias, "All is well literature" bias, and one-sided reference bias (see definitions in https://catalogofbias.org).

In industry and academia, there are both common and specific risk factors conducive to cognitive bias, and awareness of this bias can be raised with various countermeasures, including those listed in Table 2.

2.3 Initiating a Research Project and Documenting Prior Evidence

Scientists running nonclinical pharmacology studies may have different goals, depending on where they work, but initiating a research project or study is driven by questions arising from prior findings in all parts of the biomedical ecosystem. When deciding to test a new hypothesis from emergent science, or when setting up a novel experimental model or assay, scientists generally read a handful of articles or reviews, focusing on the most recent findings. Many scientists methodically formulate an answerable question, weighing the strength of the available evidence and feasibility as primary drivers. Published findings can be weighed heavily as "truth," or disregarded, based on individual scientific judgment and many other factors. When subjective factors, such as journal impact factor, author prominence, or

Table 2 Factors that contribute to manifestations of bias and potential countermeasures

Contributing factors	In biopharma companies and CROs	In academia
Awareness and knowledge of risks of bias or misconduct	Multiple levels of review and quality control can highlight unconscious biases In-house training programs on responsible conduct of research increasingly common	Growing number of online material and training programs (see examples in Sect. 2.2)
Risk factors conducive to bias or misconduct	"Pace of business": compensation linked to performance/timelines, competitive landscape, career aspirations, customer deadlines	"Publish or perish": priority given to novel findings due to academic competition, career aspirations, funding mechanisms, and durations
Measures and incentives to increase responsible conduct	Occasional individual performance metrics CROs: responsible conduct linked to credibility, a key factor of company success	Recognition, publication, citation in leading journals with strict reporting guidelines, awards for reproducibility attempts (e.g., https://www.ecnp.eu/research-innovation/ECNP-Preclinical-Network-Data-Prize.aspx)

other subjective reasons, are weighed more heavily than the strength of the evidence, a form of bias is embedded from the conception of a research project. Similarly to the flowchart approach used in EBM, where the first step is to frame the clinical question and retrieve all the related evidence, explicitly defining a question and systematically reviewing the literature should be a common practice in nonclinical pharmacology. When deciding to work on a target, biopharma scientists also have to consider whether modulating it could potentially result in adverse effects, so the background evidence to be weighed may have other aspects than for an academic research project. An obstacle to a comprehensive assessment of prior data is that data can be published, unpublished, undisclosed, or inaccessible behind a paywall or another company's firewall or simply out of reach due to past archival practices (see Sect. 2.7). Publication and selective outcome reporting biases will therefore be present in most attempts to review and weigh prior evidence. Thus, in practice, the data a scientist will evaluate at the start of a research project is often incomplete, raising the possibility of flawed experimental design, execution and interpretation, as well as the risk of confirmation and related biases.

2.4 Existence and Use of Guidelines

Recommendations on how to design and conduct nonclinical, nonregulated research studies can be found in scientific publications, in scientific society or institution guidelines, and in grant application guidelines. Although recommended "best research practices" have been around for at least a decade, there are no consensus, universal nonclinical pharmacology quality guidelines, but instead a collection of constantly evolving, context, and type-of-experiment-specific suggestions.

Biopharma companies and nonclinical CROs generally have internal guidelines. Scientists are expected to record results in real time in laboratory notebooks, should an organization or individual need to document data and timelines to establish inventorship. Guidelines produced by research quality departments therefore focus on how scientists should record the results of their research, and deviations from standard operating procedures, in order to fulfill legal and regulatory requirements, more than on study design or the use of measures to reduce experimental bias. In the private sector, research quality guidelines and best practice recommendations are generally confidential documents. In publications, research quality guidelines and implementation are rarely mentioned. While indirect, study reporting guidelines (see Sect. 2.7) are slightly more cited, but determining to what extent these were followed is far from trivial.

2.5 Use of Experimental Bias Reduction Measures in Study Design and Execution

The core principle of EBM is that the most reliable evidence comes from clinical studies with the lowest risk of bias and typically those that are designed with adequate power, randomization, blinding, and a pre-specified endpoint, in a clinically relevant patient population. There are many resources to help investigators plan human studies, such as the SPIRIT statement (http://www.spirit-statement.org), an evidence-based guideline for designing clinical trial protocols, which is being developed into a web-based protocol building tool. There are fewer resources to assist scientists in designing nonclinical studies; an example is the NC3Rs' Experimental Design Assistant (EDA, https://www.nc3rs.org.uk/experimental-design-assistant-eda) for in vivo animal studies. Experimental protocols can be found in publications or online, but they are primarily written to provide details on technical aspects, and do not always explicitly address the different sources of experimental bias.

In biopharma research, study plans which describe the study design and experimental methods in full detail, including the planned statistical methods and analyses, and any deviations to these plans as the study progresses, are usually mandatory for studies that are critical for decision-making. Study plans are more rarely written for exploratory, pilot studies. Nonclinical CROs use study plan templates that include statistical analysis methodologies, which are generally shared with customers. In our experience, CROs and academic drug discovery centers are very willing to discuss

and adapt study designs to suit customer needs. Collaboratively building a study plan is a good opportunity to share knowledge, ensure that a study is conducted and reported according to expectations, and work to identify and reduce conscious and unconscious biases. Across all sectors, planning ahead for in vivo pharmacology studies is more elaborate than for in vitro experiments, due to animal ethics requirements and the logistics of animal care and welfare. However, nonclinical study plans are not normally published, whereas clinical trial protocols are available in online databases such as the EU (https://www.clinicaltrialsregister.eu) and US (https://clinicaltrials.gov/) registers. A few initiatives, such as OSF's "preregistration challenge" (Open Science Foundation, Preregistration Challenge, Plan, Test, Discover, https://osf.io/x5w7h), have begun to promote formal preregistration of nonclinical study protocols, as a means to improve research quality (Nosek et al. 2018). However, preregistering every single nonclinical pharmacological study protocol in a public register would be difficult in practice, for confidentiality considerations, but also due to a perceived incompatibility with the pace of research in all sectors.

Overall, our experience in the field of neuroscience is that the implementation of experimental bias reduction measures is highly variable, within and across sectors, and meta-analyses of scientific publications have shown that there is clearly room for improvement, at least in the reporting of these measures (van der Worp et al. 2010; Egan et al. 2016).

Different field- and sector-related practices and weights on bias reduction measures, such as blinding and randomization (see chapter "Blinding and Randomization"), can be expected. In the clinical setting, blinding is a means to reduce observer bias, which, along with randomization to reduce selection bias, underlies the higher ranking of RCTs over, for example, open-label trials. Both blinding and randomization are relevant to nonclinical studies because the awareness of treatment or condition allocation can produce observer bias in study conduct and data analysis. Neurobehavioral measures are among the most incriminated for their susceptibility to observer bias. But even automated data capture can be biased if there are no standards for threshold and cutoff values. Observer bias is also a risk, for example, when visually counting immunolabeled cells, selecting areas for analysis in brain imaging data, and choosing recording sites or cells in manual electrophysiology experiments. Blinding has its limitations; blinding integrity may lost, such as when using transgenic mice (which are often noticeably different in appearance or behavior compared to wild-type littermates) or in pathological settings that induce visible body changes, and the experimenter's unawareness of group allocation will not be sufficient to limit the effect observing animals can have on their behavior (analogous to the Hawthorne effect in social sciences, see https://catalogofbias.org/biases/hawthorne-effect/).

Differences in resource availability will influence practices, since training experimenters, standardizing animal handling and husbandry, and earmarking suitable lab space and equipment, among other considerations, are contingent upon funding. Nonclinical CROs are most likely to have strong guidelines, or at least evidence-based standard operating procedures, and to follow them, since credibility, transparency, and customer satisfaction are business-critical. The systematic use of

inclusion/exclusion criteria and blinding should be implemented as standard practice in all sectors of the biomedical ecosystem. However, while in the industry there is a tendency to optimize workflows through standardization, and similarly in academia, strong lab "traditions," one size does not necessarily fit all. Specific technical constraints may apply, in particular for randomization. For instance, in some in vitro experiments, features such as "edge effect" or "plate effect" need to be factored into the randomization procedure (https://paasp.net/simple-randomisation); liquid chromatography-coupled mass spectrometry experiments require additional caution, since randomizing the order in which samples from different groups or conditions are tested may be counterproductive if the risk of potential cross-contamination is not addressed. Randomizing the order of procedures, while often a sound measure to prevent procedural bias, may actually increase the risk of bias, if animals behave differently depending on prior procedures or paradigms. While randomization and blinding will generally be effective in reducing risks of selection and observer bias, they have no effect on non-contemporaneous bias, when control groups or samples are tested or analyzed at a separate time from treated ones.

Thus, both in EBM and in nonclinical research, high-quality designs aim to take into account all of the known sources of bias and employ the best available countermeasures. Among these, there are two universally critical items, a pre-specified endpoint with an estimate of the predicted effect size and the corresponding adequate statistical power to detect the predicted effect, given the sample size, all of which require a prior statistical plan.

2.6 Biostatistics: Access and Use to Enable Appropriate Design of Nonclinical Pharmacology Studies

Establishing an a priori statistical plan, as part of the study design, remains far from customary in nonclinical pharmacology, mainly because scientists can lack the adequate awareness and knowledge to do so. The latest Research Integrity report by the Science and Technology Committee in the UK (https://publications.parliament.uk/pa/cm201719/cmselect/cmsctech/350/350.pdf) emphasized that scientists need to learn and understand the principles of statistics, rather than simply being told of a list of statistical tests and software that does the analyses. In our experience, biologists' statistical proficiency appears to mostly be based on local custom and varies widely even in the same field of biology. This is illustrated by misleading phrases in methods sections of publications, such as "the number of animals used was the minimum required for statistical analysis," or "post hoc comparisons were carried out between means as appropriate," or "animals were randomly assigned to 4 groups," or "the experiments were appropriately randomized" (sic). A side effect of this phenomenon is that it hampers critical assessments of published papers; biologists confronted with unfamiliar terms may struggle to capture which study designs and analyses were actually conducted.

In practice, more attention is paid to statistics once the data have been generated. In nonclinical CROs the statistical analyses are provided to the customer in the full

study reports. In biopharma companies, for clinical development candidate compounds, it is generally mandatory that the proposed statistical analyses are developed and/or validated by a statistician. Many companies have developed robust proprietary statistics software, with specific wording and a selection of internally approved tests and analysis tools. Although in-house applications are validated and updated, they are not ideal for sharing results and analyses with external partners. Overall, and despite a call for cultural change in the interactions between scientists and nonclinical statisticians (Peers et al. 2012), it seems that the nonclinical pharmacology community remains under-resourced in this area. Insight gained through discussions on data quality among partners of several European initiatives suggests that there are too few research biostatisticians in all biomedical arenas.

When a thorough process is established beforehand, choosing a pre-specified endpoint to test a hypothesis and estimating an effect size for this endpoint are essential. While both are required in EBM, these are less common in nonclinical research. Clinical studies aim to detect a predetermined effect size, or a clinically relevant direction and effect magnitude, based on prior knowledge. In contrast, scientists generally have a rough idea of values that would be negligible, due to biological variation or to inaccuracy or imprecision, but considering which values are biologically meaningful tends to be done after, rather than before, running an experiment. When generating a hypothesis, i.e., in exploratory or pilot studies, it may be possible to choose an endpoint of interest, without necessarily defining its direction and amplitude. In contrast, prior estimates of effect size are essential when the aim is to demonstrate a pharmacological effect in confirmatory studies upon which decisions about next steps are based. This distinction between exploratory and confirmatory studies (Kimmelman et al. 2014 and chapter "Resolving the Tension Between Exploration and Confirmation in Preclinical Biomedical Research") is a determining factor in study design, but remains an underused concept in nonclinical work.

Arguably the most serious consequence of insufficient planning is that nonclinical studies are too often underpowered (Table 2 in Button et al. 2013) or are of unknown power, when publications fail to reveal how sample sizes were chosen (Carter et al. 2017). Despite its central role in the null hypothesis significance testing framework, which remains the most used in nonclinical pharmacology, for many scientists, statistical power is one of the least well-understood aspects of statistics. This may be because it is generally explained using abstract mathematical terms, and its role more extensively discussed in clinical research, or in psychology, than in biology. However, recognizing that inadequately powered studies can lead to unreliable conclusions on the direction and magnitude of an effect in a sample of the whole population is just as important in nonclinical pharmacology as it is in EBM. Assay development is by definition exploratory in initial attempts; but when the assay is going to be used routinely, sample sizes to achieve a desired statistical power need to be determined. Unfortunately, this is not yet the norm in nonclinical pharmacology, where decisions are often made on so-called converging evidence from several underpowered studies with different endpoints or on a single published study of unknown power, offering little confidence that the same effect(s) would be seen in the whole population from which the sample was taken.

As discussed above (see Sect. 2.5), randomization is essential to prevent selection bias across all sectors of research. Randomization can be achieved even with limited resources and applied in many nonclinical pharmacology studies regardless of their purpose and type, without necessarily involving statistical expertise. The randomization procedure must however be part of the study design, and statistical evaluation before a study is conducted can help determine which procedure is best suited.

2.7 Data Integrity, Reporting, and Sharing

Notwithstanding the existence of vast amounts of electronic storage space and sophisticated software to ensure file integrity, retaining, and potentially sharing, original datasets and protocols is not yet straightforward. Barriers to widespread data sharing are slowly being overcome, but there remains a need for long-term funding, and the ability to browse data long after the software used to generate or store them has become obsolete.

In biopharma companies and CROs, it is customary to retain all original individual and transformed data, with information on how a study was performed, in laboratory notebooks and annexes. Scientists working in industry are all aware that the company owns the data; one does not lose or inadvertently misplace or destroy the company's property, and in audits, quality control procedures, preparation for regulatory filings, or patent litigation cases, to name a few, original data must often be produced. This also applies to studies conducted by external collaborators. For compounds that are tested in human trials (including compounds that reach the market), all data and metadata must be safely stored and retrievable for 30 years after the last administration in humans. It is thus common practice to keep the records decades after they were generated (see item GRS023 in https://india-pharma.gsk.com/media/733695/records-retention-policy-and-schedule.pdf). Such durations exceed by far the life of the software used to generate or store the data and require machine-readable formats. Paper laboratory notebooks are also stored for the duration; their contents are notoriously difficult to retrieve as time passes, and teams or companies disperse. Electronic source data in FDA-regulated clinical investigations are expected to be attributable, legible, contemporaneous, original, and accurate (ALCOA). This expectation is also applied to nonregulated nonclinical data in many biopharma companies and in nonclinical CROs. The recent FAIR (findable, accessible, interoperable, reusable) guiding principles for scientific data management and stewardship (Wilkinson et al. 2016) are intended to facilitate data access and sharing while maintaining confidentiality if needed. To this date, broadly sharing raw data and protocols from biopharma research remains rare (but see Sect. 3.1).

Generally speaking, data generated in academia destined for publication are not as strictly managed. Institutional policies (see examples of data retention policies: Harvard, https://vpr.harvard.edu/files/ovpr-test/files/research_records_and_data_retention_and_maintenance_guidance_rev_2017.pdf; MRC, https://mrc.ukri.org/documents/pdf/retention-framework-for-research-data-and-records/) may state that data should be retained for a minimum of 3 years after the end of a research

project, a period of 7–10 years or more, or as long as specified by research funder, patent law, legislative, and other regulatory requirements. Effective record-keeping and retention is limited by funding and by the rapid turnover of the scientists performing most of the experiments; a classic problem is the struggle to find the data generated by the now long-gone postdoctoral associate. Access to original, individual data can be requested by other scientists or required by journals and funding agencies or, on rare occasions, for investigations of scientific misconduct. Although academic data and metadata sharing is improving (Wallach et al. 2018), with extended supplementary materials and checklists, preprint servers, data repositories (Figshare, https://figshare.com; OSF, https://osf.io; PRIDE, https://www.ebi.ac.uk/pride/archive), and protocol sharing platforms (https://experiments.springernature.com; https://www.protocols.io), universal open access to data is yet to be achieved.

In biopharma companies, there is an enormous amount of early discovery studies, including but not limited to assay development and screening campaigns, with both "positive" and "negative" data, that are not intended per se for publication, even though many could be considered precompetitive. A relatively small proportion of conducted studies is eventually published. However, for each compound entering clinical development, all the results that are considered relevant are documented in the nonclinical pharmacology study reports that support IND and CTA filings. A summary of the data is included in the nonclinical overview of the application dossiers and in the Investigator's Brochure. From these documents it is often difficult to assess the quality of the evidence, since they contain relatively little experimental or study information (Wieschowski et al. 2018); study design features are more likely to be found in the study reports, although there are no explicit guidelines for these (Langhof et al. 2018). The study reports themselves are confidential documents that are usually only disclosed to health authorities; they are intended to be factual and include study plans and results, statistical plans and analyses, and individual data.

In academia, publishing is the primary goal; publication standards and content are set by guidelines from funders, institutions, partners, peer reviewers, and most importantly by journals and editorial policies. In recent years, journal guidelines to authors have increasingly focused on good reporting practices, implementing recommendations from landmark publications and work shepherded by institutions such as the NC3Rs with the ARRIVE guidelines (Kilkenny et al. 2010), and the NIH (Landis et al. 2012), mirroring coordinated initiatives to improve clinical trial reporting guidelines, such as the EQUATOR network (https://www.equator-network.org). Yet despite the impressive list of journals and institutions that have officially endorsed the ARRIVE guidelines, meta-research shows that there is much to be improved in terms of compliance (Jin et al. 2018; Hair et al. 2019). Moreover, there is no obligation to publish every single study performed or to report all experiments of a study in peer-reviewed journals; an important amount, possibly as much as 50%, remain unpublished (ter Riet et al. 2012).

3 Overcoming Obstacles and Further Learning from Principles of Evidence-Based Medicine

3.1 Working Together to Improve Nonclinical Data Reliability

Many conversations among researchers, basic and clinical, resemble the one between Professor Benchie and Doctor Athena (Macleod 2015), in which Athena concludes that they should be able to improve reliability and translatability, at least a little, by learning from the strengths and weaknesses of their respective backgrounds.

Strictly following the EBM and GRADE rules would require that scientists appraise *all* the available nonclinical evidence with relevance to the question being asked. This should be the case when deciding whether to take a compound to the clinic, but is unlikely to happen for other purposes. Scientists would nevertheless benefit from a basic understanding of the methodology, strengths and weaknesses of systematic review and meta-analysis. Meta-analyses are often performed in collaborations, and a recent feasibility study using crowd-sourcing for clinical study quality assessment suggests that this could be a way forward, since experts and novices obtained the same results (Pianta et al. 2018). Combined with recently developed and highly promising machine learning algorithms (Bannach-Brown et al. 2019), collaborative efforts could increase the pace and reduce human error in systematic reviews and meta-analysis.

In recent years, private sector organizations, academic institutions, disease foundations, patient associations, and government bodies have formed consortia to tackle a wide variety of complex questions, in a precompetitive manner. Many of these partnerships bring together basic and clinical researchers and also aim to share experimental procedures and unpublished findings. Collective efforts have produced consensus recommendations, based on the critical appraisal of published and unpublished data, in fields such as stroke (Macleod et al. 2009) and pain (Knopp et al. 2015; Andrews et al. 2016). In IMI Europain (home page: www.imieuropain.org), the group of scientists and clinicians working on improving and refining animal models of chronic pain, addressing the clinical relevance of endpoints used in animal models and methodologies to reduce experimental bias, held teleconference meetings roughly 10 times a year over 5 years, which represents a substantial amount of shared data and expertise. Leveraging this combined expertise and aiming to develop a novel, non-evoked outcome measure of pain-related behavior in rodents, IMI Europain partners from both academia and industry accomplished a multicenter nonclinical study (Wodarski et al. 2016), in the spirit of a phase 3 multicenter clinical trial. One of the important lessons learned during this study was that absolute standardization should not be the goal, since circumstantial differences such as site location cannot be erased, leading to pragmatic accommodations for local variations in laboratory practice and procedures. An effort to uncover evidence-based drivers of reliability in other subfields of neuroscience is ongoing in IMI EQIPD (home page: https://quality-preclinical-data.eu), with the overarching goal of building broadly applicable tools for managing nonclinical data quality. Discussions on emerging pathways of neurodegenerative disease within the IMI neurodegeneration strategic

governance group led to a single, collectively written article describing independent attempts that failed to reproduce or extend the findings of a prominent publication (Latta-Mahieu et al. 2018). A culture of collaboration is thus growing, and not only in large consortia. Co-designing nonclinical studies is now the preferred practice in bilateral partnerships or when studies are outsourced by biopharma companies to nonclinical CROS or academic drug discovery centers.

3.2 Enhancing Capabilities, from Training to Open Access to Data

Research quality training should aim to provide the ability to recognize the different forms of bias and how to minimize risks, covering the full scope of data reliability, rather than solely focusing on compliance or on scientific integrity. In the private sector, laboratory notebook compliance audits are routinely performed; checklists are used to assess whether scientists have correctly entered information in laboratory notebooks. When releasing individual audit results to scientists, these compliance checklists or, in all sectors, the Nature Reporting Summary (https://www.nature.com/documents/nr-reporting-summary.pdf) checklist can also be used as tools for continuing training.

Initial and continuing training in statistics should be an absolute priority for all biologists. Those who are privileged to work closely with biostatisticians should aim to establish a common language, and a meaningful engagement of both parties from the start, to be able to translate the scientific question to a statistical one and co-build study designs, with the most stringent criteria for confirmatory studies.

Learning to read a paper and to critically appraise evidence and keeping in mind that low-quality reporting can confound the appraisal and that even high-profile publications may have shortcomings should also be part of training, continued in journal clubs, and carried over to post-publication peer review (e.g., PubPeer, https://pubpeer.com). Paying particular attention to the methods section and any supplementary methods information, searching for sample size considerations, randomization, and blinding, before interpreting data presented in figures, is an effective way to remember that independent evaluation of the data, with its strengths and limitations, is the core responsibility of scientists in all research endeavors.

The fact that many clinical trial findings remain unpublished is still a major roadblock for EBM, which various organizations have been tackling in recent years (see links in https://www.eupati.eu/clinical-development-and-trials/clinical-study-results-publication-and-application). In biopharma companies, proprietary nonclinical data include a considerable amount of study replicates, sometimes spread over several years. Many attempts are also made to reproduce data reported in the literature (Begley and Ellis 2012; Prinz et al. 2011; Djulbegovic and Guyatt 2017), but most of these remain undisclosed. In recent years, several independent groups have been instrumental in coordinating and publishing reproducibility studies, such as the Reproducibility Initiative collaboration between Science Exchange, PLOS, figshare, and Mendeley (http://validation.scienceexchange.com/#/reproducibility-initiative), the Center for Open Science (The Reproducibility Project, a collaborative

effort by the Center for Open Science: https://cos.io), and a unique nonprofit-driven initiative in amyotrophic lateral sclerosis (Scott et al. 2008). In sectors of the biomedical ecosystem where the focus is more on exploring new ideas, generating and testing hypotheses, or confirming and extending a team's own work rather than replicating that of others, a substantial amount of work, possibly as much as 50% (ter Riet et al. 2012), remains unpublished. Thus, in the nonclinical space, the obstacles to widespread open access to data have yet to be overcome.

4 Conclusion and Perspectives

Although the term evidence-based medicine was first introduced almost 30 years ago, building upon efforts over several decades to strengthen a data-driven practice of medicine, there are still misconceptions and resistance to the approach, as well as challenges to its practical implementation, despite a number of striking illustrations of its impact (Djulbegovic and Guyatt 2017). Adapting the conceptual toolbox of EBM and using it to optimize nonclinical research practices and decision-making will likely also require time, and most importantly, strong commitment and well-targeted, well-focused advocacy from all stakeholders. Several lessons from EBM particularly deserve the attention of nonclinical scientists, such as the importance of framing a question, critically appraising prior evidence, carefully designing a study that addresses that question, and assessing the quality of the data before moving to the next step (Fig. 1).

In medicine, reviewing the evidence aims to inform the decision about how to treat a patient; in science, the decision can be about whether or not to pursue a project, about which experiment to do next, which assay to develop, whether the work is sufficient for publication, or whether the aggregated evidence supports testing a compound in humans. In all sectors, a universal framework, with customizable tools, such as those available in the clinical setting, higher standards in data, and metadata management practices and sharing would help scientists assess and generate more reliable data.

Adapting EBM principles to nonclinical research need not undermine the freedom to explore. Assessing the quality of prior work should not paralyze scientists or prevent them from thinking out of the box, and the effective implementation of measures, such as blinding and randomization, to reduce bias should not produce a bias against novelty. Exploratory studies aiming to generate new hypotheses may follow less strict designs and statistical approaches, but when they are followed by confirmatory studies, a novel body of evidence and knowledge is formed, which can propel innovation through significance and impact. Indeed, "Innovative research projects are expected to generate data that is reproducible and provides a foundation for future studies" (http://grants.nih.gov/reproducibility/faqs.htm#4831). In other words, to be truly innovative, novel findings should endure beyond the initial excitement they create. If publications were collaboratively appraised using an adaptation of GRADE ratings, journals could develop novel impact metrics to reflect these ratings and the endurance of the findings.

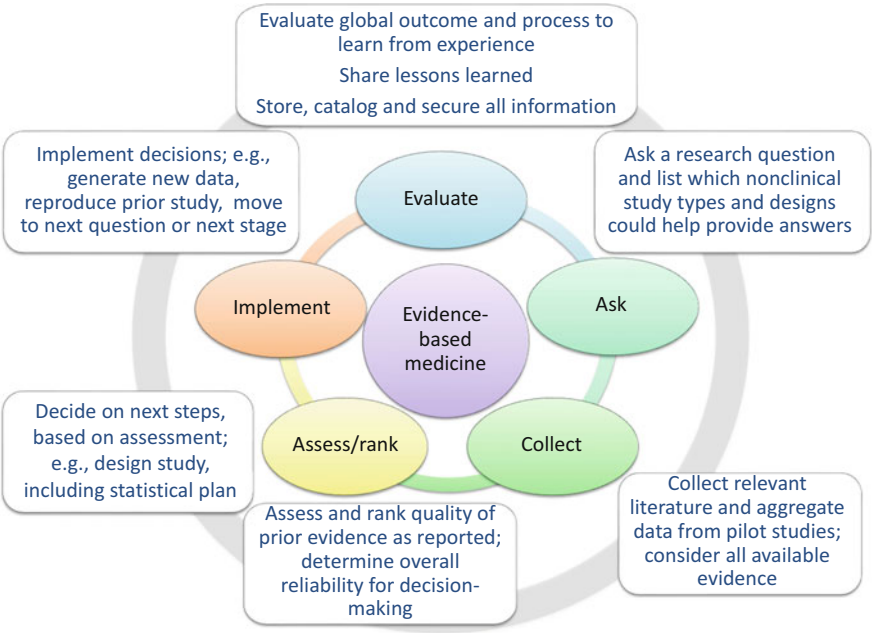

Fig. 1 Adapting the five evidence-based medicine steps to nonclinical pharmacology research

In drug discovery and development, the significance and reliability, or lack thereof, of experimental data have immediate consequences. Biopharma companies need to be able to rely on the data to determine a course of action in a research project, to shape the future of a drug discovery program, and to extrapolate doses that will be administered to humans. There is thus both a financial interest and an ethical imperative, and from the patient's perspective, an absolute requirement, to base a decision to test a compound in humans on reliable data. When the overarching goal of nonclinical pharmacology research is to bring a compound to the clinic, transitioning to an evidence-based model, using and generating evidence rated in the upper levels of the pyramid to inform decisions, would benefit discovery, and at the very least, reduce the amount of wasted experiments.

Even with high quality of evidence and better informed decision-making, it remains to be seen whether the approaches discussed here will effectively decrease the attrition rate of drug candidates and lead to more success in translating findings from nonclinical to clinical studies. There are many reasons for "failure," and only some are related to scientific rigor, reproducibility, and robustness. However, progress in understanding disease mechanisms and target tractability (https://docs.targetvalidation.org/faq/what-is-target-tractability) is linked to the ability to design experiments and clinical trials that provide reliable information. In the near future, as solutions for enhancing data access emerge and stringent reporting standards become mandatory, scientists of all sectors should be encouraged to adapt and adopt EBM principles, to better enable reliable data-driven decisions.

Acknowledgments The authors thank the members of the IMI EQIPD consortium for many thought-provoking and fruitful discussions, Catherine Deon for providing insight on liquid chromatography-coupled mass spectrometry experiments, and Stéphanie Eyquem for critically reviewing the manuscript.

References

Andrews NA, Latremoliere A, Basbaum AI et al (2016) Ensuring transparency and minimization of methodologic bias in preclinical pain research: PPRECISE considerations. Pain 157:901–909

Balshem H, Helfand M, Schünemann HJ et al (2011) GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 64(4):401–406

Bannach-Brown A, Przybyła P, Thomas J et al (2019) Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. Syst Rev 8(1):23

Begley CG, Ellis LM (2012) Drug development: raise standards for preclinical cancer research. Nature 483(7391):531–533

Button KS, Ioannidis JP, Mokrysz C et al (2013) Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14:365–376

Carter A, Tilling K, Munafò MR (2017) A systematic review of sample size and power in leading neuroscience journals. https://www.biorxiv.org/content/early/2017/11/23/217596

Djulbegovic B, Guyatt GH (2017) Progress in evidence-based medicine: a quarter century on. Lancet 390(10092):415–423

Egan KJ, Vesterinen HM, Beglopoulos V, Sena ES, Macleod MR (2016) From a mouse: systematic analysis reveals limitations of experiments testing interventions in Alzheimer's disease mouse models. Evid Based Preclin Med 3(1):e00015

Goodman SN, Fanelli D, Ioannidis JP (2016) What does research reproducibility mean? Sci Transl Med 8(341):341ps12

Hair K, Macleod MR, Sena ES, IICARus Collaboration (2019) A randomised controlled trial of an Intervention to Improve Compliance with the ARRIVE guidelines (IICARus). Res Integr Peer Rev 4:12

Hooijmans CR, Rovers MM, de Vries RBM et al (2014) SYRCLE's risk of bias tool for animal studies. BMC Med Res Methodol 14:43

Hooijmans CR, de Vries RBM, Ritskes-Hoitinga M et al (2018) GRADE Working Group. Facilitating healthcare decisions by assessing the certainty in the evidence from preclinical animal studies. PLoS One 13(1):e0187271

Jin Y, Sanger N, Shams I et al (2018) Does the medical literature remain inadequately described despite having reporting guidelines for 21 years? – A systematic review of reviews: an update. J Multidiscip Healthc 11:495–510

Kilkenny C, Browne WJ, Cuthill C et al (2010) Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. PLoS Biol 8:e1000412; ARRIVE: https://www.nc3rs.org.uk/arrive-guidelines

Kimmelman J, Mogil JS, Dirnagl U (2014) Distinguishing between exploratory and confirmatory preclinical research will improve translation. PLoS Biol 12(5):e1001863

Knopp KL, Stenfors C, Baastrup C et al (2015) Experimental design and reporting standards for improving the internal validity of pre-clinical studies in the field of pain: consensus of the IMI-Europain consortium. Scand J Pain 7(1):58–70

Landis SC, Amara SG, Asadullah K et al (2012) A call for transparent reporting to optimize the predictive value of preclinical research. Nature 490:187–191; and https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research

Langhof H, Chin WWL, Wieschowski S et al (2018) Preclinical efficacy in therapeutic area guidelines from the U.S. Food and Drug Administration and the European Medicines Agency: a cross-sectional study. Br J Pharmacol 175(22):4229–4238

Latta-Mahieu M, Elmer B, Bretteville A et al (2018) Systemic immune-checkpoint blockade with anti-PD1 antibodies does not alter cerebral amyloid-β burden in several amyloid transgenic mouse models. Glia 66(3):492–504
Macleod MR (2015) Prof Benchie and Dr Athena-a modern tragedy. Evid Based Preclin Med 2 (1):16–19
Macleod MR, Fisher M, O'Collins V et al (2009) Good laboratory practice: preventing introduction of bias at the bench. Stroke 40(3):e50–e52
Nosek BA, Ebersole CR, DeHaven AC et al (2018) The preregistration revolution. Proc Natl Acad Sci U S A 115(11):2600–2606
Peers IS, Ceuppens PR, Harbron C (2012) In search of preclinical robustness. Nat Rev Drug Discov 11(10):733–734
Pianta MJ, Makrai E, Verspoor KM et al (2018) Crowdsourcing critical appraisal of research evidence (CrowdCARE) was found to be a valid approach to assessing clinical research quality. J Clin Epidemiol 104:8–14
Prinz F, Schlange T, Asadullah K (2011) Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov 10(9):712
Scott S, Kranz JE, Cole J et al (2008) Design, power, and interpretation of studies in the standard murine model of ALS. Amyotroph Lateral Scler 9(1):4–15
Sena ES, Currie GL, McCann SK et al (2014) Systematic reviews and meta-analysis of preclinical studies: why perform them and how to appraise them critically. J Cereb Blood Flow Metab 34 (5):737–742
ter Riet G, Korevaar DA, Leenaars M et al (2012) Publication bias in laboratory animal research: a survey on magnitude, drivers, consequences and potential solutions. PLoS One 7:e43404
van der Worp HB, Howells DW, Sena ES et al (2010) Can animal models of disease reliably inform human studies? PLoS Med 7(3):e1000245
Wallach JD, Boyack KW, Ioannidis JPA (2018) Reproducible research practices, transparency, and open access data in the biomedical literature, 2015–2017. PLoS Biol 11:e2006930
Wieschowski S, Chin WWL, Federico C et al (2018) Preclinical efficacy studies in investigator brochures: do they enable risk-benefit assessment? PLoS Biol 16(4):e2004879
Wilkinson MD, Dumontier M, Aalbersberg IJ et al (2016) The FAIR guiding principles for scientific data management and stewardship. Sci Data 3:160018
Wodarski R, Delaney A, Ultenius C et al (2016) Cross-centre replication of suppressed burrowing behaviour as an ethologically relevant pain outcome measure in the rat: a prospective multicentre study. Pain 157(10):2350–2365