Journal of Neuroscience Methods xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Lacking quality in research: Is behavioral neuroscience affected more than other areas of biomedical science?

Anton Bespalov[a,b,*], Thomas Steckler[c]

[a] Partnership for Assessment and Accreditation of Scientific Practice (PAASP), Heidelberg, Germany
[b] Valdman Institute of Pharmacology, Pavlov Medical University, St. Petersburg, Russia
[c] Janssen Research & Development, Beerse, Belgium

HIGHLIGHTS

- CNS research is negatively affected by a reproducibility crisis.
- Reasons for such negative impact are essentially the same as in the other fields.
- However, consequences are much more dramatic in case of the CNS research.

ARTICLE INFO

Article history:
Received 26 March 2017
Received in revised form 18 October 2017
Accepted 19 October 2017
Available online xxx

Keywords:
CNS drug discovery
Behavioral neuroscience
Reproducibility
Research quality

ABSTRACT

There are many reasons why novel therapeutics fail in clinical trials but these failures are often attributed to lacking quality of preclinical data. These problems are not limited to any specific therapeutic area, academic or industrial research and are due in large part to several generic factors influencing research quality (e.g., related to definition of pre-specified endpoints, principles of study design and analysis, biased reporting, and lack of proper training). Yet, neuroscience drug discovery is often said to be affected more than the other fields. Within neuroscience, behavioral studies are the most blamed for being poorly designed, underpowered and mis-reported and there are indeed several factors that may be rather unique for behavioral research, such as a multitude of environmental conditions that are difficult to control and that are often not reported, ethical concerns about *in vivo* research and the pressure to reduce animal numbers, contributing to under-powered studies, and the complexity of study design and analysis, creating too much room for *post hoc* data massaging and selective reporting. Also, the blood-brain barrier as a frequently neglected complicating factor has to be considered in CNS research. The importance of these factors is increasingly recognized and urgent efforts are needed to demonstrate that behavioral methods of preclinical neuroscience research deliver results that can be as robust as with the non-behavioral methods Until this goal is achieved, behavioral neuroscience and neuroscience in general may be losing young talent, CNS drug discovery may lack the needed investment and this field may indeed be amongst the most affected by the current preclinical data quality crisis.

1. Introduction

When a clinical Phase II trial fails to meet its primary predicted endpoints, the preclinical data upon which the prediction was made is one area in the drug development chain that is often called into question. The reasons for translational failures are undoubtedly multifactorial but recent discussions increasingly focus on robustness of the preclinical data (Bespalov et al., 2016; Millan et al., 2015; Williams, 2011; Witkin, 2015). Indeed, extensive analyses in several fields such as stroke, multiple sclerosis and amyotrophic lateral sclerosis confirm that preclinical efficacy data are not always as robust as one would need to translate those findings into clinical efficacy (e.g., O'Collins et al., 2006; Perrin, 2014; Scott et al., 2008; Sena et al., 2010; Vesterinen, 2010).

These problems are not limited to any specific therapeutic area, academic or industrial research and are due to several generic factors influencing research quality (i.e. related to definition of pre-specified endpoints, principles of study design and analysis, biased reporting, and lack of proper training). These factors have been extensively reviewed elsewhere (Garner, 2014; Landis et al., 2012;

* Corresponding author at: PAASP, Am Aukopf 14/1, 69118 Heidelberg, Germany.
E-mail address: anton.bespalov@paasp.net (A. Bespalov).

https://doi.org/10.1016/j.jneumeth.2017.10.018
0165-0270/

Millan et al., 2015) and will not be discussed here since there is no reason why compliance with good research practice standards in neuroscience should be better or worse than in other fields.

Yet, it is often said that Neuroscience drug discovery is affected more than other therapeutic areas. These beliefs initially stemmed from a frustratingly low rate of positive news from clinical trials on novel CNS therapeutics. Indeed, there may be certain factors complicating CNS drug development, such as limited understanding of disease biology, resulting in difficulties to develop valid preclinical models to test novel treatment approaches, which in turn leads to higher risk of translational failures (note, however, more recent analyses indicate that this rate is low but comparable to therapeutic fields such as cardiovascular or oncology; Thomas et al., 2016).

Another reason for increased attention to neuroscience was provided by several highly publicized cases of preclinical data that turned out to be difficult to confirm in subsequent studies, such as efficacy of experimental amyotrophic lateral sclerosis therapeutics in the SOD1 mouse model (Perrin, 2014; Scott et al., 2008), efficacy of bexarotene in APP transgenic mice (e.g., Balducci et al., 2015; see also O'Hare et al., 2016), etc.

Most disturbing, however, are the concerns expressed about behavioral studies being particularly often poorly designed and underpowered, results not analyzed properly and mis-reported, and behavioral methods unavoidably suffering from lack of standardization and high variability.

The discussion below argues that behavioral studies are not at a generic risk of delivering non-robust results but explains why today's behavioral neuroscience nevertheless is the area of biomedical research that is probably the most affected by the 'reproducibility' discussion.

2. What makes behavioral research less robust?

Both behavioral and non-behavioral scientists are equally motivated to have their research results to be robust and reproducible. However, there seems to be a constellation of factors that are rather unique for behavioral research and that have a negative impact on research robustness.

2.1. Environmental conditions

Behavior is essentially a product of interaction of the organism with the environment. Therefore, environmental conditions affect behavior more than any other "readout" in biomedical research. Many of the environmental conditions are difficult to control and their impact is difficult to assess. A potential way forward could be at least to standardize and control those environmental factors known to affect behavior. However, even for those factors which can be controlled and where impact is acknowledged, excessive standardization may turn out to be counter-productive. Voelkl and Würbel (2016) argued that "because many environmental factors resist standardization between laboratories, animals *within* laboratories will be more homogeneous than animals *between* laboratories. Increasingly rigorous standardization will therefore produce results that are increasingly distinct between laboratories and hence less reproducible" (p. 510). One way to deal with this dilemma could be of course to be as transparent as possible and to at least report environmental test conditions in detail, thereby allowing readers of a paper to form an opinion about the generalizability of certain findings.

2.2. Blood-brain barrier

With very few exceptions (such as the "sink" hypothesis of anti-amyloid Aß passive immunization), CNS therapeutics need to cross the blood-brain barrier to engage the target(s) and produce desired efficacy. Thus, a mere presence of the drug in the body (or in the plasma compartment) does not guarantee that the CNS target will also be engaged. This is especially critical for single-dose studies that are surprisingly common and where one has no possibility to rely on the power of a dose-response analysis to establish a relationship between drug exposure and the measured outcome.

Relatively few novel drugs are freely soluble in water. This means that one may need to try different vehicles (some of which affect behavior on their own) to enable administration of the drug and different labs working with the same drug may use different preparation protocols (e.g., different vehicles, stirring conditions, temperature, etc.). Even the administration methods (e.g. needle gauge, injection site and technique) may affect the exposure to the drug, especially if it is not in a solution but rather a suspension and if there is a first-pass metabolism effect. Ideally, one would need to monitor the exposure (e.g., drug levels in plasma and in the target organ), at least for the critical studies. For behavioral studies, this is desirable but often difficult to implement without using satellite groups of animals.

2.3. Behavioral data for project's shiny finish

In vivo studies in general and behavioral studies in particular are often conducted at the final stages of a project that was based on a great idea, received funding, and progressed through a series of technically complicated, modern in vitro methods. Often, the only piece of evidence that separates a research team from a high-impact publication is an in vivo proof of concept. With all the time, money and efforts invested into the project, the pressure to deliver this proof of concept is mounting and the behavioral data may be published even if not robust, and findings may be over-interpreted.

Conversely, when the behavioral data do not look convincing, it does not necessarily mean that the project hypothesis failed. Instead, it may highlight problems with dedicating sufficient attention and resources to enable more robust in vivo proof of concept.

Consider an example of an interesting study where a series of genetic and biochemical experiments established a link between IL-12/IL-23 signaling and Alzheimer's disease-like pathology in mice and humans (Vom Berg et al., 2012). Amongst a number of measures, this paper also reported that intracerebroventricular delivery of a neutralizing p40-specific antibody (p40 is the common IL-12 and IL-23 subunit) reversed cognitive deficits in aged APP/PS1 mice. This conclusion was based on experiments using three cognitive tasks (discussed on pp. 1816-7): the contextual fear conditioning paradigm, novel object recognition and the Barnes maze. One may ask, however, how strong this evidence is.

Contextual fear conditioning evidence cannot be judged by the readers of the paper as the data is not shown, neither in the main paper nor in the supplementary material, because '. . .performance in the contextual fear conditioning test did not differ between p40-antibody-treated and isotype-treated APPPS1 mice'. Even though data were negative, it would have been interesting to allow readers to easily access those data in the spirit of full transparency. However, several reasons could have made this difficult, e.g., space limitations and other journal policies. In any case, there was no effect of treatment. Novel object recognition data was presented in the paper (Fig. 5a) but neither visual inspection of the data in the figure, nor statistical analysis support conclusions about 'deficits shown by APPPS1 mice treated with isotype control antibodies' (vs. corresponding WT mice) or normalization of this deficit by 'p40-antibody treatment to the levels of age-matched WT mice'. Similarly, Barnes maze data presented in Fig. 5b claim that 'the significant deficit in short-term memory retention in APPPS1 mice in the Barnes maze test was substantially ameliorated by icv treatment with p40-specific antibodies' without supporting this conclusion by data analysis. Does it mean that the behavioral proof

Please cite this article in press as: Bespalov, A., Steckler, T., Lacking quality in research: Is behavioral neuroscience affected more than other areas of biomedical science? J Neurosci Methods (2017), https://doi.org/10.1016/j.jneumeth.2017.10.018