



Meta-Analytic Methodology for Basic Research: A Practical Guide

Nicholas Mikolajewicz^{1,2} and Svetlana V. Komarova^{1,2*}

¹ Faculty of Dentistry, McGill University, Montreal, QC, Canada, ² Shriners Hospital for Children-Canada, Montreal, QC, Canada

OPEN ACCESS

Edited by:

Christopher Basciano,
Becton Dickinson, United States

Reviewed by:

Tomas Drgon,
United States Food and Drug
Administration, United States
Ulf Knothe,
TissuTex Pty Ltd., Wentworth Falls,
NSW, Australia

*Correspondence:

Svetlana V. Komarova
svetlana.komarova@mcgill.ca

Specialty section:

This article was submitted to
Computational Physiology and
Medicine,
a section of the journal
Frontiers in Physiology

Received: 22 December 2017

Accepted: 15 February 2019

Published: 27 March 2019

Citation:

Mikolajewicz N and Komarova SV
(2019) Meta-Analytic Methodology for
Basic Research: A Practical Guide.
Front. Physiol. 10:203.
doi: 10.3389/fphys.2019.00203

Basic life science literature is rich with information, however methodically quantitative attempts to organize this information are rare. Unlike clinical research, where consolidation efforts are facilitated by systematic review and meta-analysis, the basic sciences seldom use such rigorous quantitative methods. The goal of this study is to present a brief theoretical foundation, computational resources and workflow outline along with a working example for performing systematic or rapid reviews of basic research followed by meta-analysis. Conventional meta-analytic techniques are extended to accommodate methods and practices found in basic research. Emphasis is placed on handling heterogeneity that is inherently prevalent in studies that use diverse experimental designs and models. We introduce *MetaLab*, a meta-analytic toolbox developed in MATLAB R2016b which implements the methods described in this methodology and is provided for researchers and statisticians at Git repository (<https://github.com/NMikolajewicz/MetaLab>). Through the course of the manuscript, a rapid review of intracellular ATP concentrations in osteoblasts is used as an example to demonstrate workflow, intermediate and final outcomes of basic research meta-analyses. In addition, the features pertaining to larger datasets are illustrated with a systematic review of mechanically-stimulated ATP release kinetics in mammalian cells. We discuss the criteria required to ensure outcome validity, as well as exploratory methods to identify influential experimental and biological factors. Thus, meta-analyses provide informed estimates for biological outcomes and the range of their variability, which are critical for the hypothesis generation and evidence-driven design of translational studies, as well as development of computational models.

Keywords: meta-analysis, basic research, rapid review, systematic review, MATLAB, methodology

INTRODUCTION

Evidence-based medical practice aims to consolidate best research evidence with clinical and patient expertise. Systematic reviews and meta-analyses are essential tools for synthesizing evidence needed to inform clinical decision making and policy. Systematic reviews summarize available literature using specific search parameters followed by critical appraisal and logical synthesis of multiple primary studies (Gopalakrishnan and Ganeshkumar, 2013). Meta-analysis refers to the statistical analysis of the data from independent primary studies focused on the same question, which aims to generate a quantitative estimate of the studied phenomenon, for example, the effectiveness of the intervention (Gopalakrishnan and Ganeshkumar, 2013). In clinical research,

systematic reviews and meta-analyses are a critical part of evidence-based medicine. However, in basic science, attempts to evaluate prior literature in such rigorous and quantitative manner are rare, and narrative reviews are prevalent. The goal of this manuscript is to provide a brief theoretical foundation, computational resources and workflow outline for performing a systematic or rapid review followed by a meta-analysis of basic research studies.

Meta-analyses can be a challenging undertaking, requiring tedious screening and statistical understanding. There are several guides available that outline how to undertake a meta-analysis in clinical research (Higgins and Green, 2011). Software packages supporting clinical meta-analyses include the Excel plugins MetaXL (Barendregt and Doi, 2009) and Mix 2.0 (Bax, 2016), Revman (Cochrane Collaboration, 2011), Comprehensive Meta-Analysis Software [CMA (Borenstein et al., 2005)], JASP (JASP Team, 2018) and MetaFOR library for R (Viechtbauer, 2010). While these packages can be adapted to basic science projects, difficulties may arise due to specific features of basic science studies, such as large and complex datasets and heterogeneity in experimental methodology. To address these limitations, we developed a software package aimed to facilitate meta-analyses of basic research, *MetaLab* in MATLAB R2016b, with an intuitive graphical interface that permits users with limited statistical and coding background to proceed with a meta-analytic project. We organized *MetaLab* into six modules (Figure 1), each focused on different stages of the meta-analytic process, including graphical-data extraction, model parameter estimation, quantification and exploration of heterogeneity, data-synthesis, and meta-regression.

In the present manuscript, we describe each step of the meta-analytic process with emphasis on specific considerations made when conducting a review of basic research. The complete workflow of parameter estimation using *MetaLab* is demonstrated for evaluation of intracellular ATP content in osteoblasts (OB [ATP]_{ic} dataset) based on a rapid literature review. In addition, the features pertaining to larger datasets are explored with the ATP release kinetics from mechanically-stimulated mammalian cells (ATP release dataset) obtained as a result of a systematic review in our prior work (Mikolajewicz et al., 2018).

MetaLab can be freely accessed at Git repository (<https://github.com/NMikolajewicz/MetaLab>), and a detailed documentation of how to use *MetaLab* together with a working example is available in the **Supporting materials**.

VALIDITY OF EVIDENCE IN THE BASIC SCIENCES

To evaluate the translational potential of basic research, the validity of evidence must first be assessed, usually by examining the approach taken to collect and evaluate the data. Studies in the basic sciences are broadly grouped as hypothesis-generating and hypothesis-driven. The former tend to be small-sampled proof-of-principle studies and are typically exploratory and less valid than the latter. An argument can even be made that studies that report novel findings fall into this group as well, since their

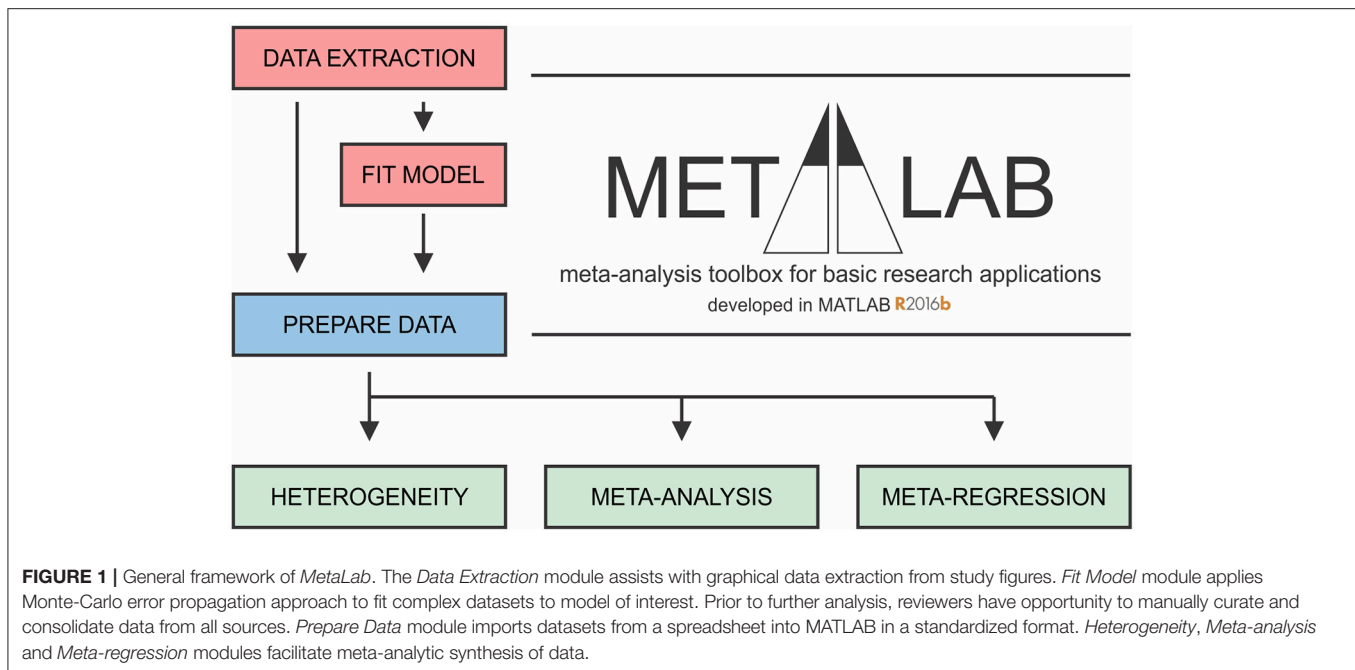
findings remain subject to external validation prior to being accepted by the broader scientific community. Alternatively, hypothesis-driven studies build upon what is known or strongly suggested by earlier work. These studies can also validate prior experimental findings with incremental contributions. Although such studies are often overlooked and even dismissed due to a lack of substantial novelty, their role in external validation of prior work is critical for establishing the translational potential of findings.

Another dimension to the validity of evidence in the basic sciences is the selection of experimental model. The human condition is near-impossible to recapitulate in a laboratory setting, therefore experimental models (e.g., cell lines, primary cells, animal models) are used to mimic the phenomenon of interest, albeit imperfectly. For these reasons, the best quality evidence comes from evaluating the performance of several independent experimental models. This is accomplished through systematic approaches that consolidate evidence from multiple studies, thereby filtering the signal from the noise and allowing for side-by-side comparison. While systematic reviews can be conducted to accomplish a qualitative comparison, meta-analytic approaches employ statistical methods which enable hypothesis generation and testing. When a meta-analysis in the basic sciences is hypothesis-driven, it can be used to evaluate the translational potential of a given outcome and provide recommendations for subsequent translational- and clinical-studies. Alternatively, if meta-analytic hypothesis testing is inconclusive, or exploratory analyses are conducted to examine sources of inconsistency between studies, novel hypotheses can be generated, and subsequently tested experimentally. Figure 2 summarizes this proposed framework.

STEPS IN QUANTITATIVE LITERATURE REVIEW

All meta-analytic efforts prescribe to a similar workflow, outlined as follows:

- 1) **Formulate research question**
 - Define primary and secondary objectives
 - Determine breadth of question
- 2) **Identify relevant literature**
 - Construct search strategy: rapid or systematic search
 - Screen studies and determine eligibility
- 3) **Extract and consolidate study-level data**
 - Extract data from relevant studies
 - Collect relevant study-level characteristics and experimental covariates
 - Evaluate quality of studies
 - Estimate model parameters for complex relationships (optional)
- 4) **Data appraisal and preparation**
 - Compute appropriate outcome measure



- Evaluate extent of between-study inconsistency (heterogeneity)
 - Perform relevant data transformations
 - Select meta-analytic model
- 5) **Synthesize study-level data into summary measure**
- Pool data and calculate summary measure and confidence interval
- 6) **Exploratory analyses**
- Explore potential sources of heterogeneity (ex. biological or experimental)
 - Subgroup and meta-regression analyses
- 7) **Knowledge synthesis**
- Interpret findings
 - Provide recommendations for future work

META-ANALYSIS METHODOLOGY

Search and Selection Strategies

The first stage of any review involves formulating a primary objective in the form of a research question or hypothesis. Reviewers must explicitly define the objective of the review before starting the project, which serves to reduce the risk of data dredging, where reviewers later assign meaning to significant findings. Secondary objectives may also be defined; however, precaution must be taken as the search strategies formulated for the primary objective may not entirely encompass the body of work required to address the secondary objective. Depending on the purpose of a review, reviewers may choose to undertake a rapid or systematic review. While the meta-analytic methodology is similar for systematic and rapid reviews, the scope of literature

assessed tends to be significantly narrower for rapid reviews permitting the project to proceed faster.

Systematic Review and Meta-Analysis

Systematic reviews involve comprehensive search strategies that enable reviewers to identify all relevant studies on a defined topic (DeLuca et al., 2008). Meta-analytic methods then permit reviewers to quantitatively appraise and synthesize outcomes across studies to obtain information on statistical significance and relevance. Systematic reviews of basic research data have the potential of producing information-rich databases which allow extensive secondary analysis. To comprehensively examine the pool of available information, search criteria must be sensitive enough not to miss relevant studies. Key terms and concepts that are expressed as synonymous keywords and index terms, such as Medical Subject Headings (MeSH), must be combined using Boolean operators AND, OR and NOT (Ecker and Skelly, 2010). Truncations, wildcards, and proximity operators can also help refine a search strategy by including spelling variations and different wordings of the same concept (Ecker and Skelly, 2010). Search strategies can be validated using a selection of expected relevant studies. If the search strategy fails to retrieve even one of the selected studies, the search strategy requires further optimization. This process is iterated, updating the search strategy in each iterative step until the search strategy performs at a satisfactory level (Finfgeld-Connett and Johnson, 2013). A comprehensive search is expected to return a large number of studies, many of which are not relevant to the topic, commonly resulting in a specificity of <10% (McGowan and Sampson, 2005). Therefore, the initial stage of sifting through the library to select relevant studies is time-consuming (may take 6 months to 2 years) and prone to human error. At this stage, it is recommended to include at least two independent reviewers to minimize

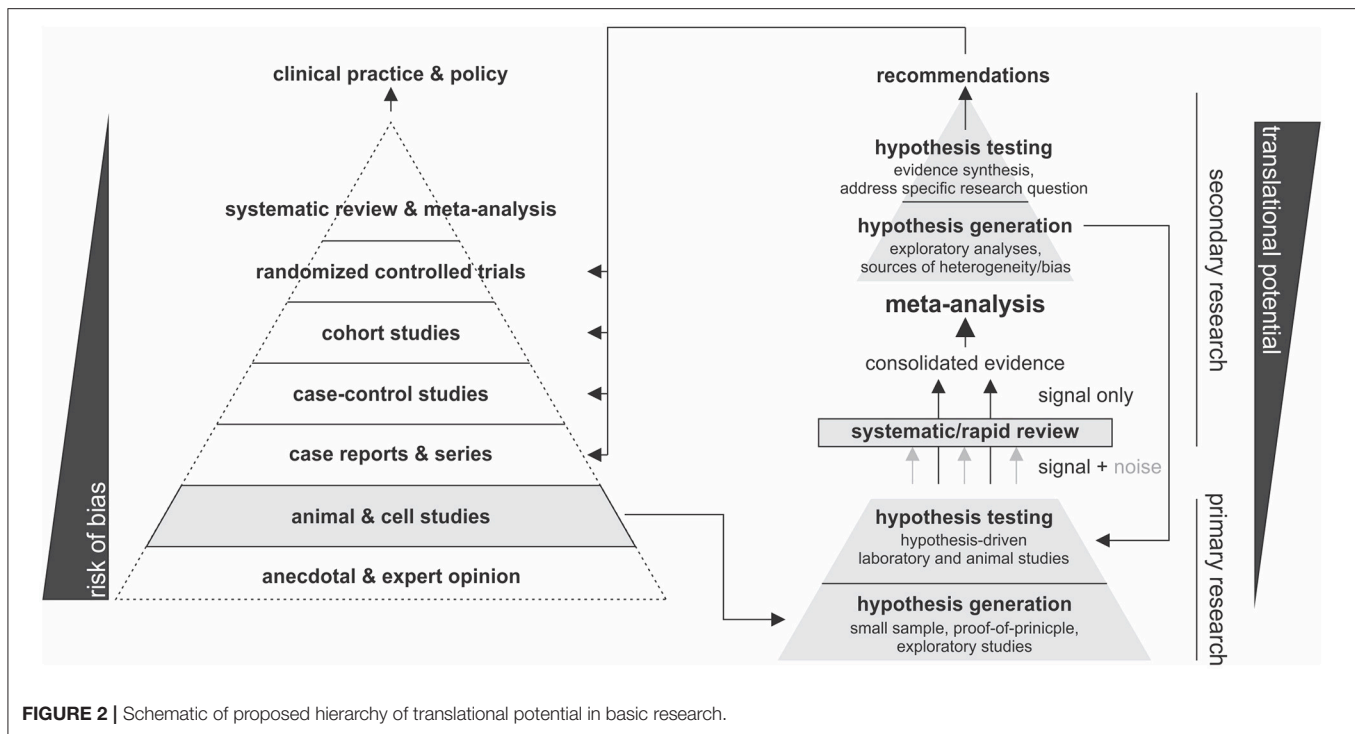


FIGURE 2 | Schematic of proposed hierarchy of translational potential in basic research.

selection bias and related errors. Nevertheless, systematic reviews have a potential to provide the highest quality quantitative evidence synthesis to directly inform the experimental and computational basic, preclinical and translational studies.

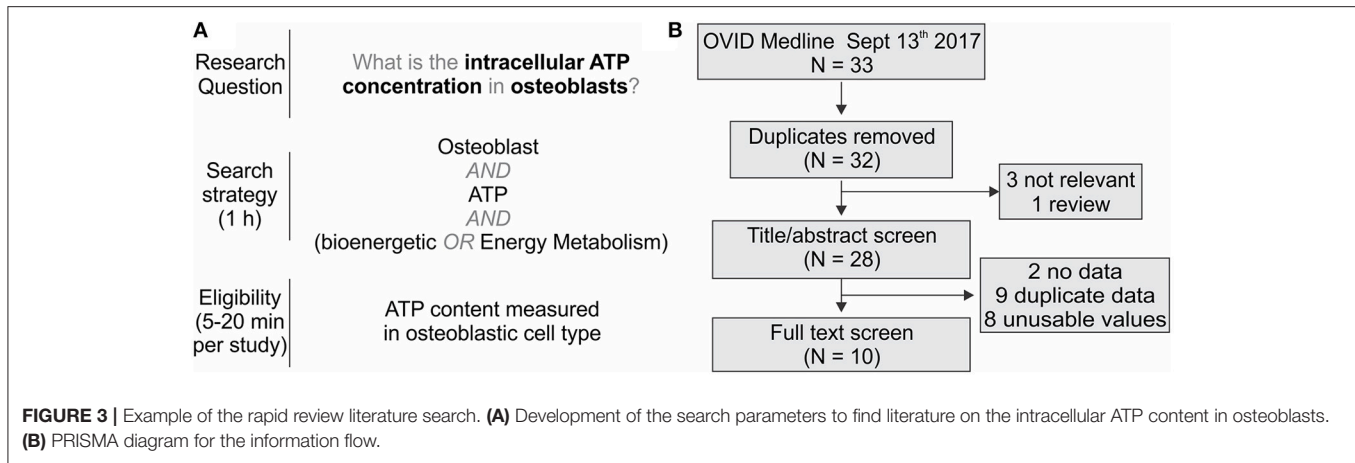
Rapid Review and Meta-Analysis

The goal of the rapid review, as the name implies, is to decrease the time needed to synthesize information. Rapid reviews are a suitable alternative to systematic approaches if reviewers prefer to get a general idea of the state of the field without an extensive time investment. Search strategies are constructed by increasing search specificity, thus reducing the number of irrelevant studies identified by the search at the expense of search comprehensiveness (Haby et al., 2016). The strength of a rapid review is in its flexibility to adapt to the needs of the reviewer, resulting in a lack of standardized methodology (Mattivi and Buchberger, 2016). Common shortcuts made in rapid reviews are: (i) narrowing search criteria, (ii) imposing date restrictions, (iii) conducting the review with a single reviewer, (iv) omitting expert consultation (i.e., librarian for search strategy development), (v) narrowing language criteria (ex. English only), (vi) foregoing the iterative process of searching and search term selection, (vii) omitting quality checklist criteria and (viii) limiting number of databases searched (Ganann et al., 2010). These shortcuts will limit the initial pool of studies returned from the search, thus expediting the selection process, but also potentially resulting in the exclusion of relevant studies and introduction of selection bias. While there is a consensus that rapid reviews do not sacrifice quality, or synthesize misrepresentative results (Haby et al., 2016), it is recommended that critical outcomes be

later verified by systematic review (Ganann et al., 2010). Nevertheless, rapid reviews are a viable alternative when parameters for computational modeling need to be estimated. While systematic and rapid reviews rely on different strategies to select the relevant studies, the statistical methods used to synthesize data from the systematic and rapid review are identical.

Screening and Selection

When the literature search is complete (the date articles were retrieved from the databases needs to be recorded), articles are extracted and stored in a reference manager for screening. Before study screening, the inclusion and exclusion criteria must be defined to ensure consistency in study identification and retrieval, especially when multiple reviewers are involved. The critical steps in screening and selection are (1) removing duplicates, (2) screening for relevant studies by title and abstract, and (3) inspecting full texts to ensure they fulfill the eligibility criteria. There are several reference managers available including Mendeley and Rayyan, specifically developed to assist with screening systematic reviews. However, 98% of authors report using Endnote, Reference Manager or RefWorks to prepare their reviews (Lorenzetti and Ghali, 2013). Reference managers often have deduplication functions; however, these can be tedious and error-prone (Kwon et al., 2015). A protocol for faster and more reliable de-duplication in Endnote has been recently proposed (Bramer et al., 2016). The selection of articles should be sufficiently broad not to be dominated by a single lab or author. In basic research articles, it is common to find data sets that are reused by the same group in multiple studies. Therefore, additional precautions should be taken when deciding



to include multiple studies published by a single group. At the end of the search, screening and selection process, the reviewer obtains a complete list of eligible full-text manuscripts. The entire screening and selection process should be reported in a PRISMA diagram, which maps the flow of information throughout the review according to prescribed guidelines published elsewhere (Moher et al., 2009). **Figure 3** provides a summary of the workflow of search and selection strategies using the OB [ATP]_{ic} rapid review and meta-analysis as an example.

Data Extraction, Initial Appraisal, and Preparation

Identification of Parameters to be Extracted

It is advised to predefine analytic strategies before data extraction and analysis. However, the availability of reported effect measures and study designs will often influence this decision. When reviewers aim to estimate the absolute mean difference (absolute effect), normalized mean difference, response ratio or standardized mean difference (ex. Hedges' g), they need to extract study-level means (θ_i), standard deviations ($sd(\theta_i)$), and sample sizes (n_i), for control (denoted θ_i^c , $sd(\theta_i^c)$, and n_i^c) and intervention (denoted θ_i^r , $sd(\theta_i^r)$, and n_i^r) groups, for studies i . To estimate absolute mean effect, only the mean (θ_i^r), standard deviation ($sd(\theta_i^r)$), and sample size (n_i^r) are required. In basic research, it is common for a single study to present variations of the same observation (ex. measurements of the same entity using different techniques). In such cases, each point may be treated as an individual observation, or common outcomes within a study can be pooled by taking the mean weighted by the sample size. Another consideration is inconsistency between effect size units reported on the absolute scale, for example, protein concentrations can be reported as g/cell, mol/cell, g/g wet tissue or g/g dry tissue. In such cases, conversion to a common representation is required for comparison across studies, for which appropriate experimental parameters and calibrations need to be extracted from the studies. While some parameters can be approximated by reviewers, such as cell-related parameters found in BioNumbers database (Milo

et al., 2010) and equipment-related parameters presumed from manufacturer manuals, reviewers should exercise caution when making such approximations as they can introduce systematic errors that manifest throughout the analysis. When data conversion is judged to be difficult but negative/basal controls are available, scale-free measures (i.e., normalized, standardized, or ratio effects) can still be used in the meta-analysis without the need to convert effects to common units on the absolute scale. In many cases, reviewers may only be able to decide on a suitable effect size measure after data extraction is complete.

It is regrettably common to encounter unclear or incomplete reporting, especially for the sample sizes and uncertainties. Reviewers may choose to reject studies with such problems due to quality concerns or to employ conservative assumptions to estimate missing data. For example, if it is unclear if a study reports the standard deviation or standard error of the mean, it can be assumed to be a standard error, which provides a more conservative estimate. If a study does not report uncertainties but is deemed important because it focuses on a rare phenomenon, imputation methods have been proposed to estimate uncertainty terms (Chowdhry et al., 2016). If a study reports a range of sample sizes, reviewers should extract the lowest value. Strategies to handle missing data should be pre-defined and thoroughly documented.

In addition to identifying relevant primary parameters, *a priori* defined study-level characteristics that have a potential to influence the outcome, such as species, cell type, specific methodology, should be identified and collected in parallel to data extraction. This information is valuable in subsequent exploratory analyses and can provide insight into influential factors through between-study comparison.

Quality Assessment

Formal quality assessment allows the reviewer to appraise the quality of identified studies and to make informed and methodical decision regarding exclusion of poorly conducted studies. In general, based on initial evaluation of full texts, each

study is scored to reflect the study's overall quality and scientific rigor. Several quality-related characteristics have been described (Sena et al., 2007), such as: (i) published in peer-reviewed journal, (ii) complete statistical reporting, (iii) randomization of treatment or control, (iv) blinded analysis, (v) sample size calculation prior to the experiment, (vi) investigation of a dose-response relationship, and (vii) statement of compliance with regulatory requirements. We also suggest that the reviewers of basic research studies assess (viii) objective alignment between the study in question and the meta-analytic project. This involves noting if the outcome of interest was the primary study objective or was reported as a supporting or secondary outcome, which may not receive the same experimental rigor and is subject to expectation bias (Sheldrake, 1997). Additional quality criteria specific to experimental design may be included at the discretion of the reviewer. Once study scores have been assembled, study-level aggregate quality scores are determined by summing the number of satisfied criteria, and then evaluating how outcome estimates and heterogeneity vary with study quality. Significant variation arising from poorer quality studies may justify study omission in subsequent analysis.

Extraction of Tabular and Graphical Data

The next step is to compile the meta-analytic data set, which reviewers will use in subsequent analysis. For each study, the complete dataset which includes parameters required to estimate the target outcome, study characteristics, as well as data necessary for unit conversion needs to be extracted. Data reporting in basic research are commonly tabular or graphical. Reviewers can accurately extract tabular data from the text or tables. However, graphical data often must be extracted from the graph directly using time consuming and error prone methods. The Data Extraction Module in *MetaLab* was developed to facilitate systematic and unbiased data extraction; Reviewers provide study figures as inputs, then specify the reference points that are used to calibrate the axes and extract the data (Figures 4A,B).

To validate the performance of the *MetaLab* Data Extraction Module, we generated figures using 319 synthetic data points plotted with varying markers sizes (Figure 4C). Extracted and actual values were correlated ($R^2 = 0.99$) with the relationship slope estimated as 1.00 (95% CI: 0.99 to 1.01) (Figure 4D). Bias was absent, with a mean percent error of 0.00% (95% CI: -0.02 to 0.02%) (Figure 4E). The narrow range of errors between -2.00 and 1.37%, and consistency between the median and mean error indicated no skewness. Data marker size did not contribute to the extraction error, as 0.00% of the variation in absolute error was explained by marker size, and the slope of the relationship between marker size and extraction error was 0.000 (95% CI: -0.001, 0.002) (Figure 4F). These data demonstrate that graphical data can be reliably extracted using *MetaLab*.

Extracting Data From Complex Relationships

Basic science often focuses on natural processes and phenomena characterized by complex relationships between a series of inputs (e.g., exposures) and outputs (e.g., response). The results are commonly explained by an accepted model of the relationship, such as Michaelis-Menten model of enzyme kinetics which

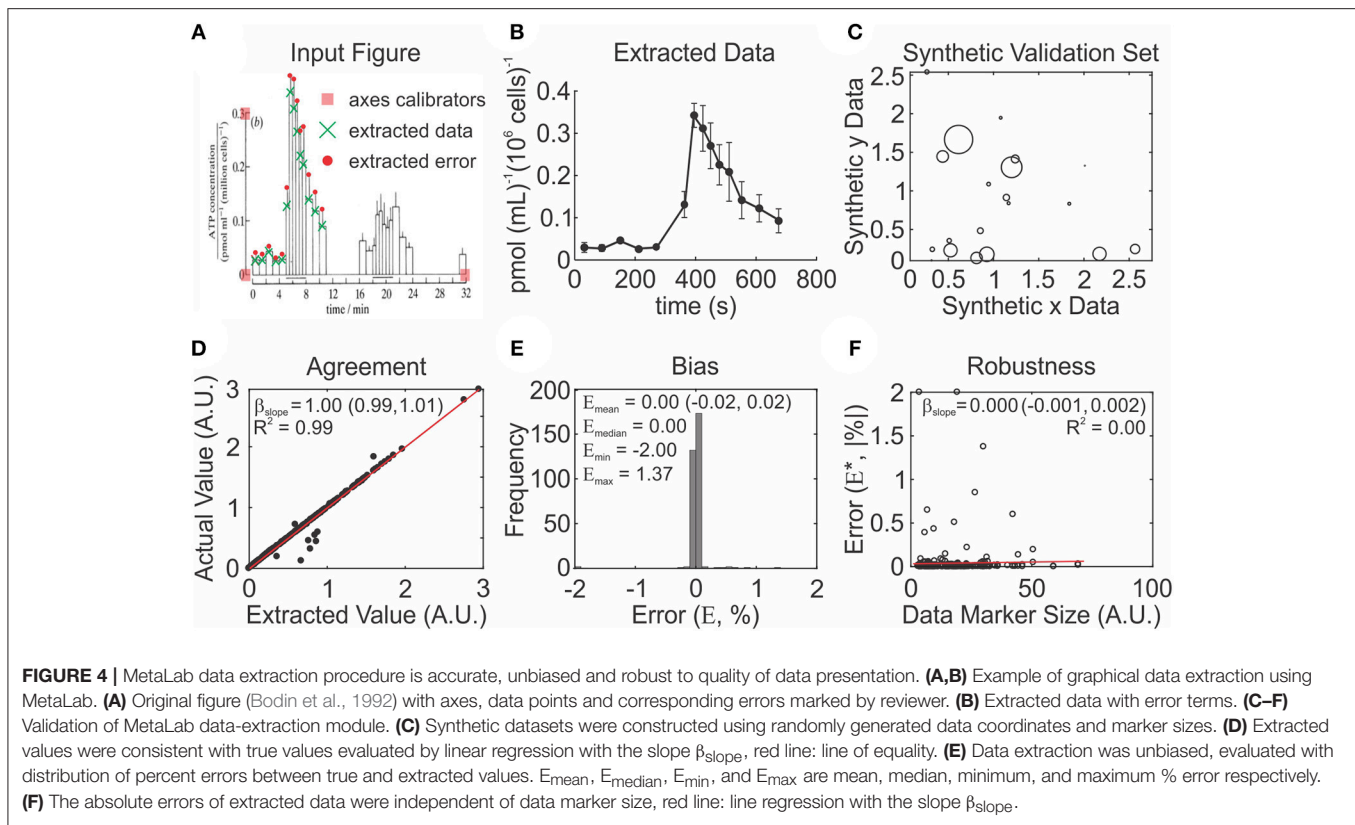
TABLE 1 | Commonly used models of complex relationships in basic sciences.

Model	Equation	Parameter meaning	Applications
Linear model	$y = \beta_1 x + \beta_2$	β_1 : slope, magnitude of relationship β_2 : intercept, response at $x = 0$	Reaction rates
Quadratic model (vertex form)	$y = \beta_1 (x - \beta_2)^2 + \beta_3$	β_1 : curvature factor β_2 : x at global max/min β_3 : global maxima/minimal	Trajectory modeling
Exponential model	$y = \beta_1 e^{\beta_2 x}$	β_1 : intercept, response at $x = 0$ β_2 : decay/growth constant	Population decay/growth
Michaelis-Menten, hyperbolic curve	$y = \frac{\beta_1 x}{\beta_2 + x}$	β_1 : max response β_2 : x at half max response	Enzyme kinetics, reaction rates, infection rates, drug clearance
Sigmoidal Emax Model, Hill Function	$y = \frac{\beta_1 x^{\beta_3}}{(\beta_2)^{\beta_3} + x^{\beta_3}}$	β_1 : max response β_2 : x at half max response β_3 : slope-related term	Dose-response relationships, pharmacodynamics

involves two parameters— V_{\max} for the maximum rate and K_m for the substrate concentration half of V_{\max} . For meta-analysis, model parameters characterizing complex relationships are of interest as they allow direct comparison of different multi-observational datasets. However, study-level outcomes for complex relationships often (i) lack consistency in reporting, and (ii) lack estimates of uncertainties for model parameters. Therefore, reviewers wishing to perform a meta-analysis of complex relationships may need to fit study-level data to a unified model $y = f(x, \beta)$ to estimate parameter set β characterizing the relationship (Table 1), and assess the uncertainty in β .

The study-level data can be fitted to a model using conventional fitting methods, in which the model parameter error terms depend on the goodness of fit and number of available observations. Alternatively, a Monte Carlo simulation approach (Cox et al., 2003) allows for the propagation of study-level variances (uncertainty in the model inputs) to the uncertainty in the model parameter estimates (Figure 5). Suppose that study i reported a set of k predictor variables $x = \{x_j | 1 \leq j \leq k\}$ for a set of outcomes $\theta = \{\theta_j | 1 \leq j \leq k\}$, and that there is a corresponding set of standard deviations $sd(\theta) = \{sd(\theta_j) | 1 \leq j \leq k\}$ and sample sizes $n = \{n_j | 1 \leq j \leq k\}$ (Figure 5A). The Monte Carlo error propagation method assumes that outcomes are normally distributed, enabling pseudo random observations to be sampled from a distribution approximated by $\mathcal{N}(\theta_j, sd(\theta_j)^2)$. The pseudo random observations are then averaged to obtain a Monte-Carlo estimate θ_j^* for each observation such that

$$\theta_j^* = \frac{1}{n_j} \sum_{m=1}^{n_j} (\theta_{j,m}^*) \quad (1)$$



where $\theta_j(m)^*$ represents a pseudo-random variable sampled n_j times from $\mathcal{N}(\theta_j, sd(\theta_j)^2)$. The relationship between x and $\theta^* = \{\theta_j^* | 1 \leq j \leq k\}$ is then fitted with the model of interest using the least-squares method to obtain an estimate of model parameters β (Figure 5B). After many iterations of resampling and fitting, a distribution of parameter estimates $\mathcal{N}(\bar{\beta}, sd(\bar{\beta})^2)$ is obtained, from which the parameter means $\bar{\beta}$ and variances $sd(\bar{\beta})^2$ can be estimated (Figures 5C,D). As the number of iterations M tend to infinity, the parameter estimate converges to the expected value $E(\beta)$.

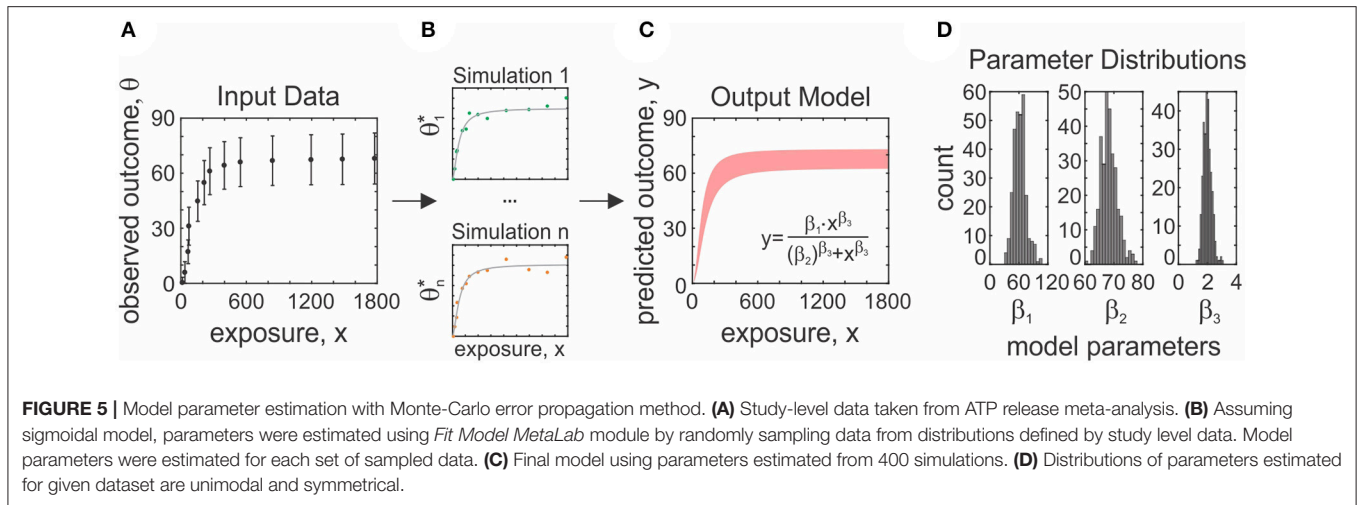
$$\lim_{M \rightarrow \infty} \frac{1}{M} (\beta_1 + \beta_2 + \dots + \beta_M) = E(\beta) \quad (2)$$

It is critical for reviewers to ensure the data is consistent with the model such that the estimated parameters sufficiently capture the information conveyed in the underlying study-level data. In general, reliable model fittings are characterized by normal parameter distributions (Figure 5D) and have a high goodness of fit as quantified by R^2 . The advantage of using the Monte-Carlo approach is that it works as a black box procedure that does not require complex error propagation formulas, thus allowing handling of correlated and independent parameters without additional consideration.

Study-Level Effect Sizes

Depending on the purpose of the review product, study-level outcomes θ_i can be expressed as one of several effect

size measures. The absolute effect size, computed as a mean outcome or absolute difference from baseline, is the simplest, is independent of variance, and retains information about the context of the data (Baguley, 2009). However, the use of absolute effect size requires authors to report on a common scale or provide conversion parameters. In cases where a common scale is difficult to establish, a scale-free measure, such as standardized, normalized or relative measures can be used. Standardized mean differences, such as Hedges' g or Cohen d , report the outcome as the size of the effect (difference between the means of experimental and control groups) relative to the overall variance (pooled and weighted standard deviation of combined experimental and control groups). The standardized mean difference, in addition to odds or risk ratios, is widely used in meta-analysis of clinical studies (Vesterinen et al., 2014), since it allows to summarize metrics that do not have unified meaning (e.g., a pain score), and takes into account the variability in the samples. However, the standardized measure is rarely used in basic science since study outcomes are commonly a defined measure, sample sizes are small, and variances are highly influenced by experimental and biological factors. Other measures that are more suited for basic science are the normalized mean difference, which expresses the difference between the outcome and baseline as a proportion of the baseline (alternatively called the percentage difference), and response ratio, which reports the outcome as a proportion of the baseline. All discussed measures have been included in *MetaLab* (Table 2).



Data Synthesis

The goal of any meta-analysis is to provide an outcome estimate that is representative of all study-level findings. One important feature of the meta-analysis is its ability to incorporate information about the quality and reliability of the primary studies by weighing larger, better reported studies more heavily. The two quantities of interest are the overall estimate and the measure of the variability in this estimate. Study-level outcomes θ_i are synthesized as a weighted mean $\hat{\theta}$ according to the study-level weights w_i :

$$\hat{\theta} = \frac{\sum_i^N (\theta_i \cdot w_i)}{\sum_i (w_i)} \quad (3)$$

where N is number of studies or datasets. The choice of a weighting scheme dictates how study-level variances are pooled to estimate the variance of the weighted mean. The weighting scheme thus significantly influences the outcome of meta-analysis, and if poorly chosen, potentially risks over-weighting less precise studies and generating a less valid, non-generalizable outcome. Thus, the notion of defining an *a priori* analysis protocol has to be balanced with the need to assure that the dataset is compatible with the chosen analytic strategy, which may be uncertain prior to data extraction. We provide strategies to compute and compare different study-level and global outcomes and their variances.

Weighting Schemes

To generate valid estimates of cumulative knowledge, studies are weighed according to their reliability. This conceptual framework, however, deteriorates if reported measures of precision are themselves flawed. The most commonly used measure of precision is the inverse variance which is a composite measure of total variance and sample size, such that studies with larger sample sizes and lower experimental errors are more reliable and more heavily weighed. Inverse variance weighting schemes are valid when (i) sampling error is random, (ii) the reported effects are homoscedastic, i.e., have equal variance and (iii) the sample size reflects the number of independent

experimental observations. When assumptions (i) or (ii) are violated, sample size weighing can be used as an alternative. Despite sample size and sample variance being such critical parameters in the estimation of the global outcome, they are often prone to deficient reporting practices.

Potential problems with sample variance and sample size

The standard error $se(\theta_i)$ is required to compute inverse variance weights, however, primary literature as well as meta-analysis reviewers often confuse standard errors with standard deviations $sd(\theta_i)$ (Altman and Bland, 2005). Additionally, many assays used in basic research often have uneven error distributions, such that the variance component arising from experimental error depends on the magnitude of the effect (Bittker and Ross, 2016). Such uneven error distributions will lead to biased weighing that does not reflect true precision in measurement. Fortunately, the standard error and standard deviation have characteristic properties that can be assessed by the reviewer to determine whether inverse variance weights are appropriate for a given dataset. The study-level standard error $se(\theta_i)$ is a measure of precision and is estimated as the product of the sample standard deviation $sd(\theta_i)$ and margin of error $\frac{1}{\sqrt{n_i}}$ for study i . Therefore, the standard error is expected to be approximately inversely proportionate to the root of the study-level sample size n_i

$$se(\theta_i) \sim \frac{1}{\sqrt{n_i}} \quad (4)$$

Unlike the standard error, the standard deviation—a measure of the variance of a random variable $sd(\theta)^2$ —is assumed to be independent of the sample size because it is a descriptive statistic rather than a precision statistic. Since the total observed study-level sample variance is the sum of natural variability (assumed to be constant for a phenomenon) and random error, no relationship is expected between reported standard deviations and sample sizes. These assumptions can be tested by correlation analysis and can be used to inform the reviewer about the reliability of the study-level uncertainty measures. For example,

TABLE 2 | Types of effect sizes.

Measure	Mean	Standard error
Absolute	$\theta_i = \begin{cases} \theta_i^r - \theta_i^c, & \text{if } \theta_i^c \text{ reported} \\ \theta_i^r, & \text{else} \end{cases}$	$se(\theta_i) = \begin{cases} \sqrt{\frac{n_i^c + n_i^r}{n_i^c n_i^r}} sd(\theta_i)^2, & \text{if } \theta_i^c \text{ reported} \\ \frac{sd(\theta_i^r)}{\sqrt{n_i^r}}, & \text{else} \end{cases}$ <p>Where $sd(\theta_i) = \sqrt{\frac{(n_i^c - 1)sd(\theta_i^c) + (n_i^r - 1)sd(\theta_i^r)}{n_i^c + n_i^r - 2}}$</p>
Standardized (Hedges' g)	$\theta_i = \frac{\theta_i^r - \theta_i^c}{sd(\theta_i)} \cdot \left(1 - \frac{3}{4(n_i^c + n_i^r) - 9}\right)$ <p>Where $sd(\theta_i) = \sqrt{\frac{(n_i^c - 1)sd(\theta_i^c) + (n_i^r - 1)sd(\theta_i^r)}{n_i^c + n_i^r - 2}}$</p>	$se(\theta_i) = \sqrt{\frac{n_i^c + n_i^r}{n_i^c n_i^r} + \frac{\theta_i^2}{2((n_i^c + n_i^r) - 3.94)}}$
Normalized	$\theta_i = \frac{\theta_i^r - \theta_i^c}{\theta_i^c}$	$se(\theta_i) = \sqrt{\frac{\left(\frac{sd(\theta_i^c)}{\theta_i^c}\right)^2}{n_i^c} + \frac{\left(\frac{sd(\theta_i^r)}{\theta_i^r}\right)^2}{n_i^r}}$
Ratio	$\theta_i = \frac{\theta_i^r}{\theta_i^c}$	$se(\theta_i) = \sqrt{\frac{(\theta_i^r)^2}{(\theta_i^c)^2} \left(\frac{sd(\theta_i^r)^2}{n_i^r (\theta_i^r)^2} + \frac{sd(\theta_i^c)^2}{n_i^c (\theta_i^c)^2} \right)}$

Provided are formulas to calculate the mean and standard error for the specified effect sizes.

a relationship between sample size and sample variance was observed for the OB [ATP]_{ic} dataset (**Figure 6A**), but not for the ATP release data (**Figure 6B**). Therefore, in the case of the OB [ATP]_{ic} data set, lower variances are not associated with higher precision and inverse variance weighting is not appropriate. Sample sizes are also frequently misrepresented in the basic sciences, as experimental replicates and repeated experiments are often reported interchangeably (incorrectly) as sample sizes (Vaux et al., 2012). Repeated (independent) experiments refer to number of randomly sampled observations, while replicates refer to the repeated measurement of a sample from one experiment to improve measurement precision. Statistical inference theory assumes random sampling, which is satisfied by independent experiments but not by replicate measurements. Misrepresentative reporting of replicates as the sample size may artificially inflate the reliability of results. While this is difficult to identify, poor reporting may be reflected in the overall quality score of a study.

Inverse variance weighting

The inverse variance is the most common measure of precision, representing a composite measure of total variance and sample size. Widely used weighting schemes based on the inverse variance are fixed effect or random effects meta-analytic models. The fixed effect model assumes that all the studies sample one true effect γ . The observed outcome θ_i for study i is then a function of a within-study error ε_i , $\theta_i = \gamma + \varepsilon_i$, where ε_i is normally distributed $\varepsilon_i \sim \mathcal{N}(0, se(\theta_i)^2)$. The standard error $se(\theta_i)$ is calculated from the sample standard deviation $sd(\theta_i)$ and sample size n_i as:

$$se(\theta_i) = \frac{sd(\theta_i)}{\sqrt{n_i}} \tag{5}$$

Alternatively, the random effects model supposes that each study samples a different true outcome μ_i , such that the combined

effect μ is the mean of a population of true effects. The observed effect θ_i for study i is then influenced by the intrastudy error ε_i and interstudy error ξ_i , $\theta_i = \mu_i + \varepsilon_i + \xi_i$, where ξ_i is also assumed to be normally distributed $\xi_i \sim \mathcal{N}(0, \tau^2)$, with τ^2 representing the extent of heterogeneity, or between-study (interstudy) variance.

Study-level estimates for a fixed effect or random effects model are weighted using the inverse variance:

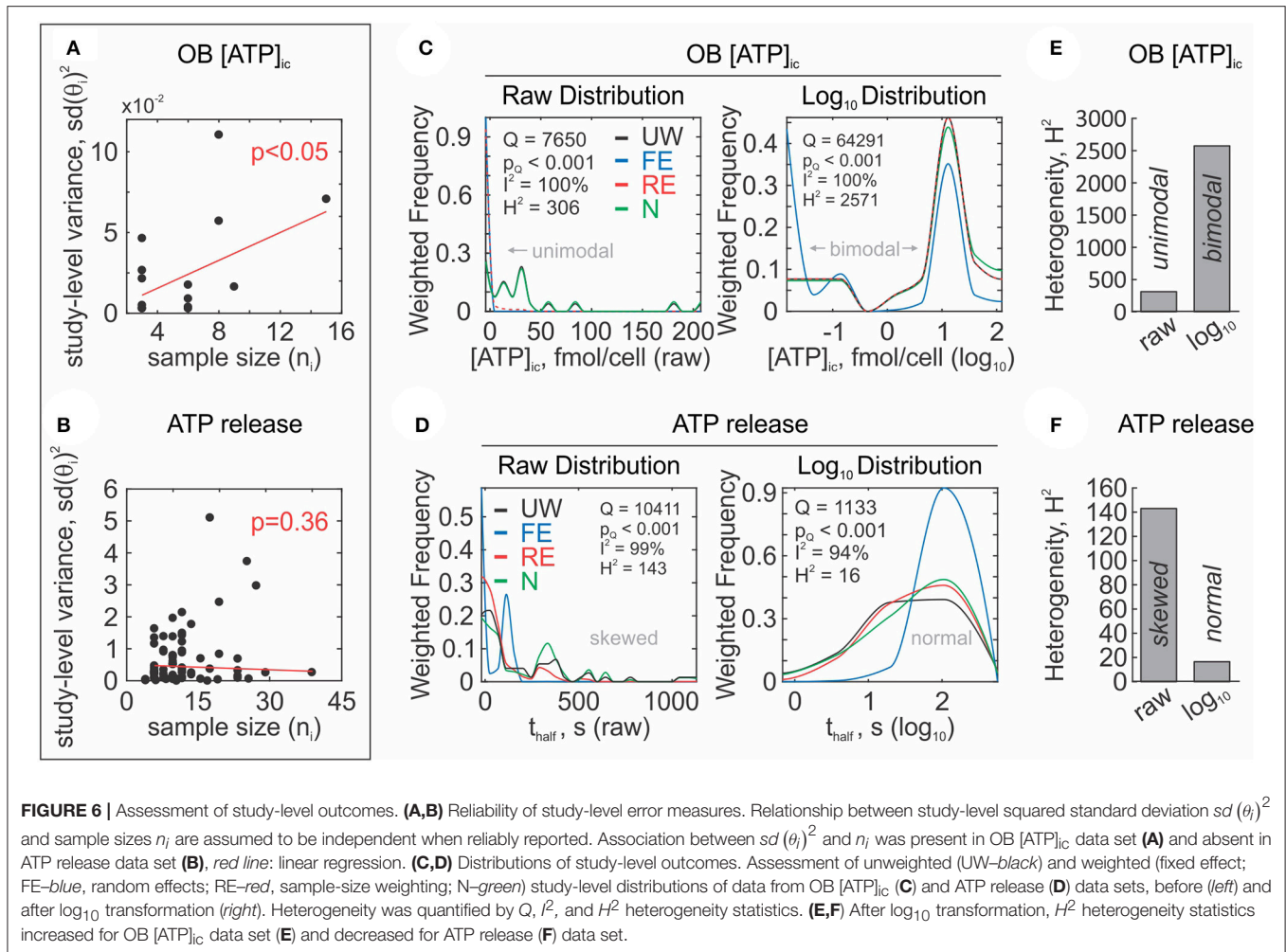
$$w_i = \begin{cases} \frac{1}{se(\theta_i)^2}, & \text{fixed effect} \\ \frac{1}{se(\theta_i)^2 + \tau^2}, & \text{random effects} \end{cases} \tag{6}$$

These weights are used to calculate the global outcome $\hat{\theta}$ (Equation 3) and the corresponding standard error $se(\hat{\theta})$:

$$se(\hat{\theta}) = \frac{1}{\sqrt{\sum_i^N w_i}} \tag{7}$$

where N = number of datasets/studies. In practice, random effects models are favored over the fixed effect model, due to the prevalence of heterogeneity in experimental methods and biological outcomes. However, when there is no between-study variability ($\tau^2 = 0$), the random effects model reduces to a fixed effect model. In contrast, when τ^2 is exceedingly large and interstudy variance dominates the weighting term [$\tau^2 \gg se(\theta_i)^2$], random effects estimates will tend to an unweighted mean.

Interstudy variance τ^2 estimators. Under the assumptions of a random effects model, the total variance is the sum of the intrastudy variance (experimental sampling error) and interstudy variance τ^2 (variability of true effects). Since the distribution of true effects is unknown, we must estimate the value of τ^2 based on study-level outcomes (Borenstein, 2009). The DerSimonian



and Laird (DL) method is the most commonly used in meta-analyses (DerSimonian and Laird, 1986). Other estimators such as the Hunter and Schmidt (Hunter and Schmidt, 2004), Hedges (Hedges and Olkin, 1985), Hartung-Makambi (Hartung and Makambi, 2002), Sidik-Jonkman (Sidik and Jonkman, 2005), and Paule-Mandel (Paule and Mandel, 1982) estimators have been proposed as either alternatives or improvements over the DL estimator (Sanchez-Meca and Marin-Martinez, 2008) and have been implemented in *MetaLab* (Table 3). Negative values of τ^2 are truncated at zero. An overview of the various τ^2 estimators along with recommendations on their use can be found elsewhere (Veroniki et al., 2016).

Sample-size weighting

Sample-size weighting is preferred in cases where variance estimates are unavailable or unreliable. Under this weighting scheme, study-level sample sizes are used in place of inverse variances as weights. The sampling error is then unaccounted for; however, since sampling error is random, larger sample sizes will effectively average out the error and produce more dependable results. This is contingent on reliable reporting of sample sizes which is difficult to assess and can be erroneous as detailed above. For a sample size weighted estimate, study-level sample sizes n_i

replace weights that are used to calculate the global effect size $\hat{\theta}$, such that

$$w_i = n_i \tag{8}$$

The pooled standard error $se(\hat{\theta})$ for the global effect is then:

$$se(\hat{\theta}) = \sqrt{\frac{\sum_i^N (se(\theta_i)^2 \cdot (n_i - 1))}{\sum_i^N (n_i - 1)}} \tag{9}$$

While sample size weighting is less affected by sampling variance, the performance of this estimator depends on the availability of studies (Marin-Martinez and Sanchez-Meca, 2010). When variances are reliably reported, sample-size weights should roughly correlate to inverse variance weights under the fixed effect model.

Meta-Analytic Data Distributions

One important consideration the reviewer should attend to is the normality of the study-level effects distributions assumed by most meta-analytic methods. Non-parametric methods that do not assume normality are available but are more computationally intensive and inaccessible to non-statisticians (Karabatsos et al.,

TABLE 3 | Interstudy variance estimators.

Estimator	τ^2 estimate
DerSimonian-Laird (DL) ^{*†}	$\tau_{DL}^2 = \frac{Q-(N-1)}{c}$
Hunter-Schmidt (HS) [*]	$\tau_{HS}^2 = \frac{Q-N}{\sum_i se(\theta_i)^{-2}}$
Hedges (H)	$\tau_H^2 = \frac{\sum_i \left(\theta_i - \left(\frac{\sum_i \theta_i}{N} \right) \right)^2}{N-1} - \frac{\sum_i se(\theta_i)^2}{N}$
Hartung-Makambi (HM) ^{*†}	$\tau_{HM}^2 = \frac{Q^2}{(2(N-1)+Q) \cdot c}$
Sidik-Jonkman (SJ)	$\tau_{SJ}^2 = \frac{\sum_i v_i^{-1} \left(\theta_i - \left(\frac{\sum_i v_i^{-1} \theta_i}{\sum_i v_i^{-1}} \right) \right)^2}{N-1}$, Where $v_i = \left(\frac{se(\theta_i)^2}{\left(\frac{\sum_i (\theta_i - \bar{\theta})^2}{N} \right)} + 1 \right)$ and $\bar{\theta} = \frac{1}{N} \sum_i \theta_i$
Paule-Mandel (PM) [#]	$\tau_{PM}^2 = \frac{\sum_i w_i (\theta_i - \hat{\theta}_{PM})^2 - \left(\sum_i w_i^2 se(\theta_i)^2 - \left(\frac{\sum_i w_i^2 se(\theta_i)^2}{\sum_i w_i} \right) \right)}{\sum_i w_i - \left(\frac{\sum_i w_i^2}{\sum_i w_i} \right)}$ Where $\hat{\theta}_{PM} = \frac{\sum_i (\theta_i \cdot w_i)}{\sum_i w_i}$

^{*} $Q = \sum_i \left(se(\theta_i)^{-2} \left(\theta_i - \frac{\sum_i se(\theta_i)^{-2} \theta_i}{\sum_i se(\theta_i)^{-2}} \right)^2 \right)$. [†] $c = \sum_i se(\theta_i)^{-2} - \frac{\sum_i (se(\theta_i)^{-2})^2}{\sum_i se(\theta_i)^{-2}}$. [#]iterative estimator.
N = number of datasets/studies.

2015). The performance of parametric meta-analytic methods has been shown to be robust to non-normally distributed effects (Kontopantelis and Reeves, 2012). However, this robustness is achieved by deriving artificially high estimates of heterogeneity for non-normally distributed data, resulting in conservatively wide confidence intervals and severely underpowered results (Jackson and Turner, 2017). Therefore, it is prudent to characterize the underlying distribution of study-level effects and perform transformations to normalize distributions to preserve the inferential integrity of the meta-analysis.

Assessing data distributions

Graphical approaches, such as the histogram, are commonly used to assess the distribution of data; however, in a meta-analysis, they can misrepresent the true distribution of effect sizes that may be different due to unequal weights assigned to each study. To address this, we can use a weighted histogram to evaluate effect size distributions (Figure 6). A weighted histogram can be constructed by first binning studies according to their effect sizes. Each bin is then assigned weighted frequencies, calculated as the sum of study-level weights within the given bin. The sum of weights in each bin are then normalized by the sum of all weights across all bins

$$P_j = \frac{\sum_i w_{ij}}{\sum_j^{nBins} \sum_i w_{ij}} \tag{10}$$

where P_j is the weighted frequency for bin j , w_{ij} is the weight for the effect size in bin j from study i , and $nBins$ is the total number of bins. If the distribution is found deviate from normality, the most common explanations are that (i) the distribution is skewed due to inconsistencies between studies, (ii) subpopulations exist

within the dataset giving rise to multimodal distributions or (iii) the studied phenomenon is not normally distributed. The source of inconsistencies and multimodality can be explored during the analysis of heterogeneity (i.e., to determine whether study-level characteristics can explain observed discrepancies). Skewness may however be inherent to the data when values are small, variances are large, and values cannot be negative (Limpert et al., 2001) and has been credited to be characteristic of natural processes (Grönholm and Annala, 2007). For sufficiently large sample sizes the central limit theorem holds that the means of a skewed data are approximately normally distributed. However, due to common limitation in the number of studies available for meta-analyses, meta-analytic global estimates of skewed distributions are often sensitive to extreme values. In these cases, data transformation can be used to achieve a normal distribution on the logarithmic scale (i.e., lognormal distribution).

Lognormal distributions

Since meta-analytic methods typically assume normality, the log transformation is a useful tool used to normalize skewed distributions (Figures 6C–F). In the ATP release dataset, we found that log transformation normalized the data distribution. However, in the case of the OB [ATP]_{ic} dataset, log transformation revealed a bimodal distribution that was otherwise not obvious on the raw scale.

Data normalization by log transformation allows meta-analytic techniques to maintain their inferential properties. The outcomes synthesized on the logarithmic scale can then be transformed to the original raw scale to obtain asymmetrical confidence intervals which further accommodate the skew in the data. Study-level effect sizes θ_i can be related to the logarithmic

mean Θ_i through the forward log transformation, meta-analyzed on the logarithmic scale, and back-transformed to the original scale using one of the back-transformation methods (Table 4). We have implemented three different back-transformation methods into MetaLab, including geometric approximation (anti-log), naïve approximation (rearrangement of forward-transformation method) and tailor series approximation (Higgins et al., 2008). The geometric back-transformation will yield an estimate of $\hat{\theta}$ that is approximately equal to the median of the study-level effects. The naïve or tailor series approximation differ in how the standard errors are approximated, which is used to obtain a point estimate on the original raw scale. The naïve and tailor series approximations were shown to maintain adequate inferential properties in the meta-analytic context (Higgins et al., 2008).

Confidence Intervals

Once the meta-analysis global estimate and standard error has been computed, reviewers may proceed to construct the confidence intervals (CI). The CI represents the range of values within which the true mean outcome is contained with the probability of $1-\alpha$. In meta-analyses, the CI conveys information about the significance, magnitude and direction of an effect, and is used for inference and generalization of an outcome. Values that do not fall in the range of the CI may be interpreted as significantly different. In general, the CI is computed as the product of the standard error $se(\hat{\theta})$ and the critical value $v_{1-\alpha/2}$:

$$\pm CI = \pm v_{1-\alpha/2} \cdot se(\hat{\theta}) \tag{11}$$

CI estimators

The critical value $v_{1-\alpha/2}$ is derived from a theoretical distribution and represents the significance threshold for level α . A theoretical distribution describes the probability of any given possible outcome occurrence for a phenomenon. Extreme outcomes that lie furthest from the mean are known as the tails. The most commonly used theoretical distributions are the z-distribution and t-distribution, which are both symmetrical and bell-shaped, but differ in how far reaching or “heavy” the tails are. Heavier tails will result in larger critical values which translate to wider confidence intervals, and vice versa. Critical values drawn from a z-distribution, known as z-scores (z), are used when data are normal, and a sufficiently large number of studies are available (>30). The tails of a z-distribution are independent of the sample size and reflect those expected for a normal distribution. Critical values drawn from a t-distribution, known as t-scores (t), also assume data are normally-distributed, however, are used when there are fewer available studies (<30) because the t-distribution tails are heavier. This produces more conservative (wider) CIs, which help ensure that the data are not misleading or misrepresentative when there is limited evidence available. The heaviness of the t-distribution tails is dictated by the degree of freedom df , which is related to the number of available studies N ($df = N-1$) such that fewer studies will result in heavier t-distribution tails and therefore larger critical values. Importantly, the t-distribution is asymptotically normal and will thus converge to a z-distribution for a sufficiently large number of studies, resulting in similar critical values. For example, for a significance level $\alpha = 0.05$ (5% false positive rate), the z-distribution will always yield a critical value $v = 1.96$,

TABLE 4 | Logarithmic Transformation Methods.

Forward-Transformation (raw to log₁₀)

	Mean	Standard error
	$\Theta_i = \log_{10}(\theta_i) - \left(\frac{se(\theta_i)^2}{2}\right)$	$se(\Theta_i) = \sqrt{\log_{10}\left(\frac{se(\theta_i)^2}{\theta_i^2} + 1\right)}$

Back-Transformation (log₁₀ to raw)

Method	Mean	Standard error
Geometric	$\hat{\theta} = 10^{\hat{\Theta}}$	$\pm CI_{1-\alpha/2}(\hat{\theta}) = 10^{\hat{\Theta} \pm v_{1-\alpha/2} \cdot se(\hat{\Theta})}$ $se(\hat{\theta}) = \frac{(+CI_{1-\alpha/2}(\hat{\theta})) - (-CI_{1-\alpha/2}(\hat{\theta}))}{2v_{1-\alpha/2}}$ Where $v_{1-\alpha/2}$ corresponds to critical value
Naïve approximately	$\hat{\theta} = 10^{\left(\hat{\Theta} + \frac{se(\hat{\Theta})^2}{2}\right)}$	$se(\hat{\theta}) = \frac{1}{\sqrt{n_i}} \left(10^{sd(\hat{\Theta})^2} - 1\right) 10^{2\hat{\Theta} + sd(\hat{\Theta})^2}$
Tailor Series approximately	$\hat{\theta} = 10^{\left(\hat{\Theta} + \frac{se(\hat{\Theta})^2}{2}\right)}$	$se(\hat{\theta}) = \sqrt{\frac{1}{n_i} 10^{(2\hat{\Theta} + sd(\hat{\Theta})^2)} sd(\hat{\Theta})^2 \left(1 + \left(\frac{sd(\hat{\Theta})^2}{2}\right)\right)}$

Forward-transformation of study-level estimates θ_i to corresponding log-transformed estimates Θ_i , and back-transformation of meta-analysis outcome $\hat{\Theta}$ to the corresponding outcome $\hat{\theta}$ on the raw scale (Higgins et al., 2008). $v_{1-\alpha/2}$: confidence interval critical value at significance level α .

regardless of how many studies are available. The t-distribution will however yield $\nu = 2.78$ for 5 studies, $\nu = 2.26$ for 10 studies, $\nu = 2.05$ for 30 studies and $\nu = 1.98$ for 100 studies, gradually converging to 1.96 as the number of studies increases. We have implemented the z-distribution and t-distribution CI estimators into MetaLab.

Evaluating Meta-Analysis Performance

In general, 95% of study-level outcomes are expected to fall within the range of the 95% global CI. To determine whether the global 95% CI is consistent with the underlying study-level outcomes, the coverage of the CI can be computed as the proportion of study-level 95% CIs that overlap with the global 95% CI:

$$\begin{aligned} |\hat{\theta} - \theta_i| &\leq \nu_{1-\frac{\alpha}{2}} \cdot se(\hat{\theta}) + \nu_{1-\frac{\alpha}{2}} \cdot se(\theta_i), & \text{covered} \\ |\hat{\theta} - \theta_i| &> \nu_{1-\frac{\alpha}{2}} \cdot se(\hat{\theta}) + \nu_{1-\frac{\alpha}{2}} \cdot se(\theta_i), & \text{not covered} \end{aligned} \quad (12)$$

The coverage is a performance measure used to determine whether inference made on the study-level is consistent with inference made on the meta-analytic level. Coverage that is less than expected for a specified significance level (i.e., <95% coverage for $\alpha = 0.05$) may be indicative of inaccurate estimators, excessive heterogeneity or inadequate choice of meta-analytic model, while coverage exceeding 95% may indicate an inefficient estimator that results in insufficient statistical power.

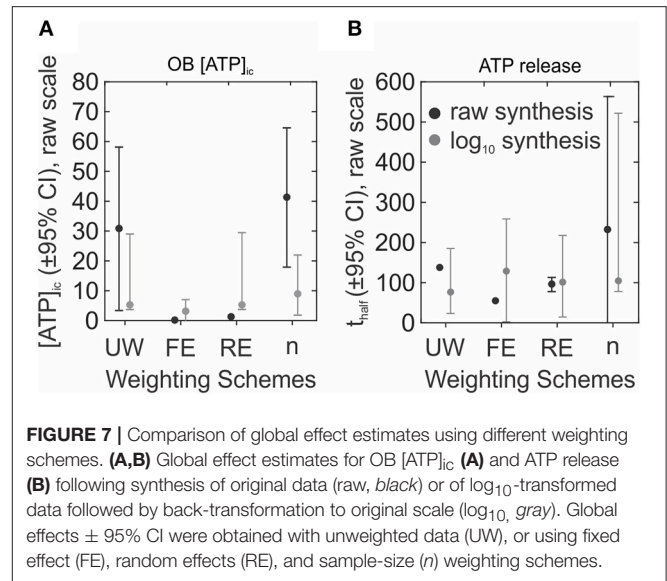
Overall, the performance of a meta-analysis is heavily influenced by the choice of weighting scheme and data transformation (Figure 7). This is especially evident in the smaller datasets, such as our OB [ATP]_{ic} example, where both the global estimates and the confidence intervals are dramatically different under different weighting schemes (Figure 7A). Working with larger datasets, such as ATP release kinetics, allows to somewhat reduce the influence of the assumed model (Figure 7B). However, normalizing data distribution (by log transformation) produces much more consistent outcomes under different weighting schemes for both datasets, regardless of the number of available studies (Figures 7A,B, log₁₀ synthesis).

Analysis of Heterogeneity

Heterogeneity refers to inconsistency between studies. A large part of conducting a meta-analysis involves quantifying and accounting for sources of heterogeneity that may compromise the validity of meta-analysis. Basic research meta-analytic datasets are expected to be heterogeneous because (i) basic research literature searches tend to retrieve more studies than clinical literature searches and (ii) experimental methodologies used in basic research are more diverse and less standardized compared to clinical research. The presence of heterogeneity may limit the generalizability of an outcome due to the lack of study-level consensus. Nonetheless, exploration of heterogeneity sources can be insightful for the field in general, as it can identify biological or methodological factors that influence the outcome.

Quantifying of Heterogeneity

Higgins and Thompson emphasized that a heterogeneity metric should be (i) dependent on magnitude of heterogeneity, (ii)



independent of measurement scale, (iii) independent of sample size and (iv) easily interpretable (Higgins and Thompson, 2002). Regrettably, the most commonly used test of heterogeneity is the Cochran's Q test (Borenstein, 2009), which has been repeatedly shown to have undesirable statistical properties (Higgins et al., 2003). Nonetheless, we will introduce it here, not because of its widespread use, but because it is an intermediary statistic used to obtain more useful measures of heterogeneity, H^2 and I^2 . The measure of total variation Q_{total} statistic is calculated as the sum of the weighted squared differences between the study-level means θ_i and the fixed effect estimate $\hat{\theta}_{FE}$:

$$Q_{total} = \sum_{i=1}^N \left(w_i \cdot (\theta_i - \hat{\theta}_{FE})^2 \right)$$

where $\hat{\theta}_{FE} = \frac{\sum_i se(\theta_i)^{-2} \theta_i}{\sum_i se(\theta_i)^{-2}}$ and $w_i = se(\theta_i)^{-2}$ (13)

The Q_{total} statistic is compared to a chi-square (χ^2) distribution ($df = N-1$) to obtain a p -value, which, if significant, supports the presence of heterogeneity. However, the Q -test has been shown to be inadequately powered when the number of studies is too low ($N < 10$) and excessively powered when study number is too high ($N > 50$) (Gavaghan et al., 2000; Higgins et al., 2003). Additionally, the Q_{total} statistic is not a measure of the magnitude of heterogeneity due to its inherent dependence on the number of studies. To address this limitation, H^2 heterogeneity statistics was developed as the relative excess in Q_{total} over degrees of freedom df :

$$H^2 = \frac{Q_{total}}{df} \quad (14)$$

H^2 is independent of the number of studies in the meta-analysis and is indicative of the magnitude of heterogeneity (Higgins and Thompson, 2002). For values <1, H^2 is truncated at 1, therefore

values of H^2 can range from one to infinity, where $H^2 = 1$ indicates homogeneity. The corresponding confidence intervals for H^2 are

$$H^2 \pm 95\% \text{ CI} = \left(e^{\ln(H) \pm 1.96 \cdot \sqrt{\frac{1}{2(df-1)} \left(1 - \frac{1}{3(df)^2}\right)}} \right)^2 \quad (15)$$

Intervals that do not overlap with 1 indicate significant heterogeneity. A more easily interpretable measure of heterogeneity is the I^2 statistic, which is a transformation of H^2 :

$$I^2 = \frac{H^2 - 1}{H^2} \cdot 100\% \quad (16)$$

The corresponding 95% CI for I^2 is derived from the 95% CI for H^2

$$I^2 \pm 95\% \text{ CI} = \frac{(H^2 \pm 95\% \text{ CI}) - 1}{(H^2 \pm 95\% \text{ CI})} \cdot 100\% \quad (17)$$

Values of I^2 range between 0 and 100% and describe the percentage of total variation that is attributed to heterogeneity. Like H^2 , I^2 provides a measure of the magnitude of heterogeneity. Values of I^2 at 25, 50, and 75% are generally graded as low, moderate and high heterogeneity, respectively (Higgins and Thompson, 2002; Pathak et al., 2017). However, several limitations have been noted for the I^2 statistic. I^2 has a non-linear dependence on τ^2 , thus I^2 will appear to saturate as it approaches 100% (Huedo-Medina et al., 2006). In cases of excessive heterogeneity, if heterogeneity is partially explained through subgroup analysis or meta-regression, residual unexplained heterogeneity may still be sufficient to maintain I^2 near saturation. Therefore, I^2 will fail to convey the decline in overall heterogeneity, while H^2 statistic that has no upper limit will allow to track changes in heterogeneity more meaningfully. In addition, a small number of studies (<10) will bias I^2 estimates, contributing to uncertainties inevitable associated with small meta-analyses (von Hippel, 2015). Of the three heterogeneity statistics Q_{total} , H^2 and I^2 described, we recommend that H^2 is used as it best satisfies the criteria for a heterogeneity statistic defined by Higgins and Thompson (2002).

Identifying bias

Bias refers to distortions in the data that may result in misleading meta-analytic outcomes. In the presence of bias, meta-analysis outcomes are often contradicted by higher quality large sample-sized studies (Egger et al., 1997), thereby compromising the validity of the meta-analytic study. Sources of observed bias include publication bias, methodological inconsistencies and quality, data irregularities due to poor quality design, inadequate analysis or fraud, and availability or selection bias (Egger et al., 1997; Ahmed et al., 2012). At the level of study identification and inclusion for meta-analysis, systematic searches are preferred over rapid review search strategies, as narrow search strategies may omit relevant studies. Withholding negative results is also a common source of publication bias, which is further

exacerbated by the small-study effect (the phenomenon by which smaller studies produce results with larger effect sizes than larger studies) (Schwarzer et al., 2015). By extension, smaller studies that produce negative results are more likely to not be published compared to larger studies that produce negative results. Identifying all sources of bias is unfeasible, however, tools are available to estimate the extent of bias present.

Funnel plots. Funnel plots have been widely used to assess the risk of bias and examine meta-analysis validity (Light and Pillemer, 1984; Borenstein, 2009). The logic underlying the funnel plot is that in the absence of bias, studies are symmetrically distributed around the fixed effect size estimate, due to sampling error being random. Moreover, precise study-level estimates are expected to be more consistent with the global effect size than less precise studies, where precision is inversely related to the study-level standard error. Thus, for an unbiased set of studies, study-level effects θ_i plotted in relation to the inverse standard error $1/se(\theta_i)$ will produce a funnel shaped plot. Theoretical 95% CIs for the range of plotted standard errors are included as reference to visualize the expected distribution of studies in the absence of bias (Sterne and Harbord, 2004). When bias is present, study-level effects will be asymmetrically distributed around the global fixed-effect estimate. In the past, funnel plot asymmetries have been attributed solely to publication bias, however they should be interpreted more broadly as a general presence of bias or heterogeneity (Sterne et al., 2011). It should be noted that rapid reviews (Figure 8A, left) are far more subject to bias than systematic reviews (Figure 8A, right), due to the increased likelihood of relevant study omission.

Heterogeneity sensitivity analyses

Inconsistencies between studies can arise for a number of reasons, including methodological or biological heterogeneity (Patsopoulos et al., 2008). Since accounting for heterogeneity is an essential part of any meta-analysis, it is of interest to identify influential studies that may contribute to the observed heterogeneity.

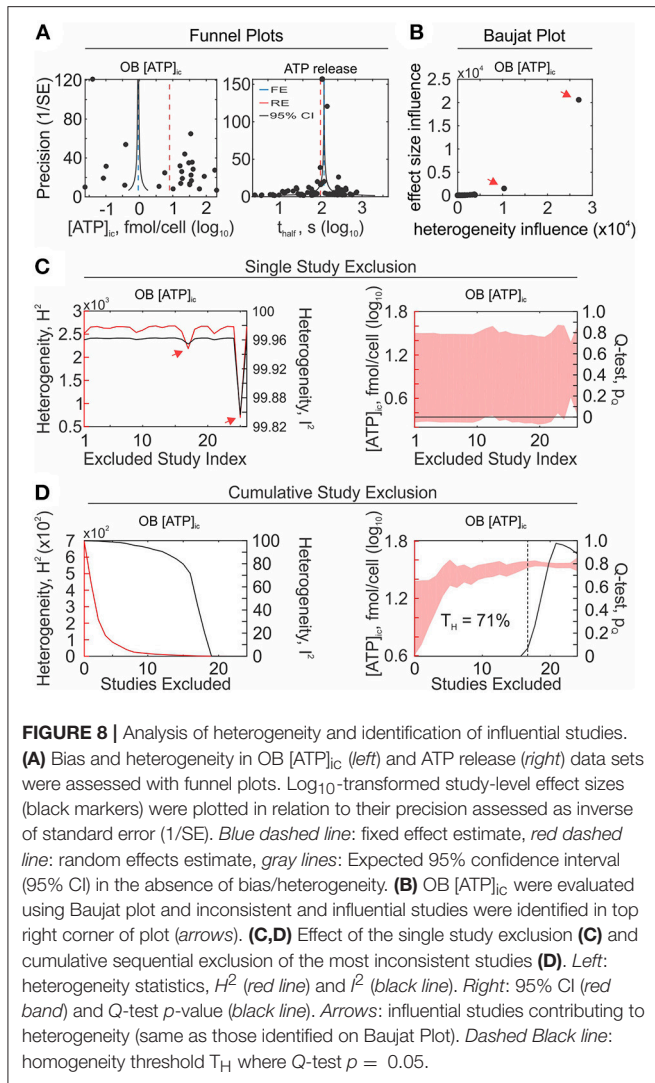
Baujat plot. The Baujat Plot was proposed as a diagnostic tool to identify the studies that contribute most to heterogeneity and influence the global outcome (Baujat, 2002). The graph illustrates the contribution Q_i^{inf} of each study to heterogeneity on the x-axis

$$Q_i^{inf} = \frac{\theta_i - \hat{\theta}_{FE}}{se(\theta_i)^2} \quad (18)$$

and contribution θ_i^{inf} to global effect on the y-axis

$$\theta_i^{inf} = \frac{\hat{\theta}_{-i} - \hat{\theta}_{FE}}{se(\hat{\theta}_{-i})^2} \quad (19)$$

Studies that strongly influence the global outcome and contribute to heterogeneity are visualized in the upper right corner of the plot (Figure 8B). This approach has been used to identify outlying studies in the past (Anzures-Cabrera and Higgins, 2010).



Single-study exclusion sensitivity. Single-study exclusion analysis assesses the sensitivity of the global outcome and heterogeneity to exclusion of single studies. The global outcomes and heterogeneity statistics are computed for a dataset with a single omitted study; single study exclusion is iterated for all studies; and influential outlying studies are identified by observing substantial declines in observed heterogeneity, as determined by Q_{total} , H^2 , or I^2 , and by significant differences in the global outcome (Figure 8C). Influential studies should not be blindly discarded, but rather carefully examined to determine the reason for inconsistency. If a cause for heterogeneity can be identified, such as experimental design flaw, it is appropriate to omit the study from the analysis. All reasons for omission must be justified and made transparent by reviewers.

Cumulative-study exclusion sensitivity. Cumulative study exclusion sequentially removes studies to maximize the decrease in total variance Q_{total} , such that a more homogenous set of studies with updated heterogeneity statistics is achieved with each iteration of exclusion (Figure 8D).

$$\hat{\theta}_{-j} \pm 95\% CI_{-j} \quad \text{where } j = \arg \max_i (Q - Q_{-i})^2 \quad (20)$$

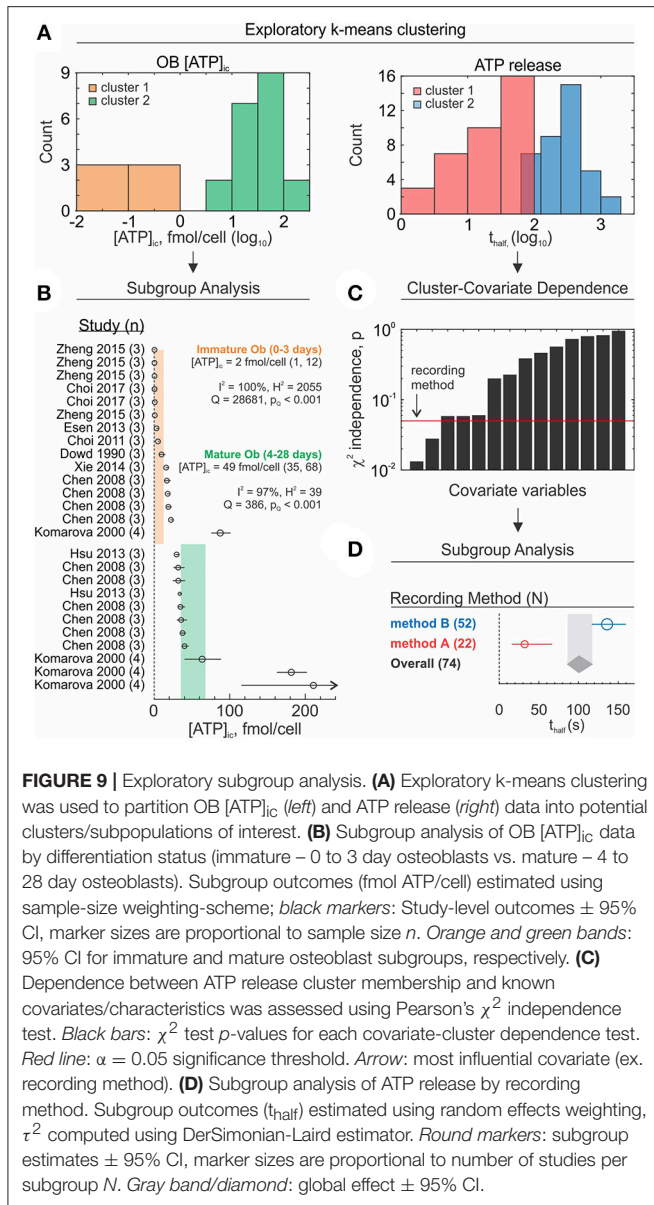
This method was proposed by Patsopoulos et al. to achieve desired levels of homogeneity (Patsopoulos et al., 2008), however, Higgins argued that its application should remain limited to (i) quantifying the extent to which heterogeneity permeates the set of studies and (ii) identifying sources of heterogeneity (Higgins, 2008). We propose the homogeneity threshold T_H as a measure of heterogeneity that can be derived from cumulative-study exclusion sensitivity analysis. The homogeneity threshold describes the percentage of studies that need to be removed (by the maximal Q-reduction criteria) before a homogenous set of studies is achieved. For example, in the OB [ATP]_{ic} dataset, the homogeneity threshold was 71%, since removal of 71% of the most inconsistent studies resulted in a homogeneous dataset (Figure 8D, right). After homogeneity is attained by cumulative exclusion, the global effect generally stabilizes with respect to subsequent study removal. This metric provides information about the extent of inconsistency present in the set of studies that is scale invariant (independent of the number of studies), and is easily interpretable.

Exploratory Analyses

The purpose of an exploratory analysis is to understand the data in ways that may not be represented by a pooled global estimate. This involves identifying sources of observed heterogeneity related to biological and experimental factors. Subgroup and meta-regression analyses are techniques used to explore known data groupings define by study-level characteristics (i.e., covariates). Additionally, we introduce the cluster-covariate dependence analysis, which is an unsupervised exploratory technique used to identify covariates that coincide well with natural groupings within the data, and the intrastudy regression analysis, which is used to validate meta-regression outcomes.

Cluster-covariate dependence analysis

Natural groupings within the data can be informative and serve as a basis to guide further analysis. Using an unsupervised k-means clustering approach (Lloyd, 1982), we can identify natural groupings within the study-level data and assign cluster memberships to these data (Figure 9A). Reviewers then have two choices: either proceed directly to subgroup analysis (Figure 9B) or look for covariates that co-cluster with cluster memberships (Figure 9C) In the latter case, dependencies between cluster memberships and known data covariates can be tested using Pearson's Chi-Squared test for independence. Covariates that coincide with clusters can be verified by subgroup analysis (Figure 9D). The dependence test is limited by the availability of studies and requires that at least 80% of covariate-cluster pairs are represented by at least 5 studies (McHugh, 2013). Clustering results should be considered exploratory and warrant further investigation due to several limitations. If the subpopulations were identified through clustering, however they do not depend on extracted covariates, reviewers risk assigning misrepresentative meaning to these clusters. Moreover,



conventional clustering methods always converge to a result, therefore the data will still be partitioned even in the absence of natural data groupings. Future adaptations of this method might involve using different clustering algorithms (hierarchical clustering) or independence tests (G-test for independence) as well as introducing weighting terms to bias clustering to reflect study-level precisions.

Subgroup analysis

Subgroup analyses attempt to explain heterogeneity and explore differences in effects by partitioning studies into characteristic groups defined by study-level categorical covariates (Figures 9B,D; Table 5). Subgroup effects are estimated along with corresponding heterogeneity statistics. To evaluate the extent to which subgroup covariates contribute to observed

inconsistencies, the explained heterogeneity $Q_{between}$ and unexplained heterogeneity Q_{within} can be calculated.

$$Q_{within} = \sum_{j=1}^S \left(\sum_{i=1}^{N_j} \left(se(\theta_i)^{-2} \cdot (\theta_i - \hat{\theta}_{(FE)j})^2 \right) \right) \quad (21)$$

where *S* is the total number of subgroups per given covariate and each subgroup *j* contains *N_j* studies. The explained heterogeneity $Q_{between}$ is then the difference between total and subgroup heterogeneity:

$$Q_{between} = Q_{total} - Q_{within} \quad (22)$$

If the *p*-value for the χ^2 distributed statistic $Q_{between}$ is significant, the subgrouping can be assumed to explain a significant amount of heterogeneity (Borenstein, 2009). Similarly, Q_{within} statistic can be used to test whether there is any residual heterogeneity present within the subgroups.

The $R^2_{explained}$ is a related statistic that can be used to describe the percent of total heterogeneity that was explained by the covariate and is estimated as

$$R^2_{explained} = \left(1 - \frac{\tau^2_{within}}{\tau^2_{total}} \right) \cdot 100\% \quad (23)$$

Where pooled heterogeneity within subgroups τ^2_{within} represents the remaining unexplained variation (Borenstein, 2009):

$$\tau^2_{within} = \frac{\sum_{j=1}^S Q_{(within)j} - \sum_{j=1}^S df_j}{\sum_{j=1}^S c_j}$$

$$\text{where } c_j = \sum_{i=1}^{N_j} se(\theta_i)^{-2} - \frac{\sum_i (se(\theta_i)^{-2})^2}{\sum_i se(\theta_i)^{-2}} \quad (24)$$

Subgroup analysis of the ATP release dataset revealed that recording method had a major influence on ATP release outcome, such that method A produced significantly lower outcomes than method B (Figure 9D; Table 5, significance determined by non-overlapping 95% CIs). Additionally, recording method accounted for a significant amount of heterogeneity ($Q_{between}$, $p < 0.001$), however it represented only 4% ($R^2_{explained}$) of the total observed heterogeneity. Needless to say, the remaining 96% of heterogeneity is significant (Q_{within} , $p < 0.001$). To explore the remaining heterogeneity, additional subgroup analysis can be conducted by further stratifying method A and method B subgroups by other covariates. However, in many meta-analyses multi-level data stratification may be unfeasible if covariates are unavailable or if the number of studies within subgroups are low.

Multiple comparisons. When multiple subgroups are present for a given covariate, and the reviewer wishes to investigate the statistical differences between the subgroups, the problem of multiple comparisons should be addressed. Error rates are multiplicative and increase substantially as the number of subgroup comparisons increases. The Bonferroni correction has

TABLE 5 | Exploratory subgroup analysis.

Subgroup summary statistics					
Group (N)	$\hat{\theta} \pm 95\% \text{ CI}$	I^2 (%)	H^2	Q	
Total (74)	101 (86, 117)	94	16	1133	
Method A (22)	32 (16, 66)	94	17	358	
Method B (52)	136 (117, 159)	92	13	669	
Accounting for heterogeneity with subgroup analysis					
	Q	df	p -value	Interpretation	
Total	1,133	73	<0.001	Data are heterogeneous	
Method A	358	21	<0.001	Data are heterogeneous	
Method B	669	51	<0.001	Data are heterogeneous	
Between	106	1	<0.001	Subgrouping explained significant heterogeneity	
Within	1,027	72	<0.001	Significant heterogeneity remains	

Effect and heterogeneity estimates of ATP release by recording method.

been advocated to control for false positive findings in meta-analyses (Hedges and Olkin, 1985) which involves adjusting the significance threshold:

$$\alpha^* = \frac{\alpha}{m} \tag{25}$$

α^* is the adjusted significance threshold to attain intended error rates α for m subgroup comparisons. Confidence intervals can then be computed using α^* in place of α :

$$\pm CI = \pm t_{1-\alpha^*/2} \cdot se(\hat{\theta}) \tag{26}$$

Meta-regression

Meta-regression attempts to explain heterogeneity by examining the relationship between study-level outcomes and continuous covariates while incorporating the influence of categorical covariates (Figure 10A). The main differences between conventional linear regression and meta-regression are (i) the incorporation of weights and (ii) covariates are at the level of the study rather than the individual sample. The magnitude of the relationship β_n between the covariates $x_{n,i}$ and outcome y_i for study i and covariate n are of interest when conducting a meta-regression analysis. It should be noted that the intercept β_0 of a meta-regression with negligible effect of covariates is equivalent to the estimate approximated by a weighted mean (Equation 3). The generalized meta-regression model is specified as

$$y_i = \beta_0 + \beta_1 x_{1,i} + \dots + \beta_n x_{n,i} + \eta_i + \varepsilon_i \tag{27}$$

where intrastudy variance ε_i is

$$\varepsilon_i \sim \mathcal{N}(0, se(\theta_i)^2) \tag{28}$$

and the deviation from the distribution of effects η_i depends on the chosen meta-analytic model:

$$\eta_i \sim \begin{cases} 0, & \text{fixed effect} \\ \mathcal{N}(0, \tau^2), & \text{random effects} \end{cases} \tag{29}$$

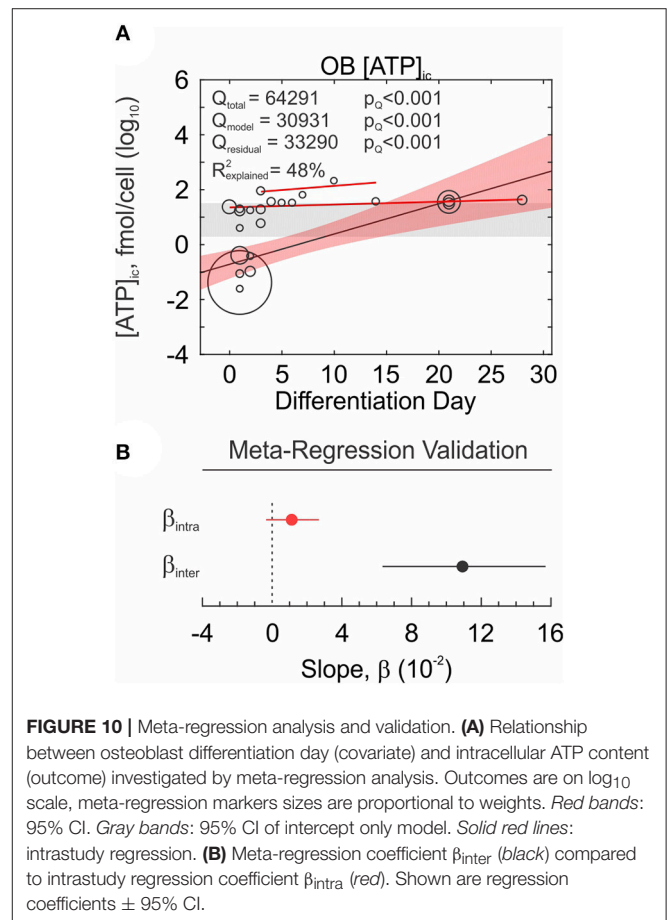


FIGURE 10 | Meta-regression analysis and validation. (A) Relationship between osteoblast differentiation day (covariate) and intracellular ATP content (outcome) investigated by meta-regression analysis. Outcomes are on log₁₀ scale, meta-regression markers sizes are proportional to weights. Red bands: 95% CI. Gray bands: 95% CI of intercept only model. Solid red lines: intrastudy regression. (B) Meta-regression coefficient β_{inter} (black) compared to intrastudy regression coefficient β_{intra} (red). Shown are regression coefficients \pm 95% CI.

The residual Q statistic that explains the dispersion of the studies from the regression line is calculated as follows

$$Q_{residual} = \sum_{i=1}^N (w_i \cdot (\theta_i - y_i)^2) \tag{30}$$

Where y_i is the predicted value at x_i according to the meta-regression model. $Q_{residual}$ is analogous to $Q_{between}$ computed during subgroup analysis and is used to test the degree of remaining unaccounted heterogeneity. $Q_{residual}$ is also used to approximate the unexplained interstudy variance $\tau_{residual}^2$

$$\tau_{residual}^2 = \frac{Q_{residual} - df}{c_{total}}$$

$$\text{where } c_{total} = \sum_i se(\theta_i)^{-2} - \frac{\sum_i (se(\theta_i)^{-2})^2}{\sum_i se(\theta_i)^{-2}} \quad (31)$$

Which can be used to calculate $R_{explained}^2$ estimated as

$$R_{explained}^2 = \left(1 - \frac{\tau_{residual}^2}{\tau_{total}^2}\right) \cdot 100\% \quad (32)$$

Q_{model} quantifies the amount of heterogeneity explained by the regression model and is analogous to Q_{within} computed during subgroup analysis.

$$Q_{model} = Q_{total} - Q_{residual} \quad (33)$$

Intrastudy regression analysis The challenge of interpreting results from a meta-regression is that relationships that exist within studies may not necessarily exist across studies, and vice versa. Such inconsistencies are known as aggregation bias and in the context of meta-analyses can arise from excess heterogeneity or from confounding factors at the level of the study. This problem has been acknowledged in clinical meta-analyses (Thompson and Higgins, 2002), however cannot be corrected without access to individual patient data. Fortunately, basic research studies often report outcomes at varying predictor levels (ex. dose-response curves), permitting for intrastudy (within-study) relationships to be evaluated by the reviewer. If study-level regression coefficients can be computed for several studies (Figure 10A, red lines), they can be pooled to estimate an overall effect β_{intra} . The meta-regression interstudy coefficient β_{inter} and the overall intrastudy-regression coefficient β_{intra} can then be compared in terms of magnitude and sign. Similarity in the magnitude and sign validates the existence of the relationship and characterizes its strength, while similarity in sign but not the magnitude, still supports the presence of the relationship, but calls for additional experiments to further characterize it. For the Ob [ATP]_i dataset, the magnitude of the relationship between osteoblast differentiation day and intracellular ATP concentration was inconsistent between intrastudy and interstudy estimates, however the estimates were of consistent sign (Figure 10B).

Limitations of exploratory analyses

When performed with knowledge and care, exploratory analysis of meta-analytic data has an enormous potential for hypothesis generation, cataloging current practices and trends, and identifying gaps in the literature. Thus, we emphasize the inherent limitations of exploratory analyses:

Data dredging. A major pitfall in meta-analyses is data dredging (also known as p-hacking), which refers to searching for significant outcomes only to assign meaning later. While exploring the dataset for potential patterns can identify outcomes of interest, reviewers must be wary of random patterns that can arise in any dataset. Therefore, if a relationship is observed it should be used to generate hypotheses, which can then be tested on new datasets. Steps to avoid data dredging involve defining an *a priori* analysis plan for study-level covariates, limiting exploratory analysis of rapid review meta-analyses and correcting for multiple comparisons.

Statistical power. The statistical power reflects the probability of rejecting the null hypothesis when the alternative is true. Meta-analyses are believed to have higher statistical power than the underlying primary studies, however this is not always true (Hedges and Pigott, 2001; Jackson and Turner, 2017). Random effects meta-analyses handle data heterogeneity by accounting for between-study variance, however this weakens the inference properties of the model. To maintain statistical powers that exceed those of the contributing studies in a random effects meta-analysis, at least five studies are required (Jackson and Turner, 2017). This consequently limits subgroup analyses that partition studies into smaller groups to isolate covariate-dependent effects. Thus, reviewers should ensure that group are not under-represented to maintain statistical power. Another determinant of statistical power is the expected effect size, which if small, will be much more difficult to support with existing evidence than if it is large. Thus, if reviewers find that there is insufficient evidence to conclude that a small effect exists, this should not be interpreted as evidence of no effect.

Causal inference. Meta-analyses are not a tool for establishing causal inference. However, there are several criteria for causality that can be investigated through exploratory analyses that include consistency, strength of association, dose-dependence and plausibility (Weed, 2000, 2010). For example, consistency, the strength of association, and dose-dependence can help establish that the outcome is dependent on exposure. However, reviewers are still posed with the challenge of accounting for confounding factors and bias. Therefore, while meta-analyses can explore various criteria for causality, causal claims are inappropriate, and outcomes should remain associative.

CONCLUSIONS

Meta-analyses of basic research can offer critical insights into the current state of knowledge. In this manuscript, we have adapted meta-analytic methods to basic science applications and provided a theoretical foundation, using OB [ATP]_i and ATP release datasets, to illustrate the workflow. Since the generalizability of any meta-analysis relies on the transparent, unbiased and accurate methodology, the implications of deficient reporting practices and the limitations of the meta-analytic methods were discussed. Emphasis was placed on the analysis and exploration of heterogeneity. Additionally, several alternative and supporting methods have been proposed, including a method for validating

meta-regression outcomes—intrastudy regression analysis, and a novel measure of heterogeneity—the homogeneity threshold. All analyses were conducted using *MetaLab*, a meta-analysis toolbox that we have developed in MATLAB R2016b. *MetaLab* has been provided for free to promote meta-analyses in basic research (<https://github.com/NMikolajewicz/MetaLab>).

In its current state, the translational pipeline from benchtop to bedside is an inefficient process, in one case estimated to produce ~1 clinically favorable clinical outcome for ~1,000 basic research studies (O'Collins et al., 2006). The methods we have described here serve as a general framework for comprehensive data consolidation, knowledge gap-identification, evidence-driven hypothesis generation and informed parameter estimation in computation modeling, which we hope will contribute to meta-analytic outcomes that better inform translation studies, thereby minimizing current failures in translational research.

AUTHOR CONTRIBUTIONS

Both authors contributed to the study conception and design, data acquisition and interpretation and drafting and critical

revision of the manuscript. NM developed *MetaLab*. Both authors approved the final version to be published.

ACKNOWLEDGMENTS

This work was supported by Natural Sciences and Engineering Research Council (NSERC, RGPIN-288253) and Canadian Institutes for Health Research (CIHR MOP-77643). NM was supported by the Faculty of Dentistry, McGill University and le Réseau de Recherche en Santé Buccodentaire et Osseuse (RSBO). Special thanks to Ali Mohammed (McGill University) for help with validation of *MetaLab* data extraction module.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fphys.2019.00203/full#supplementary-material>

REFERENCES

- Ahmed, I., Sutton, A. J., and Riley, R. D. (2012). Assessment of publication bias, selection bias, and unavailable data in meta-analyses using individual participant data: a database survey. *Br. Med. J.* 344:d7762 doi: 10.1136/bmj.d7762
- Altman, D. G., and Bland, J. M. (2005). Standard deviations and standard errors. *Br. Med. J.* 331, 903–903. doi: 10.1136/bmj.331.7521.903
- Anzures-Cabrera, J., and Higgins, J. P. T. (2010). Graphical displays for meta-analysis: an overview with suggestions for practice. *Res. Synth. Methods* 1, 66–80. doi: 10.1002/jrsm.6
- Baguley, T. (2009). Standardized or simple effect size: what should be reported? *Br. J. Soc. Psychol.* 100, 603–617. doi: 10.1348/000712608X377117
- Barendregt, J., and Doi, S. (2009). *MetaXL User Guide: Version 1.0*. Wilston, QLD: EpiGear International Pty Ltd.
- Baujat, B. (2002). A graphical method for exploring heterogeneity in meta-analyses: application to a meta-analysis of 65 trials. *Stat. Med.* 21:18. doi: 10.1002/sim.1221
- Bax, L. (2016). *MIX 2.0 – Professional Software for Meta-analysis in Excel. Version 2.0.1.5. BiostatXL*. Available online at: <https://www.meta-analysis-made-easy.com>
- Bittker, J. A., and Ross, N. T. (2016). *High Throughput Screening Methods: Evolution and Refinement*. Cambridge: Royal Society of Chemistry. doi: 10.1039/9781782626770
- Bodin, P., Milner, P., Winter, R., and Burnstock, G. (1992). Chronic hypoxia changes the ratio of endothelin to ATP release from rat aortic endothelial cells exposed to high flow. *Proc. Biol. Sci.* 247, 131–135. doi: 10.1098/rspb.1992.0019
- Borenstein, M. (2009). *Introduction to Meta-Analysis*. Chichester: John Wiley & Sons. doi: 10.1002/9780470743386
- Borenstein, M., Hedges, L., Higgins, J. P. T., and Rothstein, H. R. (2005). *Comprehensive meta-analysis (Version 2.2.027) [Computer software]*. Englewood, CO.
- Bramer, W. M., Giustini, D., de Jonge, G. B., Holland, L., and Bekhuis, T. (2016). De-duplication of database search results for systematic reviews in EndNote. *J. Med. Libr. Assoc.* 104, 240–243. doi: 10.3163/1536-5050.104.3.014
- Chowdhry, A. K., Dworkin, R. H., and McDermott, M. P. (2016). Meta-analysis with missing study-level sample variance data. *Stat. Med.* 35, 3021–3032. doi: 10.1002/sim.6908
- Cochrane Collaboration (2011). *Review Manager (RevMan) [Computer Program]*. Copenhagen.
- Cox, M., Harris, P., and Siebert, B. R.-L. (2003). Evaluation of measurement uncertainty based on the propagation of distributions using monte carlo simulation. *Measure. Techniq.* 46, 824–833. doi: 10.1023/B:METE.0000008439.82231.ad
- DeLuca, J. B., Mullins, M. M., Lyles, C. M., Crepaz, N., Kay, L., and Thadiparthi, S. (2008). Developing a comprehensive search strategy for evidence based systematic reviews. *Evid. Based Libr. Inf. Pract.* 3, 3–32. doi: 10.18438/B8KP66
- DerSimonian, R., and Laird, N. (1986). Meta-analysis in clinical trials. *Control. Clin. Trials* 7, 177–188. doi: 10.1016/0197-2456(86)90046-2
- Ecker, E. D., and Skelly, A. C. (2010). Conducting a winning literature search. *Evid. Based Spine Care J.* 1, 9–14. doi: 10.1055/s-0028-1100887
- Egger, M., Smith, G. D., Schneider, M., and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Br. Med. J.* 315, 629–634. doi: 10.1136/bmj.315.7109.629
- Fingfeld-Connett, D., and Johnson, E. D. (2013). Literature search strategies for conducting knowledge-building and theory-generating qualitative systematic reviews. *J. Adv. Nurs.* 69, 194–204. doi: 10.1111/j.1365-2648.2012.06037.x
- Ganann, R., Ciliska, D., and Thomas, H. (2010). Expediting systematic reviews: methods and implications of rapid reviews. *Implementation Sci.* 5, 56–56. doi: 10.1186/1748-5908-5-56
- Gavaghan, D. J., Moore, R. A., and McQuay, H. J. (2000). An evaluation of homogeneity tests in meta-analyses in pain using simulations of individual patient data. *Pain* 85, 415–424. doi: 10.1016/S0304-3959(99)00302-4
- Gopalakrishnan, S., and Ganeshkumar, P. (2013). Systematic reviews and meta-analysis: understanding the best evidence in primary healthcare. *J. Fam. Med. Prim. Care* 2, 9–14. doi: 10.4103/2249-4863.109934
- Grönholm, T., and Annala, A. (2007). Natural distribution. *Math. Biosci.* 210, 659–667. doi: 10.1016/j.mbs.2007.07.004
- Haby, M. M., Chapman, E., Clark, R., Barreto, J., Reveiz, L., and Lavis, J. N. (2016). What are the best methodologies for rapid reviews of the research evidence for evidence-informed decision making in health policy and practice: a rapid review. *Health Res. Policy Syst.* 14:83. doi: 10.1186/s12961-016-0155-7
- Hartung, J., and Makambi, K. H. (2002). Positive estimation of the between-study variance in meta-analysis: theory and methods. *S. Afr. Stat. J.* 36, 55–76.
- Hedges, L. V., and Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. New York, NY: Academic Press.
- Hedges, L. V., and Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychol. Methods* 6, 203–217. doi: 10.1037/1082-989X.6.3.203

- Higgins, J. P. (2008). Commentary: heterogeneity in meta-analysis should be expected and appropriately quantified. *Int. J. Epidemiol.* 37, 1158–1160. doi: 10.1093/ije/dyn204
- Higgins, J. P., and Green, S. (Eds.) (2011). *Cochrane Handbook for Systematic Reviews of Interventions*, Vol. 4. Oxford: John Wiley & Sons.
- Higgins, J. P., and Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Stat. Med.* 21, 1539–1558. doi: 10.1002/sim.1186
- Higgins, J. P., Thompson, S. G., Deeks, J. J., and Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *Br. Med. J.* 327, 557–560. doi: 10.1136/bmj.327.7414.557
- Higgins, J. P., White, I. R., and Anzures-Cabrera, J. (2008). Meta-analysis of skewed data: combining results reported on log-transformed or raw scales. *Stat. Med.* 27, 6072–6092. doi: 10.1002/sim.3427
- Huedo-Medina, T. B., Sanchez-Meca, J., Marin-Martinez, F., and Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I² index? *Psychol. Methods* 11, 193–206. doi: 10.1037/1082-989X.11.2.193
- Hunter, J. E., and Schmidt, F. L. (2004). *Methods of Meta-analysis: Correcting Error and Bias in Research Findings*. Thousand Oaks, CA: Sage.
- Jackson, D., and Turner, R. (2017). Power analysis for random-effects meta-analysis. *Res. Synth. Methods* 8, 290–302. doi: 10.1002/jrs.m.1240
- JASP Team (2018). *JASP (Version 0.9) [Computer Software]*. Amsterdam.
- Karabatsos, G., Talbot, E., and Walker, S. G. (2015). A Bayesian nonparametric meta-analysis model. *Res. Synth. Methods* 6, 28–44. doi: 10.1002/jrs.m.1117
- Kontopantelis, E., and Reeves, D. (2012). Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: a simulation study. *Stat. Methods Med. Res.* 21, 409–426. doi: 10.1177/0962280210392008
- Kwon, Y., Lemieux, M., McTavish, J., and Wathen, N. (2015). Identifying and removing duplicate records from systematic review searches. *J. Med. Libr. Assoc.* 103, 184–188. doi: 10.3163/1536-5050.103.4.004
- Light, R. J., and Pillemer, D. B. (1984). *Summing Up: The Science of Reviewing Research*. Cambridge, MA: Harvard University Press.
- Limpert, E., Stahel, W. A., and Abbt, M. (2001). Log-normal Distributions across the sciences: keys and clues: on the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability—normal or log-normal: That is the question. *AIBS Bull.* 51, 341–352.
- Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Trans. Inf. Theory* 28, 129–137. doi: 10.1109/TIT.1982.1056489
- Lorenzetti, D. L., and Ghali, W. A. (2013). Reference management software for systematic reviews and meta-analyses: an exploration of usage and usability. *BMC Med. Res. Methodol.* 13, 141–141. doi: 10.1186/1471-2288-13-141
- Marin-Martinez, F., and Sanchez-Meca, J. (2010). Weighting by inverse variance or by sample size in random-effects meta-analysis. *Educ. Psychol. Meas.* 70, 56–73. doi: 10.1177/0013164409344534
- Mattivi, J. T., and Buchberger, B. (2016). Using the amstar checklist for rapid reviews: is it feasible? *Int. J. Technol. Assess. Health Care* 32, 276–283. doi: 10.1017/S0266462316000465
- McGowan, J., and Sampson, M. (2005). Systematic reviews need systematic searchers. *J. Med. Libr. Assoc.* 93, 74–80.
- McHugh, M. L. (2013). The Chi-square test of independence. *Biochem. Med.* 23, 143–149. doi: 10.11613/BM.2013.018
- Mikolajewicz, N., Mohammed, A., Morris, M., and Komarova, S. V. (2018). Mechanically-stimulated ATP release from mammalian cells: systematic review and meta-analysis. *J. Cell Sci.* 131:22. doi: 10.1242/jcs.223354
- Milo, R., Jorgensen, P., Moran, U., Weber, G., and Springer, M. (2010). BioNumbers—the database of key numbers in molecular and cell biology. *Nucleic Acids Res.* 38:D750–3. doi: 10.1093/nar/gkp889
- Moher, D., Liberati, A., Tetzlaff, J., and Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med.* 6:e1000097. doi: 10.1371/journal.pmed.1000097
- O'Collins, V. E., Macleod, M. R., Donnan, G. A., Horkey, L. L., van der Worp, B. H., and Howells, D. W. (2006). 1,026 experimental treatments in acute stroke. *Ann. Neurol.* 59, 467–477. doi: 10.1002/ana.20741
- Pathak, M., Dwivedi, S. N., Deo, S. V. S., Sreenivas, V., and Thakur, B. (2017). Which is the preferred measure of heterogeneity in meta-analysis and why? a revisit. *Biostat Biometrics Open Acc.* 1, 1–7. doi: 10.19080/BBOAJ.2017.01.555555
- Patsopoulos, N. A., Evangelou, E., and Ioannidis, J. P. A. (2008). Sensitivity of between-study heterogeneity in meta-analysis: proposed metrics and empirical evaluation. *Int. J. Epidemiol.* 37, 1148–1157. doi: 10.1093/ije/dyn065
- Paule, R. C., and Mandel, J. (1982). Consensus values and weighting factors. *J. Res. Natl. Bur. Stand.* 87, 377–385. doi: 10.6028/jres.087.022
- Sanchez-Meca, J., and Marin-Martinez, F. (2008). Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychol. Methods* 13, 31–48. doi: 10.1037/1082-989X.13.1.31
- Schwarzer, G., Carpenter, J. R., and Rücker, G. (2015). “Small-study effects in meta-analysis,” in *Meta-Analysis with R*, eds G. Schwarzer, J. R. Carpenter, and G. Rücker (Cham: Springer International Publishing), 107–141.
- Sena, E., van der Worp, H. B., Howells, D., and Macleod, M. (2007). How can we improve the pre-clinical development of drugs for stroke? *Trends Neurosci.* 30, 433–439. doi: 10.1016/j.tins.2007.06.009
- Sheldrake, R. (1997). Experimental effects in scientific research: how widely are they neglected? *Bull. Sci. Technol. Soc.* 17, 171–174. doi: 10.1177/027046769701700405
- Sidik, K., and Jonkman, J. N. (2005). Simple heterogeneity variance estimation for meta-analysis. *J. R. Stat. Soc. Ser. C Appl. Stat.* 54, 367–384. doi: 10.1111/j.1467-9876.2005.00489.x
- Sterne, J. A., Sutton, A. J., Ioannidis, J. P., Terrin, N., Jones, D. R., Lau, J., et al. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *Br. Med. J.* 343:d4002. doi: 10.1136/bmj.d4002
- Sterne, J. A. C., and Harbord, R. (2004). Funnel plots in meta-analysis. *Stata J.* 4, 127–141. doi: 10.1177/1536867X0400400204
- Thompson, S. G., and Higgins, J. P. (2002). How should meta-regression analyses be undertaken and interpreted? *Stat. Med.* 21, 1559–1573. doi: 10.1002/sim.1187
- Vaux, D. L., Fidler, F., and Cumming, G. (2012). Replicates and repeats—what is the difference and is it significant?: a brief discussion of statistics and experimental design. *EMBO Rep.* 13, 291–296. doi: 10.1038/embor.2012.36
- Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., et al. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res. Synth. Methods* 7, 55–79. doi: 10.1002/jrsm.1164
- Vesterinen, H. M., Sena, E. S., Egan, K. J., Hirst, T. C., Churolov, L., Currie, G. L., et al. (2014). Meta-analysis of data from animal studies: a practical guide. *J. Neurosci. Methods* 221, 92–102. doi: 10.1016/j.jneumeth.2013.09.010
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.* 36, 1–48. doi: 10.18637/jss.v036.i03
- von Hippel, P. T. (2015). The heterogeneity statistic I² can be biased in small meta-analyses. *BMC Med. Res. Methodol.* 15:35. doi: 10.1186/s12874-015-0024-z
- Weed, D. L. (2000). Interpreting epidemiological evidence: how meta-analysis and causal inference methods are related. *Int. J. Epidemiol.* 29, 387–390. doi: 10.1093/ije/dyn29.3.387
- Weed, D. L. (2010). Meta-analysis and causal inference: a case study of benzene and non-hodgkin lymphoma. *Ann. Epidemiol.* 20, 347–355. doi: 10.1016/j.annepidem.2010.02.001

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2019 Mikolajewicz and Komarova. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.