

## Evidence and Quality (Part 2): Can systematic literature reviews be made more efficient by excluding poor quality studies at screening, without losing validity?

Systematic literature reviews (SLRs) provide an important evidence base for decision-making in healthcare. A truly systematic approach mandates considering the entire set of eligible evidence for extraction and reporting which can make these reviews extremely time- and resource-intensive given that the quantum of literature has been growing at a rapid pace. In resource-constrained settings, and for certain types of SLRs, one option is to consider “third screening” (or “prioritisation”) based on a certain pre-specified criteria in order to narrow down the eligible universe of studies (per the PICOS) to the highest-quality studies only. However, one may question the validity of this additional screening step and its subsequent implications while drawing conclusions. Focusing on a sub-type of SLRs that examines disease burden and unmet need using observational/real-world data, we present summary findings of an analysis to determine if the prioritised set of studies selected during third screening can provide a good representation of the overall universe of eligible studies without compromising validity.

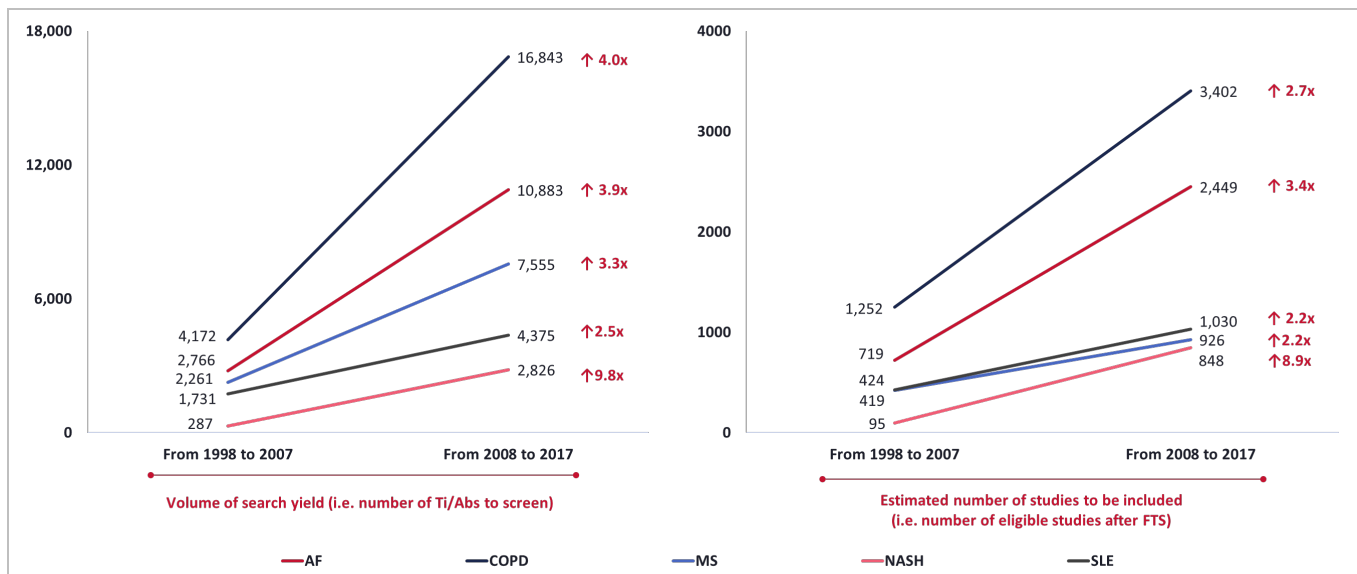
### Background

Systematic literature reviews (SLRs) are complex undertakings, yet foundational to healthcare decision making. The idea behind a “systematic” review of the literature is to eliminate bias in retrieval and reporting of existing evidence which is achieved by having a comprehensive search strategy; well-defined eligibility criteria to select studies; screening, extraction and methodological quality assessment of included studies by two reviewers independently; as well as qualitative and quantitative (if feasible) analysis of the data. While they are meant to be comprehensive, transparent and reproducible, SLRs can be time-consuming and costly, depending on the quantum of the literature available on the research question(s) of interest [1].

SLRs are conducted for different objectives and focus on different parameters.

In the case of a comparative efficacy (of the type required for many HTA submissions), it is important to capture relevant data from **all available clinical trials**, as exhaustiveness is the key driver to obtain valid results. On the other hand, for literature reviews focused on unmet need, disease burden, epidemiology, and real-world treatment patterns (i.e., reviews focused on real-world observational data), exhaustiveness might be difficult to achieve given that the volume of observational literature can be multi-fold (and growing) compared to the clinical trial literature, and potentially, representativeness or generalisability of the findings should be the key driver. **Figure 1** shows the evolution in the volume of observational studies measuring disease burden over time for different therapeutic areas.

Figure 1: Evolution of the volume of published observational studies measuring disease burden over time



Note: Systematic searches for observational studies were conducted in Embase

Source: [2]

AF = Atrial fibrillation; COPD = Chronic obstructive pulmonary disease; MS = Multiple sclerosis; NASH = Non-alcoholic steatohepatitis; SLE = Systemic lupus erythematosus

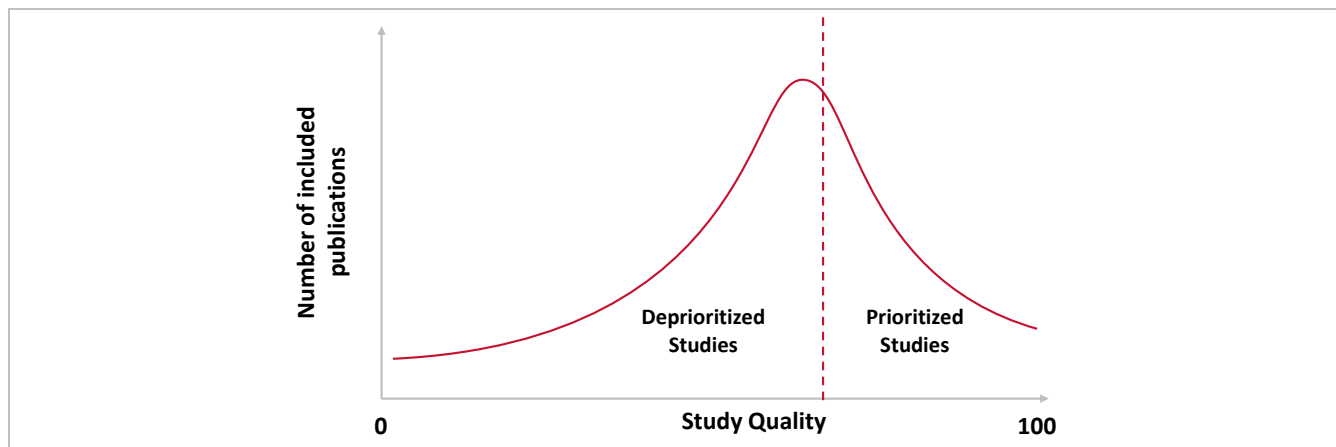
Given the size and complexity of the literature, one option to reduce the quantum of the available literature for final analysis and reporting is the use of an additional screening step called “third screening” or “prioritisation”. During third screening, the entire set of eligible studies [following full-text screening (FTS)] is re-screened and a high-quality subset of those studies is identified by using certain pre-specified parameters such as study design, sample size, follow-up period, geography of interest, adjustment for confounders, and generalizability.

Although the third screening approach can lead to efficiency gains by reducing the volume of studies to be extracted and hence saving time and cost, there remains a concern that the prioritised sample of studies might not be representative

of the entire universe of eligible studies after FTS. In other words, it might be argued that the prioritised literature is skewed or biased which might affect the validity of the conclusions drawn from the SLR.

The question for this pilot analysis was whether the sample of prioritised studies (to the right of the dotted line in the theoretical distribution below, Figure 2) is representative of all the studies that were included after FTS. We hypothesized that the gain in efficiency by using a third screening approach does not come at the cost of validity. Stated differently, we hypothesized that de-prioritisation of low-quality studies [left of the dotted line] does not lead to any material loss of information, or ability to generalise the findings.

Figure 2: Distribution of eligible studies by quality, with cut-off for highest quality studies only [hypothetical distribution curve]



## Methods

We selected a recently completed SLR that examined the disease burden of age-related macular degeneration (wAMD). The scope of the project included assessment of clinical, humanistic and economic burden of wAMD based on observational studies. After title/abstract and full-text screening, a total of **200** studies were found to be eligible. However, given the nature of the client's decision problem, as well as timeline and budget constraints, we agreed to conduct a third screening to reduce the quantum of literature for final extraction and reporting without compromising the generalizability or representativeness of the data.

In consultation with the client, the following criteria were used for third screening or prioritisation: study design (cohort studies preferred over cross-sectional studies), duration of follow-up for longitudinal studies (longer duration preferred over shorter duration), sample size and representativeness of the study sample (larger and more representative studies preferred), data collection period (studies collecting data after 2010 were given preference), whether confounders were adjusted for (multivariable analyses preferred over univariable analyses), and whether the study was based in a geographic region of interest (US, Europe, Japan, and China preferred). When there were limited or no data for any research question, *any* available relevant study was prioritised regardless of its methodological quality, so that the corresponding research question could be answered in some capacity at least.

After third screening, a total of **37** studies were prioritised and **163** were deprioritised. In line with our hypothesis stated earlier, we wanted to investigate if 37 prioritised studies were a good representation of the original 200 studies that were deemed eligible after FTS. In order to test our hypothesis, we extracted the following demographic and clinical variables for all 200 studies: average age, gender (proportion of males and females), proportion of patients with unilateral and bilateral wAMD, visual acuity at baseline and the scale used to measure it, and treatment status at baseline. While age and gender are the key demographic variables, the presence of unilateral or bilateral disease, visual acuity score, and whether patients have received prior treatment are each different indicators of the underlying disease severity. Together, these variables were chosen since they primarily define a patient population in any study investigating the burden of wAMD. All data were extracted by one reviewer and checked by another reviewer. Any discrepancies were resolved by discussing with a senior reviewer.

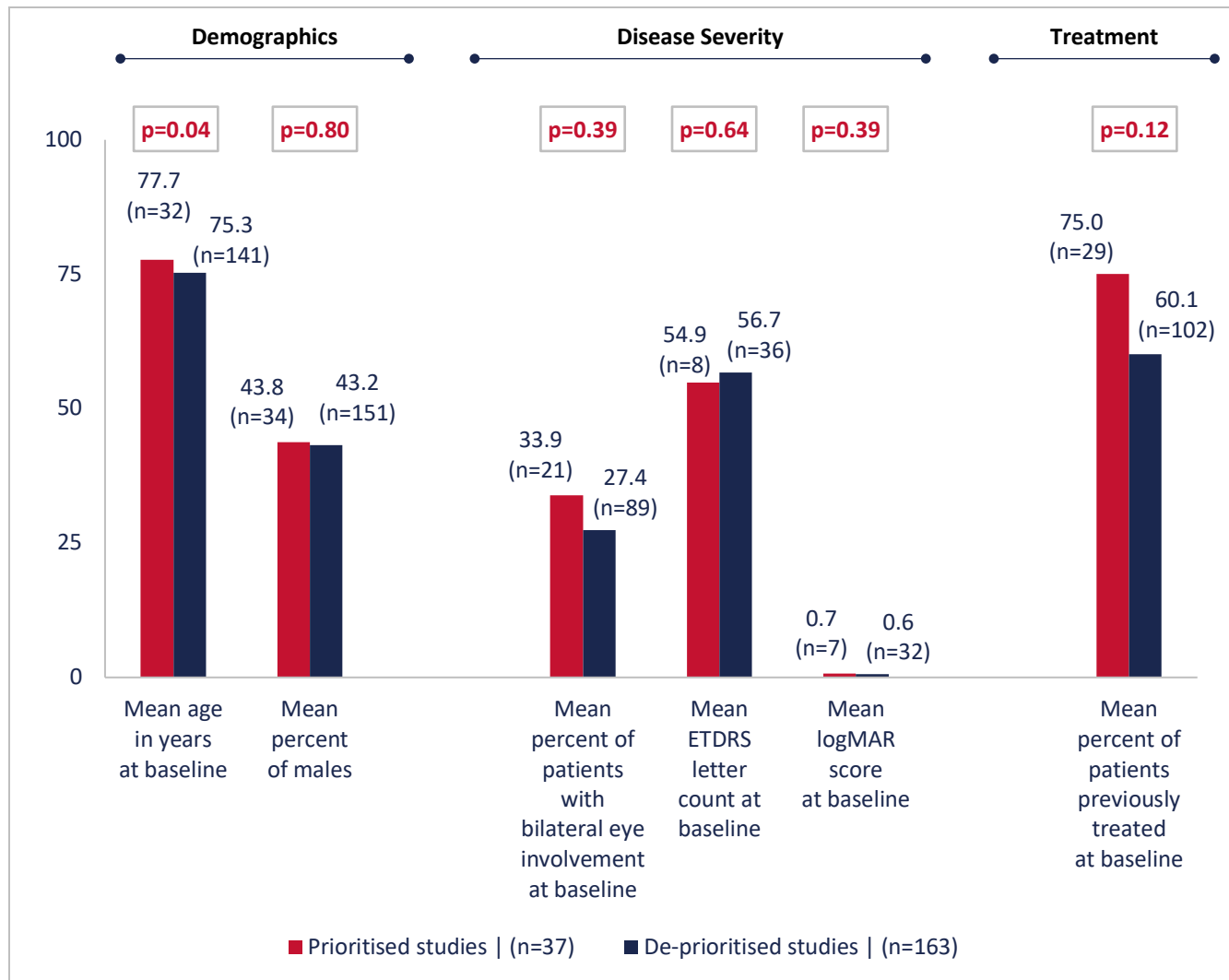
Per our hypothesis, if it could be shown statistically that there are no significant differences in patient characteristics between those two data sets (i.e., prioritised versus deprioritised studies), we could be confident that the third screening approach has not resulted in a skewed/biased sample. [Appendix 1](#) shows the detailed statistical approach used to test our hypothesis.

## Results

Of 200 studies, 37 were prioritised and 163 were deprioritised based on the criteria described earlier. [Appendix 2](#) shows the key characteristics of the 200 studies. [Figure 3](#) show the comparison of these two groups with respect to the demographic, disease severity- and treatment-related variables to evaluate the presence of any significant differences. Using the Bonferroni-adjusted alpha of 0.016 (please see [Appendix 1](#)), no statistically significant differences were found between the two groups.

The age variable, although statistically significant using the conventional p-value of 0.05, did not cross the adjusted threshold of 0.016. Moreover, a difference of 2.4 years between the two groups was not considered to be of any clinical significance. The other patient variables such as gender, disease severity and baseline treatment were neither statistically nor clinically significantly different between the two groups.

Figure 3: Comparison of patient characteristics between prioritised and deprioritised studies



ETDRS = Early Treatment Diabetic Retinopathy Study; LogMAR = Logarithmic minimal angle of resolution; VA = visual acuity

**ETDRS letter count:** Visual acuity (VA) is tested commonly using ETDRS charts and normal VA is expressed as ETDRS letter score of 85, with score ranging from 0-100; higher scores indicating better VA.

**LogMAR score:** Studies may convert Snellen decimal or ETDRS letter scores to produce a logMAR score, which can vary from -0.30 to 2.00, with a normal VA score of 0.00; lower logMAR indicates better VA.

**Note:** There were missing data of varying degree for each variable, with 7.5% (15/200) missing data for gender, 13.5% (27/200) for average age, 34.5% (69/200) for treatment status, 45.0% (90/200) for laterality, 52.5% (105/200) for baseline VA

## Discussion

This pilot research, based on a single SLR, showed that third screening or prioritisation, when conducted in an objective and rational manner, does not appear to skew the wider evidence base eligible studies.

There are important implications of this work. **First**, these results demonstrate that prioritised studies identified using a well thought-out third screening approach can provide a strong evidence base without compromising data integrity, whilst delivering more efficient SLRs. The prioritisation of literature may be particularly helpful in the following circumstances: 1) evaluating therapeutic areas with a large

amount of literature, 2) where the primary goal is to build knowledge and support decision-making within an organisation, and 3) when considering the highest quality data sources to inform future clinical study or model design. That said, the prioritisation of publications as above should be used with caution in situations when the findings are to be used for asset approvals, to support quantitative synthesis, and for guideline development. As stated clearly in the introduction, **no form of third screening should be used for SLRs focused on the comparative efficacy of treatments, where exhaustiveness of the data is the primary driver to obtain valid results.**

**Second**, this study calls for the need to formalize a method for prioritising publications for SLRs that focus on unmet need & disease burden by creating a validated framework that can be used and cited by the larger research community. Although based on a range of objective criteria such as sample size, geography and study design, the current method for prioritisation of studies is potentially rather straightforward and simplistic. Moreover, it has not been validated across multiple therapeutic settings.

The key **strengths** of our pilot study include the use of multiple quality criteria to conduct third screening and using two reviewers for data extraction to reduce the chances of errors. **Limitations** of this research should be noted. There was a varying degree of missing data in the included studies with respect to different parameters. Specifically, the studies focusing on the epidemiology or economic burden of wAMD omitted reporting data on certain patient

characteristics such as average age, visual acuity or treatment status at baseline. This research is based on data collected for only one therapeutic area and one SLR, so significantly more research is required to confirm these initial findings. It is worth noting that the key patient characteristics that define a target population could change with every therapeutic area and should therefore be defined *a priori* when assessing representativeness.

In **summary**, the use of a third screening approach to prioritise literature can potentially offer significant efficiency gains without compromising the integrity of the data. Whilst noting the pilot nature of this work, it appears that with limited budgets and growing time pressures to make decisions, third screening can be an effective way to derive valid findings from high-quality studies especially when the volume of the eligible literature is significant.

**Authors:** Digant Gupta  
(digant.gupta@bridgemedical.org) and Smeet Gala

## References

1. Borah, R., et al., Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open*, 2017. **7**(2): p. e012545.
2. Betts, M., et al., What's the Burden of Burden of Illness Reviews? *Value in Health*, 2018. **21**: p. S229.
3. Schmider, E., et al., Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution. *European Research Journal of Methods for the Behavioral and Social Sciences*, 2010. **6**(4): p. 147-151.
4. Dunn, O.J., Multiple comparisons among means. *Journal of the American statistical association*, 1961. **56**(293): p. 52-64.
5. Bland, J.M. and D.G. Altman, Multiple significance tests: the Bonferroni method. *Bmj*, 1995. **310**(6973): p. 170.

## Appendix 1: Statistical Analysis

An independent samples t-test (parametric) or a Mann-Whitney test (non-parametric) was used to compare patient characteristics between the prioritised (n=37) and de-prioritised (n=163) samples depending upon the underlying distribution of the variables. In order to inform the choice between parametric and non-parametric procedures, the normality of the variables was assessed using multiple sources of information, such as a histogram (with a superimposed normal curve), Shapiro-Wilk test, skewness and kurtosis z-values, and Normal Q-Q Plot. Note that the t-test is rather robust to deviations from normality, therefore, the absolute values of skewness and kurtosis were evaluated. If they were less than |2.0| and |9.0| respectively, the assumption of normality was assumed to be satisfied [3].

Since multiple tests of differences were carried out between the prioritised and de-prioritised samples, a Bonferroni

correction was applied to the p-values in order to keep the familywise type I error rate at approximately 0.05 or less [4, 5]. The Bonferroni correction was applied by dividing the conventional alpha level (0.05) by the number of statistical analyses performed (n=3, one for variable category: demographic, disease severity and treatment-related), resulting in a Bonferroni-adjusted alpha of 0.016. Then, the statistical analyses were performed and the obtained p-values from those analyses were compared against the adjusted critical alpha level of 0.016. All analyses were two-tailed, and a difference was considered statistically significant if the p-value was  $\leq 0.016$ . In addition to statistical significance (which is primarily a function of sample size), any differences between the prioritised and de-prioritised studies were also assessed for their clinical or practical significance. All data were analysed using SPSS version 28.0 (IBM, Armonk, NY, USA).

**Appendix 2: Study characteristics***Table 1: Key characteristics of the studies found to be eligible after FTS (N=200)*

Study Characteristic	Categories	Number (Percent)
<b>By study design</b>	Case-control	11 (5.5)
	Cross-sectional	55 (27.5)
	Prospective cohort	60 (30.0)
	Retrospective cohort	74 (37.0)
<b>By geography</b>	North America	52 (26.0)
	Europe	93 (46.5)
	Asia-Pacific	42 (21.0)
	Multi-continent	12 (6.0)
	South America	1 (0.5)
<b>By year of publication</b>	2010-2011	25 (12.5)
	2012-2013	28 (14.0)
	2014-2015	40 (20.0)
	2016-2017	45 (22.5)
	2018-2020	62 (31.0)
<b>By sample size</b>	<100	61 (30.5)
	100-199	37 (18.5)
	200-499	37 (18.5)
	500-999	18 (9.0)
	>=1000	47 (23.5)