Scientific quality issues in the design and reporting of bioscience research: a systematic study of randomly selected original *in vitro*, *in vivo* and clinical study articles listed in the PubMed database

Julija Baginskaitė

Name of supervisor:

Prof Malcolm Macleod

ABSTRACT

Over the past years concerns about scientific research quality in the biomedical research has gained a special attention. Specifically, low credibility and poor reporting quality of experimental findings has been speculated to be introducing a roadblock towards an efficient translation of preclinical discoveries into new treatments. The primary aim of this investigation was to systematically review a randomly selected sample of a major population of life science research reports in order to evaluate reporting of bioscience research and gain an insight into the quality of its study design. A total of 336 primary research publications across different study settings (*in vitro*, *ex vivo*, *in situ*, *in silico*, clinical studies) were assessed. Prevalence of the reporting of quality measures known to reduce bias in research was found to be very low. This investigation identifies with confidence that appropriate bioscience research reporting is lacking and encourages the development of new strategies that could be used to improve the quality of study reporting, design and conduct.

INTRODUCTION

"But the reason I call myself by my childhood name is to remind myself that a scientist must also be absolutely like a child. If he sees a thing, he must say that he sees it, whether it was what he thought he was going to see or not. See first, think later, then test. But always see first. Otherwise you will only see what you were expecting." Douglas (2001)

Over the last twenty years concerns about scientific research quality in the life sciences has gained a special attention. Several studies (Kilkenny *et al.*, 2009; Vesterinen *et al.*, 2011; Watters *et al.*, 1999; Yuan *et al.*, 2011) have identified that in multiple scientific publications reporting of *in vivo* and *in vitro* experimental design, conduct and outcome lack rigour. Multiple authors have discussed this issue, some suggesting that this may be the cause of currently devastatingly witnessed translational difficulties in the development on novel treatments (Drummond *et al.*, 2011; Hartshorne *et al.*, 2012; MacCallum, 2010; Macleod *et al.*, 2009; Osborne, 2011; Polyzos *et al.*, 2011; Pullen *et al.*, 2011; Zivin, 2008). Importantly, errors and omissions can potentially introduce difficulties to an adequate appraisal of studies by an independent reader and potentially provoke questioning of the validity of results. Producing valid results is a priority of a researcher as findings allow hypothesis to be evaluated. It is also apparent that the role of reports in scientific research is commonly underestimated (Michael and Amin 2002) however it could be argued that the scientific report itself is an integral part of research process and should be viewed as an ultimate goal of scientific inquiry (Bredan *et al.*, 2006).

Lack of scrupulous methodological measures in experimental design, conduct and reporting may introduce biased results, which might increase systematic variation between observed and true values of experimental findings (Weisberg, 2010). Any type of experimental design (from initial observation through to the dissemination of findings in the literature) is susceptible to dangers of different sources of bias such as design bias, selection bias, procedural bias, measurement bias, reporting bias to name a few, all of which may be overcome with careful considerations (Segall *et al.*, 2011; Weisberg, 2010).

Additionally, reproducibility of an experiment by an independent investigator may not be possible without appropriate reporting of findings and methodology. Recent study by German pharmaceutical company has directly identified serious reproducibility problems (Mullard, 2011) where it was found that 65% of drug target-validation projects were not reproducible. Recent works by Stewart et al (2012) and Siegel (2011) discusses prevalent issues of reproducibility in detail. One, however, cannot discard the possibility that

variations in independent findings may be due to the complexity of modern technology which varies across different countries and laboratories and not due to the fallacy of incomplete evidence *per se*.

Omissions in study reports and high prevalence of bias can also lead to inaccuracies in systematic reviews and meta-analyses (Hart *et al.*, 2012). This is particularly unfavourable as these studies are thought to be at the top of the evidence hierarchy and are specifically important for the evaluation of therapeutic interventions and drugs. Systematic reviews and meta-analyses are especially useful during modern times when increasing amounts of bioresearch data is being generated at an accelerated pace. Publication bias, a tendency towards reporting of positive results rather that negative, is another equally sensitive issue found to be prevalent across life-science domains (Korevaar *et al.*, 2011; Sandercock, 2012; Sena *et al.*, 2010; Weisberg, 2010).

Some approaches to increase aspects of scientific validity have been advised. For instance, certain scientific journals have decided to introduce a *Negative Results* section in order to limit prevalence of publication bias (Dirnagl *et al.*, 2010). An interesting suggestion was recently reported by Hartshorne et al (2012) where authors argue that systematic collection of evidence for/against reproducibility of already published study findings could aid identification of reliable results.

Importantly, for clinical studies and investigational *in vitro* and *in vivo* research, measures have been recommended to directly or indirectly overcome intrinsic and extrinsic research quality issues. The CONSORT (Consolidated Standards of Reporting Trials) statement (2010) advices on proper reporting of clinical trials (Schulz *et al.*, 2010), ARRIVE (Animal research: Reporting *in vivo* experiments) guidelines provide guidance on complete and transparent reporting of *in vivo* animal research (McGrath *et al.*, 2010), The Good Laboratory Practice Guide suggests measures to limit effects of bias in animal research (Macleod *et al.*, 2009), the Good Cell Culture Practice (Coecke *et al.*, 2005) advices on *in vitro* experimentation and the Gold Standard Publication Checklist offers consultation on the accurate design and conduct of animal studies (Hooijmans *et al.*, 2011), all of which are useful and freely accessible. However, it is difficult to estimate what proportion of researchers consults these.

All life science domains are affected by issues discussed here. However, success of pharmacological experimentation is particularly dependent on them, given the pressure

that many pharmacological investigations are facing. Findings of preclinical *in vitro* and *in vivo* therapeutic drug evaluation studies may be collectively used to influence the decision for candidate drugs to be assessed in clinical studies. Importantly, false positive results may lead to an over-estimation of an agent effect size. As a consequence, this may result in inappropriate clinical trials. Several systematic reviews and meta-analyses have identified a relationship between poor study quality and overestimation of agent effect size in animal models of disease (Crossley *et al.*, 2008; Pullen *et al.*, 2011; Sena *et al.*, 2010).

Aims

It appears that certain issues of quality in life science research have only been addressed in any detail in *in vivo* studies and clinical trials. With this in mind, the primary aim of this investigation was to systematically survey randomly selected life science research reports in order to assess with rigour whether there is evidence for reporting quality issues across the entire, general population of experimental bioscience study reports. To our best knowledge, this is the first study of such scope and design.

METHODS

Strategy for the sample selection

In order to limit any potential omission and inclusion bias (i.e., sampling bias), research papers for quality assessment were selected randomly, utilising an algorithm developed by Dr Malcolm Macleod, The University of Edinburgh (2012). The algorithm was used to retrieve citations from PubMed, a freely accessible database of biomedical research abstracts and related literature. The PubMed database was selected over other databases providing academic citation indexing for methodological accessibility purposes as PubMed assigns each publication with a unique identifier number PMID. Combined, PMIDs constitute a numerical population of identifiers of which a random sample can be retrieved. A random generation of 100 existing PMIDs was initiated by employing computational means of Microsoft Excel randomization function. A corresponding random selection of citations was retrieved from a database for a first, preliminary investigation. Characterization of the random sample was conducted and the efficiency of the methodology was evaluated. Subsequently, 1000 citations were randomly retrieved from PubMed for the appraisal in this study (Figure 1).

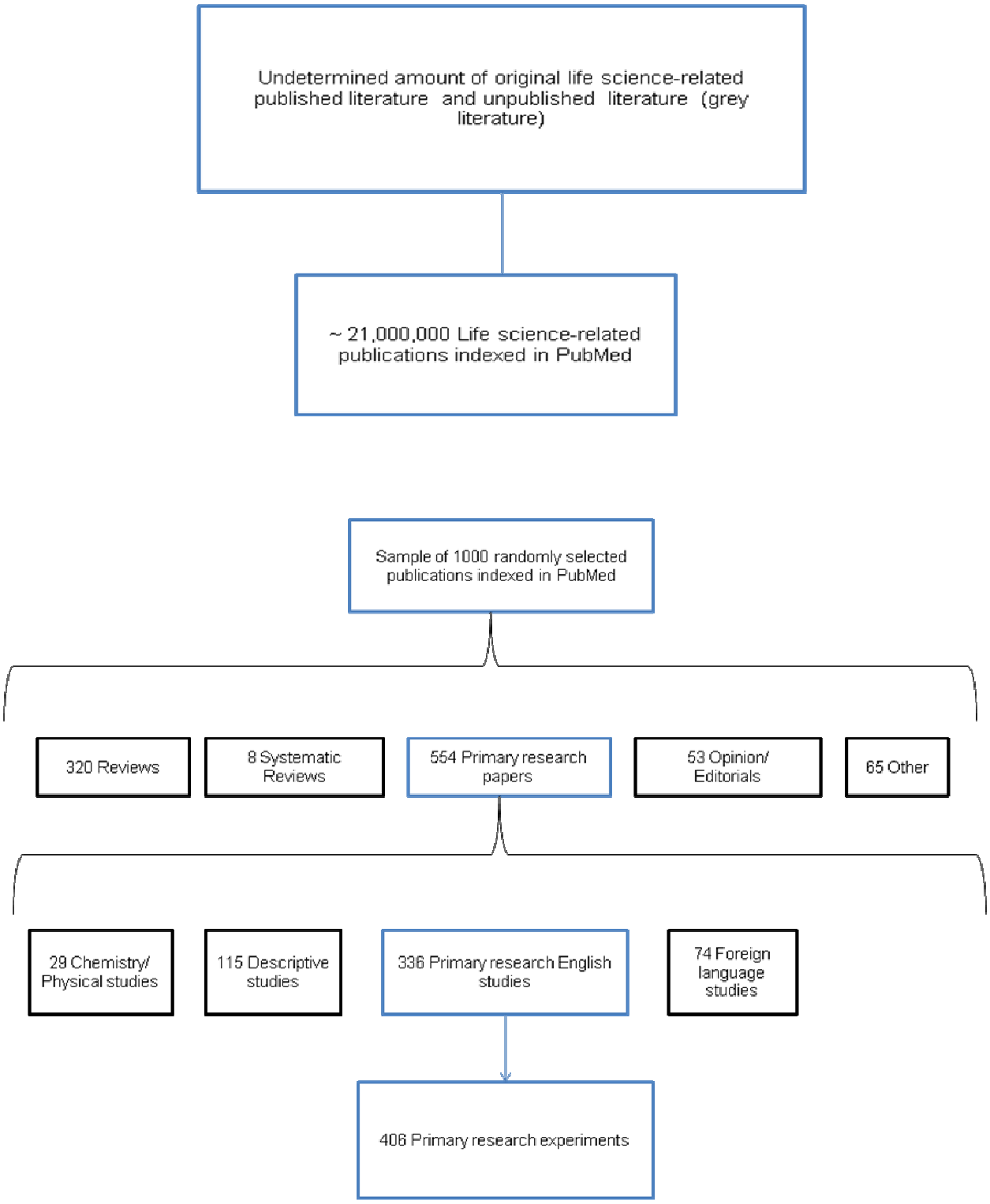

Figure1. Characterization of a sample population consisting of 1000 randomly selected scientific publications. In the end, 406 primary research experiments were identified for a full systematic review

Inclusion criteria for a full systematic study

Each publication was categorized as either primary research, review, systematic review,

opinion/editorial or other (newsletters, commentaries, biographies, bulletins, fact sheets, summaries). Publications were excluded if they were written in a non-English language, if they reported chemistry- or physics-related research or if they were descriptive research reports, case reports, developmental studies, correlation studies, retrospective studies or surveys (Figure 1). All primary research papers reporting findings of experiments in English were included in this systematic review. More than one experiment reported in a publication was analysed if the experiments differed methodologically or were performed in different settings (human, animal, *in vitro* etc).

Criteria for a quality of study

In order to evaluate the quality of selected study reports, each primary research publication underwent a detailed search for any information which could indicate inclusion or omission of fundamental features of appropriate study reporting. (Table1). It is recognized that different types of studies (*in vitro, in vivo, ex vivo, in silico, in situ,* clinical) might vary substantially in the ways in which they are conducted due to differences in materials, methodological applicability and varying degrees of complexity. Therefore, comparison between them is not straightforward. However, it is agreed that what every study has in common is an intrinsic threat of different types of bias that can be overcome by taking certain measures. Therefore, for every relevant study, prevalence of six different quality measures known to reduce bias and four characterization values were amended from previous research (Macleod *et al* (2009) and McGrath *et al* (2010)) and applied in this study (Table 1). Importantly, one should note the difference between concealment of allocation sequence and blinded conduct of an experiment, where blinded conduct of an experiment could be conducted by an independent investigator even if there were no two subjects or preparation groups to be compared in-between or if groups were phenotypically different thus precluding randomized allocation to groups.

For each measure a value of “Yes”, “No” or “Not Applicable” was assigned, depending on whether the measure was reported and if not – whether it was applicable at that particular experimental setting. The decision to assign a value of “Not Applicable” was based on logical reflection, previous experience and academic knowledge.

Experiments were described as observational if the investigator was not utilising interventional measures and they were described as experimental if a preparation, model or subject studied was subjected to an intervention. A selection of main disciplinary or

interdisciplinary domains to be assigned to each experiment was determined arbitrarily.

A.

Quality measures
○ Sample size calculation
○ Randomization
○ Concealment of sequence allocation
○ Blinded conduct of an experiment
○ Blinded ascertainment of outcome
○ Conflict of interest statement

B.

Setting	Basis	Design	Domain
Human	In vivo	Experimental	Pharmacology
Animal	In vitro	Observational	Neuroscience
Plant	Ex vivo		Endocrinology
Microbe	In silico		Immunology
Other	Other		Cardiovascular
			Oncology
			Aging
			Neuropsychology
			Bacteriology
			Virology
			Haematology
			Cellular/molecular
			Physiology
			Genetics
			Public Health
			Other

Table 1. A) Six quality measures assessed for each study report. Adapted from Macleod et al (2009) and McGrath et al (2010). B) Characterization values assigned to each study.

Statistics

Confidence intervals at 95% were calculated for each of the findings on the prevalence of each quality measure. Sample size calculations were performed to assess to what extent our randomly selected study sample would be a representative of the whole life science-related original literature.

RESULTS

Sample characterization

Of a randomly selected sample, 55% of articles were identified as primary research, 32% as reviews, 5% as opinion/editorial articles, 1% as systematic reviews, and 7% as other publications.

Interestingly, the number of retrieved publications was found to be increasing in time (Figure 2).

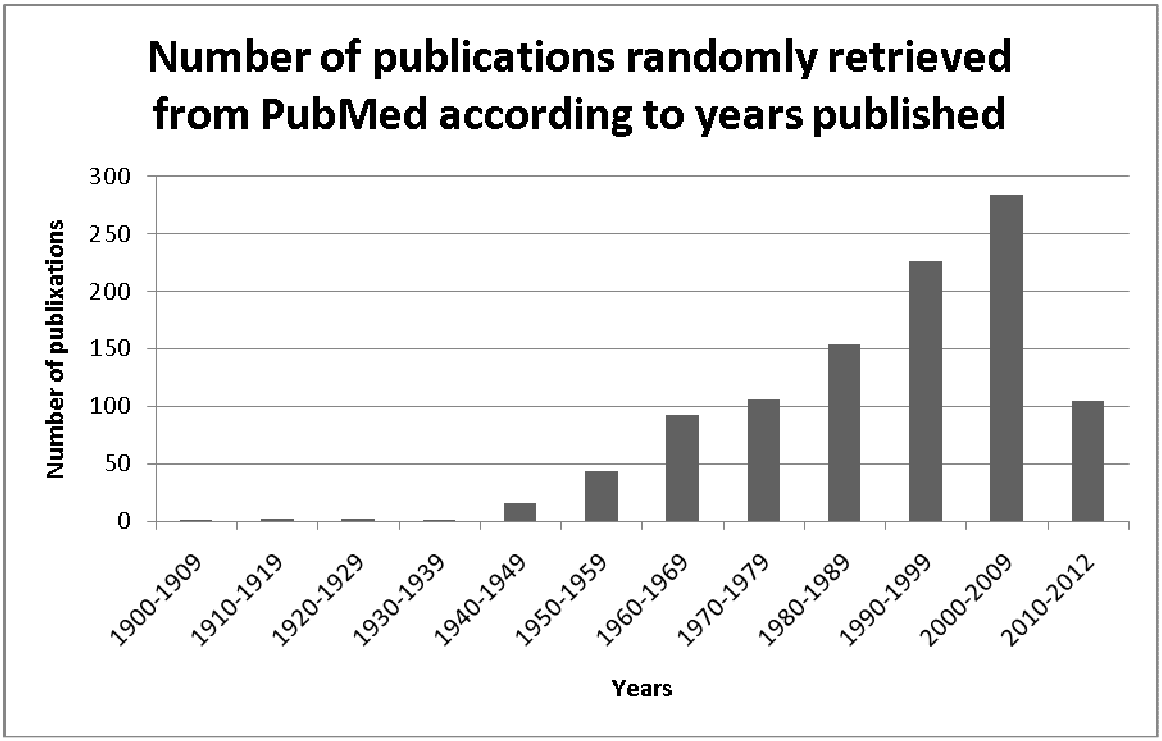

Figure 2. Number of publications randomly retrieved from PubMed and published over years. Note: the right corner year interval is covering only two years (2010-2012)

Of 554 primary research articles, 5% were from fields of chemistry and physics, 13% were non-English studies, 21% were descriptive studies and 61% (336/554) were primary research non-descriptive studies published in English, the latter representing a proportion that was included in this systematic review.

This systematic study includes 336 relevant articles representing 185 *in vitro* (46%), 130 *in vivo* (32%), 49 *ex vivo* (12%), 5 *in silico* (1%) and 37 *other* (9%) experiments. Amongst these, 159 experiments were conducted in humans or human cells and tissues (39%), 174 in animals or animal cells and tissues (43%), 60 in microorganisms (15%) and 13 in plants or plant cells ant tissues (3%) across different life-science fields (summarised in Table 2 and Table 3). Reporting of quality measures across various life-science fields was not found to differ to any great extent.

		Sample size calculation		Randomized allocation		Allocation concealment		Blinded conduct of experiment		Blinded ascertainment of outcome		Conflict of interest statement	
n		Yes		Yes		Yes		Yes		Yes		Yes	
96	pharmacology	1		14		0		2		3		11	
43	cellular/molecular biology	0		3		0		0		0		0	
38	oncology	0		2		0		0		0		5	
34	microbiology	0		1		0		0		0		2	
29	other	0		7		0		0		0		3	
28	public health	2		7		0		0		0		3	
24	neuroscience	0		3		1		0		1		3	
23	cardiovascular research	0		2		0		1		1		4	
20	immunology	0		0		0		0		0		2	
18	physiology	0		4		0		0		2		2	
17	genetics	0		0		0		0		0		2	
13	endocrinology	0		4		0		0		0		1	
13	virology	0		0		0		0		0		2	
10	haematology	0		1		0		0		0		0	

Table 2. A number of experiments conducted across different life-science fields and prevalence of 6 quality measures in reports included in the systematic assessment.

	Sample size calculation		Randomized allocation		Allocation concealment		Blinded conduct of experiment		Blinded ascertainment of outcome		Conflict of interest statement	
	Yes	N/A	Yes	N/A	Yes	N/A	Yes	N/A	Yes	N/A	Yes	N/A
Human n = 61	1	5	12	39	0	46	3	42	5	12	11	-
Animal n = 62	0	0	11	12	0	15	0	12	1	2	4	-
Plant n = 6	0	0	1	0	0	2	0	2	0	0	0	-
In vitro n = 186	1	1	9	59	0	82	0	43	0	5	14	-
Ex vivo n = 49	0	2	5	19	1	27	0	16	1	3	4	-
In silico n = 5	0	0	1	2	0	4	0	4	0	0	2	-
Other n = 37	1	0	10	13	0	24	0	20	0	8	5	-

Table 3. Prevalence of 6 quality measures across different experimental settings reported in the selected random sample.

Characterization of experimental settings

Overall, it was found that the most highly prevalent measure across different study settings of different years was randomization (Figure 3). Randomization was described least commonly in *in vitro* studies and its reporting prevalence was higher in *in vivo* human and animal as well as *in situ* studies (Table 3).

The highest prevalence of all quality measures was found in studies published in the year block 2000-2012 as compared to other year intervals (Figure 3).

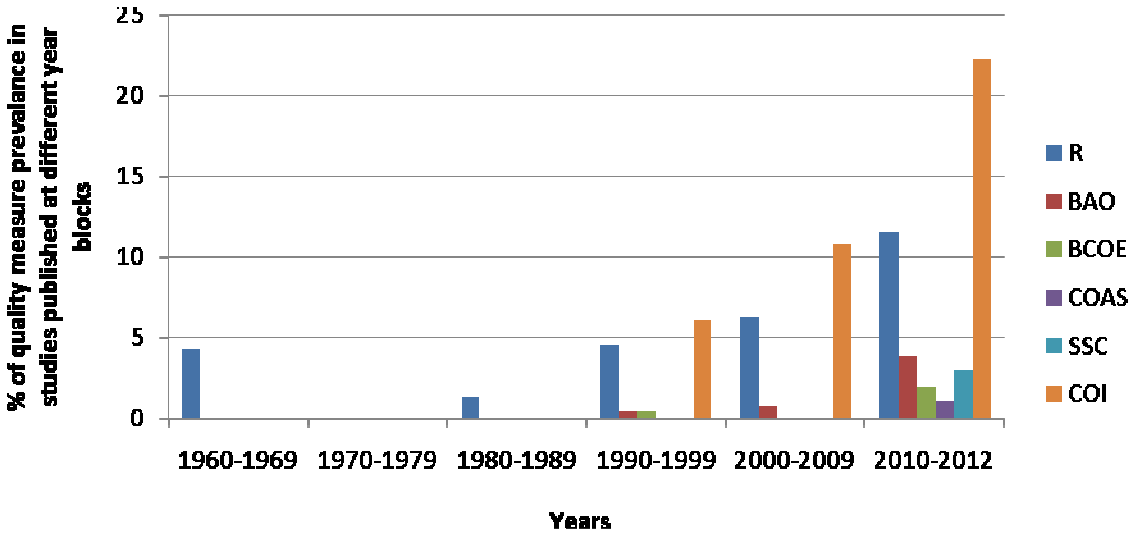

Figure 3. Prevalence of reporting quality measures across blocks of years. R = Randomization, BAO = Blinding ascertainment of outcome, BCOE = Blinded conduct of experiment, COAS = Concealment of allocation sequence, SSC = Sample size calculation, COI = Conflict of interest

Study quality

Only 1% (95% CI = 10.9) of publications reported a sample size calculation where applicable. In 8 studies, however, sample size calculation was not possible (Table 3).

12% (95% CI = 8.5) of studies reported randomized allocation to groups or an attempt to select a random sample from a population studied and thus increase internal and external validity respectively. Strikingly, in 15% study settings randomized allocation to groups was found to be inapplicable because there was no control group reported in that experimental setting even though the use of control group was possible.

Prevalence of the reporting of blinded ascertainment of outcome was found to be very low: 2% (95% CI = 10).

Reporting of concealment of allocation was found in only 1 study of neuroscience discipline, *ex vivo* basis, published in 2012: 0.2% (95% CI= 20)

In only 1% (95% CI = 11) of experimental settings an investigator was reported to be blinded towards control and experimental groups or towards expectations of particular experiment.

10% (95% CI = 9) of all studies reported a statement of conflict of interest. 82% of these declared to have no conflict of interest.

336 publications were reviewed here. This sample size was assumed to be of truly random nature and was calculated to represent entire life science literature indexed in PubMed population with 95% CI = 5 and with 99% CI = 7.

DISCUSSION

Integrating findings across different experimental types and settings revealed recurrent and prevailing tendencies. Marked omissions of reporting of all six quality measures were observed. This observation is consistent with previous findings of similar studies (Kilkenny *et al.*, 2009; Vesterinen *et al.*, 2011; Watters *et al.*, 1999). Prevalence of poor reporting of quality measures appears to be universal across different experimental settings and scientific disciplines. However, even if substantial differences of reporting quality were observed for different disciplines, generalized assumption that certain disciplines perform more robust research would not be sensible simply from this particular systematic investigation.

Similarly, findings of current investigation do not, by any means, suggest that experiments investigated here are of poor quality. However, prevalence of quality measures is considered to give insight into the quality of experimental design, conduct and reporting (Macleod *et al.*, 2009). One could argue that a suspicion may be provoked when evaluating credibility of results, given the lack of reporting of whether attempts were made to reduce the likelihood of potential bias in a study. Appropriate reporting may spare time and resources of potential future investigators interested in replicating or continuing a specific study.

It was also observed from the characterization of our random sample that the overall proportion of quality measures reported is higher in studies published at the year block ranging from 2000 to 2012 as compared to previous years. This might indicate that the quality of reporting might be improving. However, it is arguably still very poor.

Interestingly, reporting of *in vitro* studies was found to be least robust. This finding raise a possibility that some *in vitro* environments where tissue and cell culture preparations or reagents are being tested might not be as tightly controlled as for example *in vivo* studies. *In vitro* studies are often viewed at the base of the hierarchy of evidence. However, even

from the current systematic investigation it appears that *in vitro* systems constitute a high proportion of all experiments conducted, therefore generating high amounts of data that in many cases is being interpreted for further studies or replications. The role *in vitro* studies in pharmacological setting is immense (Coecke *et al.*, 2005) and the use of *in vitro* systems might be predicted to increase, given the sophistication of modern technology and the desirability for the reduction of animal use in life sciences.

In vitro, ex vivo, in vivo, in silico, in situ studies may all be reciprocally related while investigating, for instance, certain questions of complex disease pathways, toxicology, etc. With this in mind, lack of clarity, transparency and robustness in either reporting or design/conduct of an experiment may generate a "chain reaction" of misinterpretations and possibly waste of resources and time. The quality of *in vitro* study reporting as well as reporting of other type of studies needs to be improved.

Interestingly, prevalence of one of the quality measures, that is disclosure of competing interest, seems to be increasing. This might be due to the fact that more journals are integrating conflict of interest statement as a routine practice. If disclosed genuinely, it may facilitate assessment of scientific integrity in the original bioscience work (Bekelman *et al.*, 2003).

Prevalence of reporting of randomization appears to be spread rather similarly across the course of years as compared to other measures. The finding is not surprising, given that this simple measure can substantially increase either intrinsic or extrinsic validity (or indeed both) (Weisberg, 2010). Also, there might be the case that randomization might have been detected to be reported (described) more often than other measures due to a relatively consensus language that might be used describe randomization.

Particularly low occurrence of reporting of sample size calculation or of any blinding measures is very disappointing. If experiment is conducted in an unblinded manner, the likelihood of investigator committing fallacy of incomplete evidence is increased (Weisberg, 2010). These findings should be treated as an alarm, encouraging transformations in the way science community is currently tackling threats of internal consistency and logic.

Justification for appropriateness of the database in this study

There is a strong argument in support that the findings of this study can be generalized to the whole population of bioresearch. The PubMed database from which our random

sample was retrieved comprises approximately 21,000,000 indexed life science-related records to date (U.S. National Library of Medicine, 2012) and therefore could be argued to be a good representation of the whole bioscience research population. Database coverage ranges throughout different life science fields. Interestingly, a calculation was performed on information retrieved from PubMed database for the last ten years and it shows that approximately 633,296 new indexes are being added to PubMed database each year. In contrast, World of Knowledge citation database is commonly believed to comprise more comprehensive citation indexes than PubMed and currently holds approximately 50,000,000 records. However, indexes in Web of Knowledge database range through fields of Sciences, Social Sciences, Arts and Humanities and Chemistry, and evidently a proportion of its records is corresponding to a non-life science research.

Finding of a sensible degree of heterogeneity of studies in our random sample, as evident from Table 2, Table 3 and Figure 1, demonstrates that this random sample might be a good representation of the entire bioscience literature population. Another interesting observation that the current systematic investigation revealed is that the random sample of PubMed indexes of bioscience literature comprises increasing proportions of studies published over the years, with increasing tendency towards present times. As our sample is random, tendency observed should agree with tendencies observed regarding the growth of scientific publications. Indeed, this coincides with scientiometric study by Larsen et al (2010) where numbers of scientific publications have been shown to be steadily increasing over the last 100 years.

Limitations

However interesting the findings, this investigation has several limitations. Firstly, biomedical literature archive in the PubMed database does not contain the entire population of studies published by life science research community. Also, a great proportion of unpublished studies (grey literature) is an important part of bioscience research and should not be excluded. Moreover, in our study primary research papers in foreign languages were excluded due to time constraints. Altogether these factors introduce certain level of exclusion bias. However, it could be argued that introduction of bias here is rather small and does not affect accuracy of our findings to great extent but affects precision. In addition, it should be noted that this investigation was conducted by only one reviewer.

Future perspectives

Taking into account these limitations, another similar systematic investigation would be useful in drawing broader conclusions. Of interest would be a study of greater scope, greater precision (larger sample size) and one investigating the prevalence of more research quality measures considered to increase reliability of study discoveries. Ideally, two or more independent reviewers should conduct such investigation. It would also be useful to develop a method to retrieve a random sample from amore complete scientific literature coverage.

Conclusions

Taken together, findings of this systematic investigation are significant in that they confidently reveal the lack of transparent reporting across various settings of bioscience research. Information collected from systematic investigations like this should encourage and contribute to the development of effective strategies to overcome current roadblocks such as challenges in the translation of new therapies in the future. Development of systematic approaches towards robust assessment of the reliability of the original scientific work could be useful if the quality of experimental research was to be improved to great extent.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the great expertise, guidance, inspiration and contribution from my supervisor Dr Malcolm Macleod, kind advices from Hanna M. Vesterinen, constant support throughout the academic year from my Director of Studies Dr Andrew Hall, Pharmacology Honours Course organizer Prof James McCulloch and my father Prof Armantas Baginskas

CONFLICT OF INTEREST STATEMENT

Author declares no competing interest. An investigation was conducted out of mere curiosity.

REFERENCES

Bekelman JE, Li Y, Gross CP (2003). Scope and Impact of Financial Conflicts of Interest in Biomedical Research. *JAMA: The Journal of the American Medical Association* **289**(4)**:** 454-465.

Bredan AS, van Roy F (2006). Writing readable prose: when planning a scientific manuscript, following a few simple rules has a large impact. *EMBO Rep* **7**(9)**:** 846-849.

Coecke S, Balls M, Bowe G, Davis J, Gstraunthaler G, Hartung T*, et al.* (2005). Guidance on good cell culture practice. a report of the second ECVAM task force on good cell culture practice. *Altern Lab Anim* **33**(3)**:** 261-287.

Crossley NA, Sena E, Goehler J, Horn J, van der Worp B, Bath PMW*, et al.* (2008). Empirical Evidence of Bias in the Design of Experimental Stroke Studies. *Stroke* **39**(3)**:** 929-934.

Dirnagl U, Lauritzen M (2010). Fighting publication bias: introducing the Negative Results section. *J Cereb Blood Flow Metab* **30**(7)**:** 1263-1264.

Douglas A (2001). *So Long, and Thanks for All the Fish*. edn. Pan Macmillan.

Drummond GB, Vowler SL (2011). Show the data, don't conceal them. *British Journal of Pharmacology* **163**(2)**:** 208-210.

Hart B, Lundh A, Bero L (2012). Effect of reporting bias on meta-analyses of drug trials: reanalysis of meta-analyses. *BMJ* **344**.

Hartshorne JK, Schachner A (2012). Tracking replicability as a method of post-publication open evaluation. *Front. Comput. Neurosci.* **6**.

Hooijmans CR, de Vries R, Leenaars M, Curfs J, Ritskes-Hoitinga M (2011). Improving planning, design, reporting and scientific quality of animal experiments by using the Gold Standard Publication Checklist, in addition to the ARRIVE guidelines. *British Journal of Pharmacology* **162**(6)**:** 1259-1260.

Kilkenny C, Parsons N, Kadyszewski E, Festing MFW, Cuthill IC, Fry D*, et al.* (2009). Survey of the Quality of Experimental Design, Statistical Analysis and Reporting of Research Using Animals. *PLoS ONE* **4**(11)**:** e7824.

Korevaar DA, Hooft L, ter Riet G (2011). Systematic reviews and meta-analyses of preclinical

studies: publication bias in laboratory animal experiments. *Laboratory animals* **45**(4)**:** 225-230.

Larsen P, Ins M (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics* **84**(3)**:** 575-603.

MacCallum CJ (2010). Reporting Animal Studies: Good Science and a Duty of Care. *PLoS Biol* **8**(6)**:** e1000413.

Macleod MR, Fisher M, O'Collins V, Sena ES, Dirnagl U, Bath PM*, et al.* (2009). Good laboratory practice: preventing introduction of bias at the bench. *Stroke* **40**(3)**:** e50-52.

McGrath JC, Drummond GB, McLachlan EM, Kilkenny C, Wainwright CL (2010). Guidelines for reporting experiments involving animals: the ARRIVE guidelines. *British Journal of Pharmacology* **160**(7)**:** 1573-1576.

Mullard A (2011). Reliability of 'new drug target' claims called into question. *Nat Rev Drug Discov* **10**(9)**:** 643-644.

Osborne N (2011). Reporting animal research - worthy of a rethink? *Learn. Publ.* **24**(1)**:** 5-7.

Polyzos NP, Valachis A, Patavoukas E, Papanikolaou EG, Messinis IE, Tarlatzis BC*, et al.* (2011). Publication bias in reproductive medicine: from the European Society of Human Reproduction and Embryology annual meeting to publication. *Hum. Reprod.* **26**(6)**:** 1371-1376.

Pullen N, Birch CL, Douglas GJ, Hussain Q, Pruimboom-Brees I, Walley RJ (2011). The translational challenge in the development of new and effective therapies for endometriosis: a review of confidence from published preclinical efficacy studies. *Human Reproduction Update* **17**(6)**:** 791-802.

Sandercock P (2012). Negative results: why do they need to be published? *International Journal of Stroke* **7**(1)**:** 32-33.

Schulz KF, Altman DG, Moher D (2010). CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMJ* **340**.

Segall M, Chadwick A (2011). The risks of subconscious biases in drug-discovery decision making. *Future Med. Chem.* **3**(7)**:** 771-774.

Sena ES, van der Worp HB, Bath PMW, Howells DW, Macleod MR (2010). Publication Bias in Reports of Animal Stroke Studies Leads to Major Overstatement of Efficacy. *PLoS Biol* **8**(3)**:** e1000344.

Siegel V (2011). Reproducibility in research. *Disease Models & Mechanisms* **4**(3)**:** 279-280.

Vesterinen HV, Egan K, Deister A, Schlattmann P, Macleod MR, Dirnagl U (2011). Systematic survey of the design, statistical analysis, and reporting of studies published in the 2008 volume of the Journal of Cerebral Blood Flow and Metabolism. *J Cereb Blood Flow Metab* **31**(4)**:** 1064-1072.

Watters MP, Goodman NW (1999). Comparison of basic methods in clinical studies and in vitro tissue and cell culture studies reported in three anaesthesia journals. *Br J Anaesth* **82**(2)**:** 295-298.

Weisberg HI (2010). *Bias and Causation: Models and Judgment for Valid Comparisons*. edn: John Wiley & Sons.

Yuan JCC, Shyamsunder N, Barao VAR, Lee DJ, Sukotjo C (2011). Publication Bias in Five Dental Implant Journals: An Observation from 2005 to 2009. *Int. J. Oral Maxillofac. Implants* **26**(5)**:** 1024-1032.

Zivin JA (2008). Identifying Low-Quality Preclinical Studies. *Stroke* **39**(10)**:** 2697-2698.