BLINDED TITLE PAGE:

Do Radiological Research Articles Apply the Term “Pilot Study” Correctly?

Systematic Review.

INTRODUCTION

It has been suggested that radiological research exhibits less methodological expertise than other medical disciplines[1,2]. Methodological expertise is reflected partly by a clear understanding of different study designs and their specific purpose. To statisticians and methodologists, the term “pilot study” has precise implications, namely a “small-scale test of the methods and procedures to be used on a larger scale” subsequently[3]. Pilot studies therefore aim to evaluate the viability and applicability of critical research components and assumptions needed for larger studies to be successful. They play a pivotal role in mitigating unwanted surprises, saving cost and time when later undertaking large trials. They may prevent studies that are doomed to fail, for example due to inadequate recruitment. For these reasons, the UK Medical Research Council guidance on designing and evaluating complex interventions recommends that pilot studies be conducted before any definitive large-scale evaluation[4].

By virtue of their design, radiological pilot studies are not powered to determine diagnostic accuracy (or a “pilot study” would be unnecessary). Rather, their endpoints should revolve around metrics necessary to determine if a larger study/trial ought to proceed. The most obvious endpoint is recruitment rate, which tends to be over-optimistic when planning research. Test method feasibility in a trial setting and data capture and retention are examples of other common endpoints for pilot studies.

However, we have noted anecdotally that the description “pilot study” is used by radiological studies that instead report measures of diagnostic accuracy while often appearing underpowered. If so, this would imply that the researchers do not appreciate the precise implications of this study design and are using the term “pilot” to excuse

underpowering. Therefore, by systematic review we aimed to determine the proportion of radiological “pilot” studies that use this description correctly.

MATERIAL AND METHODS

Ethical approval

Our institution does not require ethical approval for systematic review of indexed literature.

Review question

We aimed to determine what proportion of radiological “pilot” studies genuinely used this study design.

Search strategy and study eligibility

Following investigator discussion, we identified studies from four well-recognised, representative, indexed radiological journals: Radiology (RADIOL), European Radiology (ER), American Journal of Roentgenology (AJR), and British Journal of Radiology (BJR). We decided to limit our review to four well-established, internationally visible journals believing that these would provide data representative of pilot study reporting in the radiological domain, and that little additional information would be gained by extending the search to all potential journals. Limiting the search allowed us to study each individual journal in greater depth. To be eligible for inclusion the study title had to include the word “pilot” and investigate human subjects (of any age). We elected to

exclude narrative and systematic reviews if encountered. We decided *a priori* that a maximum of 20 studies per journal (for a maximum sample size of 80) would achieve sufficient saturation for information to be representative, i.e. we anticipated that by that stage our findings would be clear and that a larger sample would add no further, useful information.

Search strategy and string

The senior author (BLINDED) searched The National Library of Science via PubMed using the following terms: “Radiology [ta] AND pilot”, “European Radiology [ta] AND pilot”, “American Journal of Roentgenology [ta] AND pilot”, “British Journal of Radiology [ta] AND pilot”. Articles were selected from present day (search performed 20th September 2018) retrospectively, without date restriction.

Search process and citation management

Potentially eligible citations were passed to a junior researcher (BLINDED) who then screened the electronic abstracts, discarding “clearly unsuitable” articles (e.g. “pilot” absent from the study title, radiotherapy studies, animal studies, interventions without imaging). An exclusion log was kept for all excluded papers (Figure 1). The junior researcher had no prior experience in performing systematic reviews and so was supervised closely by more senior members of the research team. Duplicate extraction of the first 20 articles was performed by another member of the team (BLINDED) to check consistency, which was acceptable. The remaining extraction was performed by the junior researcher thereafter. Any uncertainty was discussed face-to-face with the

senior author. Potentially eligible citations were entered into a datasheet designed specifically for the review (Microsoft Excel for Mac 2011 v. 14.5.9, Microsoft Corporation, Washington).

Data extraction

The junior researcher extracted data from component primary studies into the datasheet using the following categories: Citation; Prospective/retrospective design; Single/multi-centre; Total subjects recruited; Sample size rationale and/or power calculation stated (yes/no); Primary endpoint stated explicitly (yes [description]/no); Secondary endpoints stated explicitly (yes [description]/no); If no primary endpoint stated explicitly, what appeared to be the primary endpoint; Did the endpoint suggest a pilot study (yes/no); Were diagnostic test accuracy data presented (yes [description, e.g. sensitivity/specificity]/no); Ultimately, in the Reviewer's opinion, was the study a genuine "pilot" (yes/no [description]). We judged studies to be genuine pilots if it was a, "small-scale test of the methods and procedures to be used on a larger scale" subsequently[3]. For example, the design could potentially investigate recruitment rate for a subsequent trial, intervention allocation (e.g. by randomisation) and acceptability (i.e. willingness to be randomised), assessment procedures, data retention, etc. If believed to be another study design, the nature of this was recorded. Any uncertainty was resolved via face-to-face discussion with other members of the team.

Statistical analysis

We reported our review according to PRISMA guidelines[5]. No risk of bias assessment was performed because methodological quality of component studies was not our primary concern and meta-analysis (where risk of bias is used to assess precision of the point estimate) was not anticipated. Data were expressed as simple frequencies and proportions.

RESULTS

The PRISMA flow diagram for studies identified by the systematic review is shown in Figure 1. The search string identified a total of 658 records over the four targeted journals as follows: 236 RADIOL, 158 ER, 146 AJR, 118 BJR. Our target of 20 consecutive articles was achieved for all journals with the exception of BJR, where only 18 articles were eligible from the 118 identified, giving a total sample size of 78. The 78 component articles selected for the systematic review are listed in Appendix 1.

55 (70.5%) studies were prospective, 21 (26.9%) retrospective and 2 (2.6%) mixed. The overwhelming majority of studies were conducted in single centres; 76 (94.7%). 62 studies (79.5%) investigated a single imaging modality (including 26 MRI; 18 CT; 12 ultrasound; 2 PET-CT; 2 PET-MRI), 13 studies (16.7%) investigated multiple imaging modalities, two studies investigated radio-frequency ablation and a single study investigated a biliary stent.

5,572 patients were reported across 77 studies; the remaining study reported individual lesions only and omitted the number of individual patients[6]. Median sample size per study was 20 patients, range 7[7] to 1666[8]. The majority of studies (55, 70.5%) stated no rationale for their sample size. Furthermore, no study presented a power calculation to justify sample size. A primary endpoint was stated explicitly by 70 (89.7%) studies and could be inferred in all the remaining 8 studies. Secondary endpoints were stated explicitly by 20 (25.6%) studies.

Ultimately, we judged that no individual study qualified as a genuine pilot study when assessed against our *a priori* criteria. The individual study types encountered are described in Table 1. Notably, 66 (84.6%) studies presented measures of test accuracy and were framed as studies of diagnostic test accuracy. 12 studies presented elements of feasibility, and 8 elements of technology assessment (Table 1).

DISCUSSION

This systematic review investigated our anecdotal observation that most radiological studies describing themselves as “pilot” studies are actually small, underpowered studies reporting diagnostic test accuracy. Our hypothesis was correct; not one of the 78 studies ultimately included in our systematic review satisfied our *a priori* criteria for a genuine “pilot” study, i.e. a “small-scale test of the methods and procedures to be used on a larger scale” subsequently[3]. As predicted, we found that the greatest proportion of studies described as pilots actually attempted to investigate diagnostic accuracy,

frequently presenting inferential statistics and performing hypothesis testing. Furthermore, while good studies of test accuracy should be powered sufficiently to estimate diagnostic performance (for example sensitivity and specificity, and/or area under the receiver operating characteristic curve), no study presented a power calculation.

While the majority of study designs we encountered were focussed on diagnostic test accuracy, we did encounter other designs and it is worth discussing the distinction between these and true “pilot” studies. We found that “feasibility studies” either alone or in combination with other study types were the second largest group identified. However, while the terms “feasibility” and “pilot” study are often used interchangeably, this is erroneous since they describe different study designs. The UK National Institute for Health Research defines a pilot study as, “a smaller version of the main study used to test whether the components of the main study can work together. It is focussed on processes of the main study, for example to ensure that recruitment, randomisation, treatment, and follow-up assessments all run smoothly”[9]. A pilot study, therefore, does not aim to test whether the primary intervention is effective. Where uncertainty exists regarding whether a radiological test “works” or not, i.e. does it measure (or do) what is intended, then the appropriate design is a “feasibility” study.

Feasibility studies examine, “Characteristics of the proposed outcome measure and in some cases feasibility studies might involve designing a suitable outcome measure”[9]. A radiological example might be a study to determine whether diffusion weighted MR

identifies cancer cells. However, the sensitivity and specificity with which that is achieved can only be determined via a properly powered diagnostic test accuracy study, which might require a “pilot” beforehand to see if adequate numbers of representative patients can be accrued. Feasibility studies ask, “can this study be done?”[9], and have a narrower focus than pilots. Notably, while “Feasibility studies do not evaluate the outcome of interest”[9], pilot studies do so but not in numbers sufficient for a properly powered study of diagnostic accuracy. However, pilots may contribute individual data to a subsequent main trial, in which case they are termed “internal” pilots (vs, “external” pilots, which exclude their data from any main trial).

We also encountered designs that aimed to investigate whether a specified technology worked for its intended purpose or not, which can be termed, "technology assessment". In its broadest sense, "health technology assessment" can incorporate evaluation of intended and unintended effects of an intervention, often including clinical and cost effectiveness. "Efficacy assessment" study designs attempt to determine the relative balance of benefits versus harms and are often concerned about safety. We identified one “incidence study”, described as a “pilot”[10]. Incidence studies are epidemiological designs that measure incidence of an outcome, myocardial infarction in this case, in a population following their exposure[11].

An initial objective of our research was to investigate the sample size of radiological pilot studies in order to determine how closely they matched general recommendations for pilot sample sizes (hypothesising that they would be too small generally) but our

failure to identify genuine pilots precluded this. Notwithstanding this, we found a median sample size of just 20 individuals recruited and 75% of studies described samples below 40. As noted previously, no article presented a power calculation to justify their sample. Several recommendations to estimate pilot sample size are available, for example using a confidence interval around the anticipated standard deviation[12], or 3% the anticipated size of the definitive trial[13]. Several other alternatives have been proposed[14-16]. The sample sizes we observed were largely below these recommendations. Furthermore, the few authors presenting a rationale for sample size, usually based this on practicality and convenience rather than factors important to plan a future definitive study. Even those researchers who argue that sample size calculation may not always be necessary, agree that some justification should always be given by authors[17]. It has been argued frequently that reporting underpowered research is unethical because such studies can encourage clinical practice based on invalid results[18-21]. While it is self-evident that all genuine pilot studies must be prospective, we nevertheless identified 21 purely retrospective studies. It is also self-evident that planning for a subsequent large trial might occur over several centres but the overwhelming majority of studies in our review were single centre.

While we believe ours is the first study to investigate pilot studies of radiological interventions, other authors have found this term misused commonly by other medical disciplines. For example, Shanyinde and co-workers randomly sampled 50 medical studies described as pilots and found that the large majority actually focussed on efficacy[22]. Kannan and Gowri investigated 93 Indian “pilot” studies, finding, like us, that

none satisfied this description[23]. Similarly, Arain and co-workers concluded that, “Pilot studies are still poorly reported, with inappropriate emphasis on hypothesis testing”[24]. While it could be argued that Reviewers (and certainly Journal Editors) will have more knowledge of study design and terminology than the average researcher, our (and others’) data suggest that deficiencies in study description are not being rectified after acceptance for publication.

Our study does have limitations. Most obviously, we considered only those articles using “pilot” in their title. It is possible that a study may be declared a “pilot” elsewhere, for example in the methods section. For example, reporting of a large scale multi-centre study may conceivably incorporate an internal pilot. However, it has been suggested that the large majority (87%) of studies describing themselves as pilots do so in their title[25]. It may have been informative to identify how many “pilot” studies were followed by a subsequent larger trial but our failure to identify genuine pilots diminished our enthusiasm to do this. *A priori*, we decided to restrict our search to four prominent radiological journals, hypothesising that if the issue of misinterpretation was a problem for major journals (who we might assume enjoy greatest reviewer and Editorial skills), then it would be plausible to assume the problem would be even greater for “lesser” journals. Restricting a systematic search to specific journals is a well-recognised methodological tactic where this facilitates the review while being unlikely to impact the outcome. Lancaster and co-workers[26] restricted their search of pilot studies preceding randomised controlled trials to seven journals, a tactic also adopted by Arain and co-workers[24].

In summary, by systematic review we observed that, in reality, radiological “pilot” studies mostly report underpowered studies of diagnostic test accuracy. In order to have scientific credibility, we encourage authors, reviewers, and Editors of radiologic journals to familiarise themselves with different methodological study designs and their precise implications.

REFERENCES

1. Dwyer AJ, Doppman JL, Black WC. The poor quality of early evaluations of MRI. JAMA 1988;260:2661-4.

2. Smidt N, Rutjes AW, van der Windt DA, et al. Quality of reporting of diagnostic accuracy studies. Radiology 2005;235:347-53.

3. Thabane L, Ma J, Chu R, et al. A tutorial on pilot studies: the what, why and how. BMC Med Res Methodol 2010;10:1.

4. Craig P, Dieppe P, Macintyre S, et al. Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ 2008;337:a1655.

5. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009;151:264-9, W64.

6. Aubry S, Dubut J, Nueffer JP, Chaigneau L, Vidal C, Kastler B. Prospective 1-year follow-up pilot study of CT-guided microwave ablation in the treatment of bone and soft-tissue malignant tumours. Eur Radiol 2017; 27:1477-85.

7. Elfatairy KK, Filson CP, Sanda MG, Osunkoya AO, Geller RL, Nour SG. In-bore MRI-guided biopsy: can it optimize the need for periodic biopsies in prostate cancer patients undergoing active surveillance? A pilot test-retest reliability study. Br J Radiol 2018;91:20170603.

8. Venturini E, Losio C, Panizza P, et al. Tailored breast cancer screening program with microdose mammography, US, and MR Imaging: short-term results of a pilot study in 40-49-year-old women. Radiology 2013;268:347-55.

9. National Institute for Health Research: Supporting informatio for applicants applying to the HTA programme. https://www.nihr.ac.uk/funding-and-support/documents/current-funding-opportunities/hta/hta-supporting-information.pdf Accessed 18th April 2019.

10. Paraschin K, Guerra De Andrade A, Rodrigues Parga J. Assessment of myocardial infarction by CT angiography and cardiovascular MRI in patients with cocaine-associated chest pain: a pilot study. Br J Radiol 2012;85:e274-8.

11. Pearce N. Classification of epidemiological study designs. International Journal of Epidemiology 2012;41:393-397.

12. Julious SA, Patterson SD. Sample sizes for estimation in clinical research. Pharm Stat 2004;3:213-5.

13. Stallard N. Optimal sample sizes for phase II clinical trials and pilot studies. Stat Med 2012; 31:1031-42.

14. Browne RH. On the use of a pilot sample for sample size determination. Stat Med 1995;14:1933-40.

15. Julious SA. Sample size of 12 per group rule of thumb for a pilot study. Pharm Stat 2005;4:287-91.

16. Teare MD, Dimairo M, Shephard N, Hayman A, Whitehead A, Walters SJ. Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study. Trials 2014;15:264.

17. Billingham SA, Whitehead AL, Julious SA. An audit of sample sizes for pilot and feasibility trials being undertaken in the United Kingdom registered in the United Kingdom Clinical Research Network database. BMC Med Res Methodol 2013;13:104.

18. Altman DG. Statistics and ethics in medical research: III How large a sample? Br Med J 1980;281:1336-8.

19. Altman DG. The scandal of poor medical research. BMJ 1994;308:283-4.

20. Halpern SD, Karlawish JH, Berlin JA. The continuing unethical conduct of underpowered clinical trials. JAMA 2002;288:358-62.

21. Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2:e124.

22. Shanyinde M, Pickering RM, Weatherall M. Questions asked and answered in pilot and feasibility randomized controlled trials. BMC Med Res Methodol 2011;11:117.

23. Kannan S, Gowri S. Pilot studies: Are they appropriately reported? Perspect Clin Res 2015;6:207-10.

24. Arain M, Campbell MJ, Cooper CL, Lancaster GA. What is a pilot or feasibility study? A review of current practice and editorial policy. BMC Med Res Methodol 2010;10:67.

25. Kaur N, Figueiredo S, Bouchard V, Moriello C, Mayo N. Where have all the pilot studies gone? A follow-up on 30 years of pilot studies in Clinical Rehabilitation. Clin Rehabil 2017; 31:1238-48.

Lancaster GA, Dodd S, Williamson PR. Design and analysis of pilot studies: recommendations for good practice. J Eval Clin Pract 2004;10:307-12.

Table 1

Study design for 78 radiological studies described as being a “pilot” in their title.

Studies can be composed of more than one design type

Study design type	Number	Percentage (%)
Diagnostic test accuracy	66	84.6
Feasibility	12	15.4
Technology assessment	8	10.3
Efficacy assessment	5	6.4
Incidence study	1	1.3
Pilot	0	0