Topoi
https://doi.org/10.1007/s11245-017-9535-z

Randomized Controlled Trials for Diagnostic Imaging: Conceptual and Pratical Problems

Elisabetta Lalumera[1] · **Stefano Fanti**[2]

Abstract
We raise a problem of applicability of RCTs to validate nuclear diagnostic imaging tests. In spite of the wide application of PET and other similar techniques that use radiopharmaceuticals for diagnostic purposes, RCT-based evidence on their validity is sparse. We claim that this is due to a general conceptual problem that we call Prevalence of Treatment, which arises in connection with designing RCTs for testing any diagnostic procedure in the present context of medical research, and is particularly apparent in this case. We also identify three practical reasons why RCTs do not qualify as the best option for PET validation, which have to do with specific characteristics of nuclear diagnostic imaging, and of radiopharmaceuticals. The paper is meant to contribute both to the philosophical discussion on the EBM hierarchy of evidence, and on the specific debate on radiopharmaceuticals in nuclear medicine.

Keywords EBM · RCT · Diagnosis · Nuclear medicine · PET · Diagnostic imaging · Radiopharmaceuticals

1 Introduction

The defining core of Evidence-Based Medicine (EBM) is to base clinical care and health policy decisions on the best available evidence. Since the birth of the movement, the golden standard of evidence has been identified with randomized controlled trials (RCTs), and their meta-analyses or systematic reviews. In a RCT the correlation between a specific intervention and its outcome is tested on populations of subjects (at least two groups), in strictly controlled and ideally unbiased conditions, in order to minimize confounding factors. RCTs are recommended to assess new pharmaceuticals, combined therapies, surgical procedures, and increasingly also for validating healthcare policies, and diagnostic tests (Bluhm 2016; Howick 2011; Guyatt 1991; Sackett 2000).

The very founders of the movement, however, soon acknowledged that RCTs can not be all there is to EBM (Sackett et al. 1996). More specifically, during the years the debate on the central role of RCTs has identified two broad kinds of problems. There are *evidence problems*, regarding the nature of the evidence that a RCT can bring to bear to a specific question (the external validity of RCTs, the epistemic quality of meta-analyses, the need to integrate other kinds of evidence with RCTs, such as observational studies, the causal knowledge obtained from mechanistic reasoning, and the expert knowledge, at least in specific contexts) (Charlton and Miles 1998; Clarke et al. 2007, 2014; Landewé and Van Der Heijde 2007; Worrall 2007). But there are also *domain problems*, concerning the range of application of RCTs. This means that even granting that RCTs could provide good evidence in principle, in practice they can't be applied across the board to all domains of medical and healthcare decisions. For example, it has been noted that there can be no RCT of the effectiveness of many interventions common in medical practice, such as the Heimlich manoeuvre, general anaesthesia, tracheostomy (Howick 2011, p. 5; Glasziou et al. 2007) and, ironically, of "Parachute use to prevent death and major trauma related to gravitational challenge" (Smith and Pell 2003.) Arguably, ethical and practical problems would arise in such cases, thereby limiting the domain of application of RCTs (Sehon and Stanley 2003).

Given that RCTs are the golden standard of evidence for all interventions in the EBM paradigm, it is an important

✉ Elisabetta Lalumera
elisabetta.lalumera@gmail.com

[1] Department of Psychology, Milano-Bicocca University, Milano, Italy

[2] Department of Nuclear Medicine, S.Orsola University Hospital, Bologna, Italy

Published online: 14 December 2017

consequence of the domain question, that if a certain procedure or intervention qualifies as unfit for being adequately tested with RCTs, it is unlikely to be approved by regulating agencies, and reimbursed by healthcare providing subjects. This simple fact is becoming more and more relevant in healthcare systems largely dominated by reimbursement issues, such as US (Ter Meulen et al. 2005).

In this paper we raise a domain problem for RCTs by discussing the case of diagnostic nuclear medicine. Diagnostic nuclear medicine is a branch of medical imaging that uses radioactive materials for diagnosis. The patient receives, generally intravenously, biologically active molecules with radionuclides in the form of *radiopharmaceuticals*, also called *tracers*. Then a PET, a PET-CT, a PET-MR scan or other equipment records the distribution of the radiopharmaceutical in the body (thus including the target organ), and an image is produced, from which the specialist can obtain information on functional or metabolic characteristics, for example whether tumoral cells are present, or if they are more numerous than they were in an earlier scan. FDG PET is nowadays commonly employed to diagnose cancer, cancer staging, and response to therapies. It is it is recommended in the guidelines of many medical societies because it is more accurate and/or less invasive than other tests, and permits faster treatment decisions. The most widely used PET radiopharmaceutical is ^{18}F-FDG (2-deoxy-2-[18F] fluoro-D-glucose), but very few other have been approved, such as ^{18}F-NaF ([18F] sodium fluoride) while a number of new ones have been proposed and are currently tested (See e.g. Gambhir et al. 2001; Schober and Riemann 2012).

In spite of the wide use of nuclear diagnostic imaging, RCTs assessing radiopharmaceuticals and PET against standard competitors are relatively sparse, and sometimes even contentious (Hicks et al. 2012; Siepe et al. 2014; Ware and Hicks 2011; Weber 2011). As a result, nuclear medicine imaging techniques rarely reach the status of standard test according to Health Technology Assessment studies (see e.g. Hastings and Adams 2006). Is this situation reflecting a blind spot of the EBM methodology, or a systematic lack of rigour of a medical sub-community? We think neither of the two, and we shall argue that there are structural and practical reasons why RCTs do not qualify as the best methodology for assessing diagnostic imaging techniques in nuclear medicine.

This paper has therefore two aims. On a general level, we point at a domain problem for RCTs: they do not qualify as golden standards for assessing the efficacy of diagnostic imaging techniques in nuclear medicine. On a more specific level, our aim is to take side in a methodological debate within diagnostic imaging. With respect to the first aim, we shall conclude by agreeing on the widespread view that there has to be more to EBM than RCTs. With respect to the second aim, we shall claim that RCTs are no "proofs of certainty" for PET, and other forms of validation should be (as in fact already are) considered as providing best evidence.

We shall organize our discussion as follows. In Sect. 2 we shall present a general problem that arises when RCTs are employed to validate tests, which are to every extent determining an indirect medical intervention on the patient. The problem is that what inevitably gets assessed is the test together with the medical decision it brings about (i.e. surgery, therapy, palliative care, or any other action), thereby generating a Quinean situation of indetermination (Quine REF). We call this problem *Predominance of treatment*. It is a general problem because it has to do with the nature of RCTs, and with the aims of validation. In Sect. 3 we shall describe other, more specific difficulties that arise in connection with RCT testing of nuclear imaging techniques, in particular of PET: *Approval of Ethical Committees, Difficulties in recruiting, Lack of sponsors*, and the *Problem of interpretation*. As it already appears from the labels, some of those difficulties are of practical nature, while others have to do with the very nature of the methodology of RCTs, and of EBM in general. We shall provide examples and discuss their import on a case-by-case basis. Section 4 contains our conclusions, and a reflection on the concept of intervention.

2 Predominance of Treatment in Test Evaluation

The RCT methodology is best suited for testing new pharmaceuticals. In a typical phase III drug trial, a group of subjects receives the new product, the other group is administered the standard therapy, and the differential response or outcome is assessed. Whether it is aimed at assessing the efficacy of the new therapy, or its effectiveness in clinical scenarios, ideally a RCT measures how the independent variable directly modifies the dependent one (Lesaffre and Verbeke 2005). Setting aside different philosophical characterization of direct versus indirect causation (Woodward 2003; Pearl 2009), the point here is that typical RCTs investigate medical interventions that have a *direct* impact on the outcome.

What about diagnostic tests? According to current EBM prescriptions, assessing their specificity and sensitivity, and in general their accuracy with respect to the standard in use, is ideally not enough for their validation (Lord et al. 2006; Leeflang et al. 2008).[1] The point can be expressed as simply as that: "If a test fails to improve patient-important outcomes, there is no reason to use it, whatever its

[1] There are many accuracy studies in the literature, that compare FDG PET-CT with standard diagnostic tests in specific areas (See e.g. Czernin et al. 2013 for a review).

accuracy." (Schünemann et al. 2008, 149). Though this idea has always been part of the EBM paradigm, it happens to be more prominent now, that matters of reimbursement by national healthcare systems of funding agencies are gaining central stage in everyday clinical decisions, and the role of Health Technology Assessment studies is becoming prominent (Vach et al. 2011; Hicks et al. 2012). Patient-important outcomes are survival rate, quality of life, disease remission, and anything that falls under the category of *clinical benefit*. In other words, ideally diagnostic tests need to be validated via D-RCTs (RCTs for Diagnostic tests) that assess their clinical benefit. In a D-RCT, participants are randomized to receive a new diagnostic test versus no test or control.

A diagnostic test, however, rarely if ever produces clinical benefit directly. A test brings about a further medical action, that may or may not be beneficial: with respect to patient-relevant outcomes, it constitutes and *indirect intervention*. The clinical benefit of an imaging diagnostic test mainly depends on whether the findings of the test lead to a change in patient management with positive outcomes - for short, let us say it depends on *treatment*.[2] In the case of PET, subsequent treatment can be chemotherapy, palliative care, or surgery, which may or may not improve one of the patient-relative indicators.[3] Thus, any RCT aimed at assessing the clinical benefits of PET compared to another test would target the pair test-and-treatment. This is an unwanted Quinean situation (Quine 1953), and it brings about a problem of *Predominance of treatment*: it invariably is the treatment that eventually will make a difference in terms of clinical benefit, and eventually in the validation of the test itself.[4] We submit that Predominance of treatment explains why D-RCTs are in general neither feasible, nor successful. A recent empirical assessment of 140 D-RCTs indicates significant improvements in patient outcomes occurred in only 18% of the tests assessed (the list included common diagnostic tests such as ultrasound for prenatal care, ultrasonography for trauma, and PSA testing for prostate cancer). Furthermore, the effects of testing on patient outcomes did not correlate with the effects on further diagnostic and therapeutic interventions or with the diagnostic accuracy of tests. The authors of the review note that "the proportion of evaluated tests with significant results is much smaller than what is seen in RCTs of drugs where proportions of significant results approaching 50% are seen in efficacy trials" (Siontis et al. 2014, 612). This constitutes confirmation of our initial remark that RCTs are optimal for direct interventions.

The Problem of predominance of treatment is reflected on the nature of RCTs involving PET that are actually completed and published on scientific journals. A significant portion of them involves a so-called enrichment design, in which the results of PET are used to determine the eligibility of patients before randomization, and then alternative treatment strategies are evaluated for their outcomes in patients stratified by PET. In such studies it is possible to conclude that the combination of PET as a biomarker and a particular intervention strategy does or does not improve patient outcomes. However, it is the combination of test plus treatment that is being evaluated (Hicks et al. 2012, 1820–1821; see e.g. NIMH 2014).

The problem of Predominance of treatment becomes even more apparent when treatment options are absent, or changing faster than the time needed to complete a RCT study for the couple test-and-treatment. This is the case for the typical areas of application of medical imaging tests, such as oncology and neurology. A RCT takes years to be planned, completed and published, and can be focused on only one specific combination of PET with treatment. As soon as the treatment options change, a complete trial would have to be repeated to confirm that PET is also clinically useful in combination with a new form of therapy (Vach et al. 2011).

An example can be illuminating in this respect. One of the main directions of research on Alzheimer's disease (AD) is early diagnosis via the detection of biomarkers, and studies show that FDG PET can detect hypometabolic regions in the posterior cingulate gyri, precuneus, and parietotemporal association cortices (Ishii 2014; Schilling et al. 2016). Such studies, however, cannot be employed to demonstrate that this imaging technique gives more clinical benefits than a diagnostic competitor, because there is currently no treatment for Alzheimer, and diagnosis alone is no measurable benefit: for these reasons FDG has not been approved in the US and Canada as a tracer for dementia (Dubois et al. 2013; Rowe and Villemagne 2013; Soucy et al. 2012).

[2] Other factors, alternative to treatment, that may intervene in the causal chain from test to clinical benefit are change in time-frame of management (i.e. speeding up decisions), and alterations of patients' and clinicians' perception of the situation. Such factors are even more difficult to measure with an RCT than test-plus-intervention clinical benefit (Di Ruffano and Deeks 2016).

[3] Arguably, there are cases in which PET can have a direct clinical benefit, namely when it replaces an invasive procedure like surgical staging. Is it possible to design a RCT to assess the direct clinical benefits of nuclear medicine imaging techniques, in such cases? Yes, but it would not be necessary, and therefore not appropriate, as the positive clinical (avoidance of the invasive procedure) is immediate, and there is no need to monitor the two groups of patients (invasive versus non-invasive procedure) over time. Here, the characteristics of RCTs methodology make them not recommendable (Vach et al. 2011).

[4] We are not claiming that the Predominance of treatment problem is generalizable a priori to all diagnostic tests. That there is a Predominance of treatment problem, derives from the requirement that medical diagnostic tests are evaluated at least partially in terms of their clinical benefit, which is a contingent feature of current healthcare and scientific regulative practices Therefore, it is surely possible that no such problem (*mutatis mutandis*) arises with other diagnostic tests, e.g. in engineering or robotics. We thank an anonymous reviewer for this point.

The same pattern of problem just described for D-RCTs, and for PET testing in particular, shows up at the more specific level of radiopharmaceutical testing. Here again, the EBM prescription for validating the new product would require a 5-step study, from in-vitro modeling to a RCT measuring patient outcomes (Fryback and Thornbury 1991). Let us consider the AD example again. More recently several PET radiopharmaceuticals have been introduced to image the amyloid plaques, which are well known to be associated with Alzheimer's disease, and have been approved for imaging of the brain to estimate β amyloid neuritic plaque density in adult patients with cognitive impairment (namely Fluorine-18 Florbetaben, Fluorine-18 Flutemetamol and Fluorine-18 Florbetapir) (Syed and Deeks 2015). However, none of these radiopharmaceuticals have been demonstrated to be effective in a formal RCT trial, again due to the fact that a direct result of such imaging on patients would have hampered by lack of therapies. The major problem here is that any RCT comparing PET and treatment T with the standard procedure and treatment T would be failing, as treatment T is not effective.

We could compare PET imaging to a sight, while a typical drug can be regarded as a rifle. The sight itself has no capability to shot, just like the radiopharmaceutical has no direct therapeutic effect. To prove the quality of a sight it is considered adequate to demonstrate the optical quality of it; similarly to test the efficacy of an imaging test it should be reasonable to measure its accuracy. When EBM is requiring to perform D-RCTs, it would be as testing the sight and the rifle together, and it is clear that most of the effect is related to the shooting capabilities of the rifle. It is noteworthy that in medicine most of diagnostic tests are used as sight, such as imaging (ultrasound for example) or blood tests, or pathology diagnosis. But nuclear imaging tests are the only to use a drug (i.e. a radiopharmaceutical), and for some reason this has become prevalent in the acceptance process. Many diagnostic tests have recently gained a widespread diffusion (such as genetic tests available on internet) without the need of any formal FDA approval, simply because no pharmaceutical is involved.

3 Practical Problems (Approval of Ethical Committees, and Difficulties in Recruiting, Lack of Sponsors, and the Problem of Interpretation)

In the previous section we presented the Predominance of treatment problem as a general and structural difficulty for D-RCTs, which is particularly serious for PET evaluation because of the relative underdevelopment of treatment options, with respect to diagnostic options. In this section we describe further difficulties that may be deemed "practical", but nevertheless follow from the structure of D-RCTs.

3.1 Approval of Ethical Committees and Difficulties in Recruiting

Diagnostic tests that employ pharmaceuticals, such as PET, are substantially different from ordinary drug trials. As pointed out by Hicks et al. (2012), RCTs are most useful when the mechanism of action of treatments is not fully understood, or when there is uncertainty about the benefits versus risks. But whenever there is abundant evidence that the diagnostic accuracy of PET is superior to the standard—for example to conventional staging approaches in many cancers—the need for a RCT decreases, and it becomes even unethical, and therefore not eligible for approval by an ethics committee. This is due to two different reasons. First, ethics committees require that there is a genuine uncertainty about the relative benefits of interventions, and that there is no other way to proceed to evaluate such benefits, before authorizing that a high number of participants is left without a certain intervention. This is the Principle of Equipoise, adopted by most ethical committees (Freedman 1987; Biesheuvel et al. 2006). Applied to our case, if accuracy studies show that PET is superior to the standard test on a certain diagnostic task, an ethics committee would rather not approve a D-RCT where one group of patients is denied the PET scan.[5]

The second reason why a blind or double-blind D-RCT on PET imaging is unlikely to be approved by an ethics committee, is the possible presence of incidental findings in the scan. These are occasional findings that may be evidenced by a diagnostic test aimed at another clinical question (for example, a colon suspect malignancy detected in a subject scanned for breast cancer) (Ishimori et al. 2005). Whenever there is evidence that such findings are more common with PET than with the diagnostic competitor, it would be unethical to harm a group of subject with a potential lack of relevant health information, and therefore a RCT project would not be approved.

The same reasons that make D-RCTs on PET possibly unethical, count as difficulties in recruiting subjects. If they are informed by their clinicians that there is some (although non-conclusive, non golden-standard) evidence that the new test is superior to the standard, patients are likely to refuse to get involved in a blind study. Even recruited, some subjects

[5] As suggested by a reviewer of this Journal, a possible direction of inquiry could be on the possibility of qualifying or partially suspending the Principle of Equipoise, either when applied to tests on diagnostic procedures, and not on treatments, or when research is conducted with the declared aim of providing treatment. This suggestion deserves to be discussed on a separate paper.

may attempt to have the new test in another hospital or city, outside the scope of the trial, thereby modifying their clinical benefits outcome and invalidating the test (Graham and Weber 2016). Furthermore it is common notion that clinicians supposed to enrol patients for D-RCTs will be quite sceptical if a new test is denied to half of the patients. The use of a new imaging test is generally perceived as non-harmful, and thus recruitment of patients in purely D-RCTs has frequently been difficult.

3.2 Lack of Sponsors

A further practical difficulty arising for the project of proposing and completing D-RCTs for medical imaging tests such as PET is the relative lack of sponsors. What needs to be tested frequently is the application of new pharmaceuticals in specific application, but the industry partners of the nuclear medicine scientific community produce scanning equipments, not pharmaceuticals. Unfortunately, radiopharmaceuticals are often produced by very small companies, as the overall business of radiopharmaceuticals is relatively small and not appetizing big pharma companies. For example ^{18}F-FDG was developed by academia and is not even patent-protected; none of the top ten big pharma companies has any interest in the nuclear medicine field.

Randomized trials aiming to determine the impact of imaging on clinical benefit outcomes (such as survival or quality of life) require hundreds of patients, different research centers involved, and years of follow-up. As such, they are prohibitively expensive for the available stakeholders (academia, small companies, or national healthcare systems) (Vach et al. 2011; Hicks et al. 2012, Graham and Weber 2016).

3.3 Problem of Interpretation

The last difficulty we shall examine is peculiar to the nature of radiopharmaceuticals, when compared to non-diagnostic drugs. This difficulty arises whenever what needs to be tested is a new kind of PET diagnostic test, namely, one that employs a different active molecule (rather than a new clinical application of an already known tracer). In textbook drug testing trials, there is a phase I assessing safety and dose ranging, a phase II testing efficacy and side effects, and then a phase III aimed at determining a drug's therapeutic effect, typically by means of a RCT.

As imaging is involved, between phase II and phase III there should be a phase of validation of the criteria for reading and reporting the scan. This phase should involve some hundreds of patients, with known disease, to be studied in order to evaluate the typical findings, and to build up a shared criteria of interpretation. Such a study would definitely be very expensive, complex and time consuming, and it has been rarely if ever carried over for radiopharmaceuticals. Usually a simplified methodology is employed, namely, a consensus conference or committee deliberation, where few experts decide on criteria of interpretation, basing their judgment on their previous experience with some other drug—and therefore making an inductive leap, where acquired knowledge on a topic is projected to a new one, by analogy. While practically advantageous, this methodology involves many epistemological problems, which would deserve a careful scrutiny by the philosophical and medical community (Carne and Arnaiz 2000).[6]

4 Conclusion

RCTs are idealized as the golden standard for new drug introduction and approval. While they work reasonably well for most new pharmaceuticals, in some areas RCTs are inherently limited for being applied. We discussed the general problem of RCTs when applied to diagnostic tests, and in particular the problem of Predominance of treatment. Furthermore in the specific field of PET radiopharmaceuticals there are several practical problems that we addressed.

For these reasons we conclude that using RCTs to test the efficacy of diagnostic imaging is highly problematic. The 'proof of certainty' of such tests should be searched with different approaches, possibly related to the evaluation of diagnostic accuracy of the new imaging methods. While this is largely accepted in conventional radiology, where no drugs are used, it is much more debated in nuclear medicine, where radiopharmaceuticals are employed. As a consequence there has been a very limited number of new radiopharmaceuticals approved by authorities and reimbursed by healthcare providers in the last 15 years, especially compared to the large number of radiopharmaceuticals proposed for phase I and phase II trials. It is interesting to notice, though, that in very recent times some radiopharmaceutical have been approved by FDA without the existence of a formal D-RCT trial, such as ^{18}F-Fluciclovine and 68 Ga-Dotatate: this is likely due to the grown awareness of regulatory agencies. It would be important in the next future to recognize and formalize the special status of radiopharmaceuticals as compared to interventional drugs, in order to possibly increase the rate and speed of introduction.

[6] For a philosophical discussion of consensus conferences, see Solomon (2007).

References

Biesheuvel CJ, Grobbee DE, Moons KGM (2006) Distraction from randomization in diagnostic research. Ann Epidemiol 16:540–544

Bluhm R (2016) Evidence, meta-analysis, and systematic review. In: Solomon M, Simon J, Kincaid H (eds) The Routledge companion to philosophy of medicine. Routledge, Oxford

Carne X, Arnaiz JA (2000) Methodological and political issues in clinical pharmacology research by the year 2000. Eur J Clin Pharmacol 55:781–785

Charlton BG, Miles A (1998) The rise and fall of EBM. QJM 12:371–374

Clarke M, Hopewell S, Chalmers I (2007) Reports of clinical trials should begin and end with up-to-date systematic reviews of other relevant evidence: a status report. J R Soc Med 100:187–190

Clarke B, Gillies D, Illari P, Russo F, Williamson J (2014) Mechanisms and the evidence hierarchy. Topoi 33(2):339–360

Czernin J, Allen-Auerbach M, Nathanson D, Herrmann K (2013) PET/CT in Oncology: Current Status and Perspectives. Curr Radiol Rep 1:177–190

Di Ruffano LF, JJ Deeks (2016) Test-treatment RCTs are sheep in wolves' clothing (Letter commenting on: J Clin Epidemiol. 2014;67:612–21). J Clin Epidemiol 69:266–267

Dubois B, Feldman HH, Jacova C et al (2013) Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. Lancet Neurol 13(6):614–629

Freedman B (1987) Equipoise and the ethics of clinical research. N Engl J Med Overseas Ed 317:141–145

Friedman LM, Furberg CD, DeMets DL (1998) Fundamentals of clinical trials, 3rd edn. Springer, New York

Fryback DG, Thornbury JR (1991) The efficacy of diagnostic imaging. Med Decis Making 11(2):88–94

Gambhir SS, Czernin J, Schwimmer J, Silverman DH, Coleman RE, Phelps ME (2001) A tabulated summary of the FDG PET literature. J Nucl Med 42(Suppl):1S– 93S

Glasziou P, Chalmers I, Rawlins M, McCulloch P (2007) When are randomised trials unnecessary? Picking signal from noise. BMJ 334(7589):349–363

Graham MM, Weber WA (2016) Evaluation of the efficacy of targeted imaging agents. J Nucl Med 57(4):653–659

Guyatt G (1991) Evidence-based medicine. ACP J Club A-16:114–119

Hastings J, Adams EJ (2006) Joint project of the international network of agencies for health technology assessment—part 1: survey results on diffusion, assessment, and clinical use of positron emission tomography. Int J Technol Assess Health Care 22:143–148

Hicks RJ, Hofman MS, Ware RE (2012) Not-so-random errors: randomized controlled trials are not the only evidence of the value of PET. J Nucl Med 53(11):1820–1822

Howick J (2011) The philosophy of evidence-based medicine. Wiley, Oxford

Ishii K (2014) PET approaches for diagnosis of dementia. AJNR 35:2030–2038

Ishimori T, Patel PV, Wahl RL (2005) Detection of unexpected additional primary malignancies with PET/CT. JNM 46(5):752–757

Landewé R, Van Der Heijde D (2007) Primer: challenges in randomized and observational studies. Nat Rev Rheumatol 3(11):661–666

Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM (2008) Systematic reviews of test accuracy. Ann Intern Med 149(12):889–897

Lesaffre E, Verbeke G (2005) Clinical trials and intervention studies. In: Everitt B, Howell DC (eds) The encyclopedia of statistics in behavioral science. Wiley, New York

Lord SJ, Irwig L, Simes RJ (2006) When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Intern Med 144(11):850–855

National Institute of Mental health (2014) Imaging predictors of treatment response in depression NCT00367341, Retrieved from https://clinicaltrials.gov/show/NCT00367341

Pearl J (2009) Causality. Cambridge University Press, New York

Quine WV (1953) From a logical point of view. Harvard University Press, Cambridge Mass

Rowe CC, Villemagne VL (2013) Amyloid imaging with PET in early Alzheimer disease diagnosis. Med Clin North Am 97(3):377–398

Sackett DL (2000) Evidence-based medicine: how to practice and teach EBM, 2nd edn. Churchill Livingstone, Edinburgh

Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS (1996) Evidence based medicine: what it is and what it isn't. BMJ 312:71–72

Schilling LP, Zimmer ER, Shin M, Leuzy A, Pascoal TA, Benedet AL, Rosa-Neto P (2016) Imaging Alzheimer's disease pathophysiology with PET. Dement Neuropsychol 10(2):79–90

Schober O, Riemann B (eds) (2012) Molecular imaging in oncology, vol 187. Springer Science & Business Media, Dordrecht

Schünemann AHJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R et al (2008) Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 336:1106–1110

Sehon SR, Stanley DE (2003) A philosophical analysis of the evidence-medicine debate. BMC Health Serv Res 3:14–24

Siepe B, Hoilund-Carlsen PF, Gerke O, Weber WA, Motschall E, Vach W (2014) The move from accuracy studies to randomized trials in PET: current status and future directions. J Nucl Med 55(8):1228–1234

Siontis KC, Siontis GC, Contopoulos- Ioannidis DG, Ioannidis JP (2014) Diagnostic tests often fail to lead to changes in patient outcomes. J Clin Epidemiol 67(6):612–621

Smith GC, Pell JP (2003) Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. BMJ 327:1459–1461

Solomon M (2007) The social epistemology of NIH consensus conferences. In: Kincaid H, McKitrick J (eds) Establishing medical reality. Philosophy and Medicine, vol 90. Springer, Dordrecht

Soucy JP, Bartha R, Bocti C et al (2012) Clinical applications of neuroimaging in patients with Alzheimer's disease: a review from the Fourth Canadian Consensus Conference on the Diagnosis and Treatment of Dementia. Alzheimers Res Ther 5(Suppl 1):S3

Syed YY, Deeks E (2015) [18F] Florbetaben: a review in β-Amyloid PET imaging in cognitive impairment. CNS Drugs 29(7):605–613

Ter Meulen R, Biller-Andorno N, Lenk C, Lie R (2005). Evidence-based practice in medicine and health care: a discussion of the ethical issues. Springer Science & Business Media, Dordrecht

Vach W, Hoilund-Carlsen PF, Gerke O, Weber WA (2011) Generating evidence for clinical benefit of PET/CT in diagnosing cancer patients. J Nucl Med 52(Supplement 2):77S–85S

Ware RE, Hicks RJ (2011) Doing more harm than good? Do systematic reviews of PET by health technology assessment agencies provide an appraisal of the evidence that is closer to the truth than the primary data supporting its use? J Nucl Med 52(Supplement 2):64S–73S

Weber WA (2011) Is there evidence for evidence-based medical imaging? J Nucl Med 52(Supplement 2):74S–76S

Woodward (2003) Making things happen: a theory of causal explanation. Oxford University Press, Oxford

Worrall J (2007) Why there's no cause to randomize. Br J Philos Sci 58:451–488