Study Guide 3: How to Critically Appraise a Paper

Dr David Chinn,
Research, Innovation and Knowledge Department,
Queen Margaret Hospital, Dunfermline, Fife.
david.chinn@nhs.scot 01383 623623 (ext 20943)
Alternative contact: Prof Frances Quirk frances.quirk@nhs.scot 01383 623623 (ext 20941)

Contents Page

(1) Overview and learning outcomes

This guide is designed to help health care staff identify the strengths and weaknesses of research papers. This is particularly important for staff planning services and deciding when to change practice as these decisions should be based on robust evidence. The guide will also help those designing their own projects and writing up their results in ad hoc reports and peer-reviewed papers. After reading this guide you should be able to:

- Define what is meant by critical appraisal
- Describe the key features of quantitative & qualitative research
- Understand the strengths & weaknesses of qualitative research designs
- Understand the strengths & weaknesses of quantitative research designs
- Identify the correct study design for a given research question
- Be familiar with the hierarchy of evidence
- Be familiar with critical appraisal check lists
- Understand sources of bias
- Gain practical experience in critical appraisal of a research paper

Associated NHS Fife study guides:

1	How to devise a research question and choose a study design
9	Introduction to qualitative research
10	Introduction to medical statistics
11	How to calculate sample size and statistical power
13	How to make sense of numbers

(2) Introduction

Even if you never do any research of your own it is critically important that you can judge the quality of papers you read. Every published study will have its own strengths and weaknesses. Critical appraisal is the process by which these are identified to establish whether the results are valid and their interpretation is reliable. Journal editors will subject submitted papers to a peer-review process by asking one or more referees to comment on the article. In general, the process works well but the process can fail and poor quality research can be published even in high quality journals. Publication of a flawed paper can have serious consequences as patients may receive the wrong treatment from a change in practice, the original research team may continue to use unsound methodologies, and other research teams may adopt flawed practices in the belief that the methodology is robust because the paper has been published after peer-review. Hence, it is important for all practitioners to develop their own robust skills in critical appraisal.

There are many important features to consider when appraising a paper. These include the methods adopted, the potential biases in the study design, the data obtained and any statistical treatment applied, the interpretation of the results in relation to what is already known on the topic, and the conclusions. Generally, research papers can be classified into one of three types depending on their quality and reliability:

(a) Relevant, well thought out, valid methods properly carried out with clearly presented results which address the research question(s).
(b) Patently obvious errors at different levels (methodology, analysis, interpretation). Unfortunately, such papers do get published but can be dismissed easily.
(c) Relevant and interesting (intellectually stimulating) but poorly carried out or otherwise fundamentally flawed. Such papers may appear good but are invalid once subjected to a proper critical appraisal. These can be dangerous because decisions about treatments and allocation of resources may be made on the basis of the results. There may also be ethical issues over treatments for future patients.

Hence, publication of a paper does not guarantee the study and its conclusions are sound; you should make your own assessment. Research claims are only justified if the methods are reliable and valid.

TIP: always check the journal's correspondence section for the months following the publication to see if others have raised concerns about the paper. As an example, check:

Bagenal FS, Easton DF, Harris E, Chilvers CED, McElwain TJ. Survival of patients with breast cancer attending the Bristol Cancer Help Centre. Lancet, 1990; 336: 606-610.
This was an observational study comparing survival of women with breast cancer attending the Bristol Cancer Help Centre with that of women attending the Royal Marsden and other hospitals. The study design and analysis was flawed and generated a lot of criticisms subsequently published in the journal's correspondence pages:
Sheard TAB. *Lancet* 1990; 336: 683.
Heyes-Moore L. *Lancet* 1990; 336: 743.
Wright S. *Lancet* 1990; 336: 743.
Munro J, Payne M. *Lancet* 1990; 336: 743-744.
James N, Reed A. *Lancet* 1990; 336: 744.
Boulter PS. *Lancet* 1990; 336: 744.
Bennet G. *Lancet* 1990; 336: 744.
Tonkin R, Tee D. *Lancet* 1990; 336: 744
Lewith G. *Lancet* 1990; 336: 744.
Hayes RJ, Smith PG, Carpenter L. Bristol cancer help centre. *Lancet* 1990; 336: 1185.
Sheard TAB. *Lancet* 1990; 336: 1185-1186.
Bodmer W. *Lancet* 1990; 336: 1188.
Tobias JS, Baum M. *Lancet* 1990; 336: 1323.
Bourke I, Goodare H. *Lancet* 1991; 338: 1401.
Goodare KJ. *Lancet* 1992; 340: 248.

See the reflective paper published 3 years later:
Weir MW. Bristol Cancer Help Centre: success and setbacks but the journey continues. *Complementary Therapies in Medicine* 1993; 1: 42-45.

(3) Research Methodologies

The research question determines the study design and its methodology. Qualitative and quantitative methods are considered by some authorities as being complementary. Others consider the two approaches are in logical conflict through their underlying assumptions. Qualitative research is concerned with developing explanations of social phenomena, including people's lived experiences, their views and attitudes. The data are non-numerical and typically relate to words. Quantitative research includes estimation of a numerical value such as a proportion (prevalence), the strength of association (correlation) between variables, or testing of a hypothesis. A study which seeks to answer the question 'How many women are recalled following a cervical smear' is clearly quantitative in nature, and that which asks the question 'What are the unrecognised concerns of women recalled following a cervical smear' is then qualitative in nature. A research study can include both qualitative and quantitative methods and studies using mixed methods are becoming more popular having been recommended by authorities such as the Medical Research Council.

(4) Qualitative studies

Qualitative studies have an important role particularly when first approaching a topic about which little is known. They use standard, observational methods to explore people's beliefs, experiences and knowledge. Techniques include document analysis (nursing notes, emails, minutes of meetings, diaries etc), participant observation (of a parent's interaction with their child, for example), one to one interviews (semi-structured, unstructured) and focus groups where small groups of individuals are asked open questions by a facilitator who records and interprets the conversations. Some qualitative, observational studies may include collection of quantitative data. For example, a study of parent's engagement with their child during a particular play activity may compare the father with the mother in the number of instances they each touch the child.

Each method of data collection has its own strengths and weaknesses (Table 1).

Table 1. Strengths and weaknesses of common data collection methods in qualitative research.

Document analysis	
Strengths	• Low cost • Convenient (assuming documents are accessible) • Potential for unbiased data collection • Good for prospective studies (e.g. diaries of symptoms, medication adherence) • Potentially comprehensive records • May allow retrospective review of change over time in populations if source material has been collected rigorously and to high standards of completion (e.g. care home nursing notes) • Potential source of contemporary, independent evidence • May be the only source of evidence for long-term historical research
Weaknesses	• Missing documents or content a threat • Possible restricted accessibility (confidentiality) • If multiple observers beware writing styles and content may vary • Potential ineligibility of written content • Accuracy and authenticity of content not guaranteed • Selective reporting (e.g. of unfavourable events) • Potential change in standards/practice over time (historical studies) • Context in which content is recorded may not be appropriate for the research • Information recorded may not be germane to research question • Volume of data may be excessive (hence, collection and analysis time consuming)
Direct observation	•
Strengths	• Can provide objective evidence on behaviours and interactions, verbal and non-verbal, in a natural setting if participants unaware they are being observed • Observations made within context and environment under study • Use of film and / or audio can provide remote and independent data unbiased by the presence of an observer • Researcher can be a participant (provides further in-depth analysis of context)

Weaknesses	• Hawthorne effect – participants may alter their behaviour, knowingly or unknowingly, if aware they are being observed. • Can be time consuming • May be subject to practical constraints • Rigorous training of multiple observers necessary • Potential conflict of interest if observer notes unethical or unprofessional behaviour
1:1 interviews	•
Strengths	• Can be semi-structured or unstructured • Gives opportunity to probe in-depth using ‘open questions’ • Interviewer can clarify any uncertainty over question wording • Question sequence can be varied to suit interviewee • Questions can be left out if considered irrelevant • Can use less precise wording suited to the interviewee • Potential use of audio or video recording to collate data
Weaknesses	• Potentially expensive • Can be lengthy and collection of data, and its analysis time consuming • Not anonymous, though interviewer can give reassurance • Results subject to response bias but also to observer bias (training an important issue if using multiple interviewers) • Consent to record interview may be withheld and then need to record field notes can be distractive • Does not provide evidence of interaction between participants
Focus groups	•
Strengths	• Can offer more efficient data collection than 1:1 interviews • Groups can be made up of participants who know one other (e.g. work colleagues) who share an experience or participants who are total strangers (to elicit ‘social, group norms’) • Provides evidence on the interaction between participants • Improved access to ‘hard to engage’ groups • Allows interaction between respondents to explore similarities and differences in views • Replicates the cultural context in which people discuss issues, particularly sensitive ones • Venues can be chosen to offer a ‘safe’ environment • Can study how opinions are formed from the flow of conversations within the group
Weaknesses	• All must consent to audio record interview as any dissenting participant may risk accuracy of data collection • Analysis time consuming • Possible lack of disclosure of sensitive attitudes in a group setting • Not anonymous, maintaining confidentiality between participants can be uncertain • Potential discord between participants from disclosure of ‘unsavoury’ attitudes in a group setting • Potential suppression of views from ‘power’ relationships in groups where participants are known to one another • Risk of loss of control of the group and direction of conversations by the facilitator from (i) extraneous distractions at the venue and (ii) dominance of a single participant

The non-numerical data are described as ‘rich’. Any theory may emerge from the data collected unlike quantitative research where a hypothesis may be established first and subjected to challenge by experiment.

The analysis of qualitative data can be onerous, time consuming and subject to bias from the person undertaking the analysis. In general, the analysis is best undertaken by the person collecting the data as review of any transcripts of audio recordings can be misunderstood by those not privy to the way a patients' views may have been expressed. As an example, the simple statement in a transcript "she was alright" may be interpreted differently according to any emphasis made on individual words, or pauses made during its expression. Hence, "<u>she</u> was alright" is different from "she <u>was</u> alright" which, in turn is different from "she was [pause] alright [with the latter word expressed as a question]".

Qualitative studies can be made before, during or after a quantitative study. For example, when a new intervention or service is introduced, qualitative studies may be used (1) beforehand, to interview staff to identify their concerns or potential barriers about the new service, (2) once in operation to interview service users about their experiences of using the new service, and (3) once the service has been embedded for some time to interview staff on residual problems arising from any new working arrangements.

As stated before, any theory emerges from, or is refined based on the data. Qualitative studies use a number of different analytical and theoretical approaches. These include discourse analysis, grounded theory, ethnographic and phenomenological approaches. A detailed discussion of the many varied different approaches is beyond the scope of this guide and the reader is referred to one of the many textbooks on the subjects (see Further Reading). However, critically appraising a qualitative study does require the reader to assess if the correct theoretical approach has been used to answer the research question(s), to be aware of the strengths and weaknesses of the various data collection methods (Table 1), and to assess if the correct interpretation has been made from the data collected. Further detail on critically appraising qualitative research is given in section 7.

(5) Quantitative studies

Quantitative studies can be of two types, broadly observational and experimental. Observational studies are descriptive; the subjects do not receive any treatments or experimental interventions. The measures of interest are recorded with no attempt to influence the measurement. Experimental studies, by comparison, involve some intervention to change a variable and monitor the effect of this change on some function, for example the effect of a drug on blood pressure. The drug may be compared with another drug or with a placebo, though the latter may not be possible, or considered ethical if seen to be withholding a proven treatment without patient consent. The need for a placebo may be a particular difficulty for community interventions in, for example, studies promoting a change in diet or exercise behaviour.

Experimental studies can be conducted using separate groups for treatment, control and placebo conditions (independent groups design) or by using the same group to receive all conditions (within-groups, repeated measures design – crossover design). Cross-over studies have added benefits with regard to statistical power and require fewer participants compared with independent groups.

Quantitative studies have a role to play in investigating relationships, causal pathways and testing hypotheses. They use structured data collection methods including questionnaires (structured, semi-structured) and instruments to record physiological attributes with data being collected in a measurement scale and based on a robust protocol. Data may be subjective, for example from self-report, or objective, for example from a physiological measurement. The data may be subjected to simple descriptive statistical analysis or to complex analyses to compare groups using parametric and non-parametric techniques depending on the distribution of the data (bell-shaped or skewed).

The research question determines the study design each of which has its own strengths and weaknesses. The designs include:

(i) cross-sectional,
(ii) case control (also known as a case referent design),
(iii) cohort or longitudinal,
(iv) cross-over,
(v) randomised controlled trial (considered the ‘gold standard’ design for investigating hypotheses).

(5.1) Cross-sectional

In a cross-sectional study each participant is examined at one point in time. Such studies are relevant for estimating the prevalence of a disease, symptom, or risk factor, or for investigating associations between a disease and putative causal factors. Cross-sectional studies can identify associations but cannot be used to investigate causal pathways as it is unknown which came first, the disease or the exposure to the putative causal factor. An early example from the 1940s was the observation that many patients with lung cancer happened to be tobacco smokers. The fact that two features are associated does not imply they are causally related. Consider the observation that many blind people happen to own a dog, and not just any dog but a Labrador! Could there be a link between dog ownership leading to blindness? In this example, the reality is the other way around, an effect called ‘reverse causality’.

The results of a cross-sectional study can be used to set up a hypothesis. In the case of smoking and lung cancer the hypothesis that tobacco smoking leads to the development of lung cancer can be tested in a cohort (or longitudinal, follow-up) study of smokers and non-smokers to compare disease frequency in the two groups (a pseudo experiment with naturally occurring groups). Such a causal association was firmly established in the classic work of Richard Doll and Austin Bradford-Hill in the 1950s.

One problem with cross-sectional studies concerns the participants studied because they are seen at only one point in time. A study was designed by the Health and Safety Executive to investigate respiratory ill health associated with occupational exposure at cotton mills dealing with waste cotton where the exposure to cotton dust in the atmosphere was well above the accepted ‘safe’ levels (called the threshold limit value, TLV). The 60 workers were examined once but no evidence of occupational disease due to their occupational exposure (byssinosis) was noted. However, it was clear that the turnover of staff at the mill was high and anyone who had difficulty working in those conditions simply left. The workers who remained were (apparently) unaffected by the adverse conditions and, effectively, were ‘survivors’. Accordingly, any association between respiratory ill health and exposure to cotton dust was missed (*see* Further reading, Chinn et al, 1976).

The strengths and weaknesses of the cross-sectional design are given in Table 2.

Table 2 Strengths and weaknesses of the cross-sectional study design

Strengths:	• Convenient, as carried out at one point in time • Often low cost • Can be easily set up and results obtained quickly • Useful for generating hypotheses by determining associations • Can be repeated in different settings
Weaknesses:	• Cannot study cause and effect relationships (temporality issues) • Prone to bias if studying only 'survivors' • Not good for rare conditions as numbers needed to study will be excessive • Cannot predict future health outcomes as any associations identified may be spurious rather than causally related

(5.2) Case-control

Case control studies are appropriate when studying factors associated with the development of rare conditions. Individuals with the condition of interest (cases) are identified and 'matched' with one or more individuals without the condition (controls). Features such as lifestyle factors and exposures can then be compared between cases and controls to identify suspect causative agents. An example from the 1990s was the study of the relationship between diet (specifically beef consumption) and Bovine Spongiform Encephalitis ('Mad Cow disease').

Case control studies can be subject to selection bias. For example, a population-based study of patients with upper aero-digestive tract (UAT) cancers relied on recruiting patients from a regional radiotherapy centre. Patients with a UAT cancer attending for radiotherapy were identified and their personal lifestyle and occupational exposures were compared with control patients. However, some patients with UAT elected not to undergo radiotherapy, so were unidentified from amongst the population and hence data capture was incomplete.

Another form of selection bias is called 'Berkson's fallacy' that can occur with hospital-based studies when cases and controls differ systematically in their risk of admission to hospital due to a combination of exposure and disease. The combination may increase or decrease the exposure rate amongst the cases that will distort the statistical results relating the exposure to disease occurrence. An example is the admission criteria applied before patients become eligible for surgery; some surgeons will only consider patients ready for coronary artery bypass operations after they have discontinued smoking.

The strengths and weaknesses of the case control design are given in Table 3.

Table 3 Strengths and weaknesses of a case control study.

Strengths:	• Good for studying rare conditions • Cases should be easily identifiable (and presumably available) • Relatively cheap • Can be done from hospital setting • Can be easily set up and results obtained quickly • May be statistical consideration in that fewer subjects required compared with cross-sectional and cohort studies • Can look at several potentially causative factors in the same study
Weaknesses:	• Highly dependent on suitable controls • A need for careful matching for known confounders, e.g. age, gender, etc • The greater the number of matching criteria the greater is the difficulty of finding suitable controls • Results can only support the suggestion of, but not prove a causal association (problem of temporality – which came first, the disease or the exposure?) • Subject to reporting bias, e.g. from patient's memory or notes. Cases can have selective memory e.g. mothers of children with autism may have greater recall of past events, which might be considered causative, compared to mothers of controls • Cases recruited from hospital may not be 'representative' of all cases with the disease (selective survival)

(5.3) Cohort

A cohort study seeks to follow-up a group of individuals over time to measure some aspect of change. Several groups may be involved with different exposure to a putative risk factor. Cohort studies may be prospective or retrospective. In a prospective study a group (cohort) of individuals are followed over time to investigate the development of a disease or relapse of symptoms, for example. Cohorts may include occupationally exposed individuals, infants and children (as in growth studies), patients discharged from hospital etc. In a retrospective study a cohort is defined from the past and the individuals followed-up to the present day (also called an historical cohort study). Such studies include those looking at the association between birth weight, early life exposures and subsequent health outcomes in adulthood (heart disease, stroke, diabetes).

Cohort studies can identify causal associations as, unlike cross-sectional studies they can address temporal relationships by recording which came first, the exposure or the disease. They can quantify the attributable risk of developing a disease (for example, the development of lung cancer in smokers) and hence the impact on population health status of eliminating the causative factor.

Cohort studies can extend over many years and can suffer bias in data collection due to selective loss to follow-up if individuals move away, die, or drop out for reasons associated with the condition being investigated. Hence, patients who develop symptoms may decline to participate in a follow-up examination thereby distorting the measures of relative risk between groups. However, an assessment can be made in the data analysis to estimate the effects of bias from unbalanced loss to follow-up.

The strengths and weaknesses of the cohort design are given in Table 4.

Table 4 Strengths and weaknesses of a cohort (longitudinal) study design

Strengths	• Addresses issue of temporality (which came first, the exposure or the disease) • Good for studies of causation (identification of putative risk factors) • Can quantify the risk of developing a condition • Can quantify attributable risk and, therefore, the likely impact on health status from eliminating the causative factor. • Less prone to observer bias in data collection at the start of the study (investigators will not know which participants are likely to develop the condition under investigation) • Can assess multiple outcomes in the same study
Weaknesses	• Requires long-term commitment to maintain standards (quality control) • Can be expensive though not necessarily • Results may not be available for years, during which time exposure conditions may have changed (e.g. in industry) • Serious threat of bias from incomplete follow-up due to selective loss from the cohort • No control over changes (e.g. in the environment) which may affect the relationship between the disease and putative risk factor being investigated e.g. change in tobacco taxation or legislative changes (seat belts) • Not relevant for rare diseases because follow-up must be prolonged to capture enough cases to make comparisons meaningful (threat to statistical power)

(5.4) Cross-over study

In a cross-over study each participant is subjected to both interventions being compared. A participant receives one intervention then, after a suitable washout period is switched to the second intervention. The order in which participants receive the interventions is randomised. One advantage of this study design is that, effectively, each participant acts as their own control. In consequence, the number of participants required to achieve a given statistical power is less than that required for other randomised designs involving parallel groups of different participants. This is because variability *within-patients* is less than that *between-patients.*

There are limitations, however, and cross-over trials are only useful when the effect sought is short-term and the washout period is short. The strengths and weaknesses of cross-over studies are given in Table 5.

Table 5 Strengths and weaknesses of a cross-over study design

Strengths	• Useful for studies of short-acting drugs in chronic (stable) diseases • Allows for a randomised design, hence reducing potential bias • Convenient design where each participant acts as their own control • Requires fewer participants than a traditional randomised controlled trial involving parallel groups.
Weaknesses	• Requires a washout period between treatments • May be residual effects from first treatment that interact with second treatment • Possible ethical and clinical concerns regarding withdrawal of treatment during the washout period • Less suitable for long-term drug effect studies • Less suitable for acute diseases if the condition varies naturally between treatments • Cannot be used for diseases which can be cured • Potential for bias in analysis failing to identify treatment order effects

(5.5) Randomised Controlled Trial

The randomised controlled trial (RCT) is considered best scientific evidence (when it works). Participants are randomised to receive one of two or more treatments. Randomisation works in the long-term to smooth out differences between groups but cannot guarantee balanced groups when the number of participants is low. An RCT is not always possible because of ethical issues if assigning patients to what may be considered an inferior treatment (for example, use of a placebo drug in patients with asthma), or when there is potential to do harm (for example, when studying the effect of alcohol intake on pregnancy outcome).

RCTs work best when the number of patients recruited and followed-up to completion satisfies the initial power calculation to test the hypothesis. Measurements should be objective, valid, reproducible, and made contemporaneously in both groups, as well as double-blinded (i.e. neither the patient nor the researcher assessing the treatment effect is aware of which treatment the patient is on). Groups should be balanced at the start of treatment (by comparing baseline data) and the data should be analysed using an *'intention to treat'* analysis whereby participants remain in the groups to which they were allocated. Analyses where data are analysed according to the treatment participants actually received is called a *'per-protocol'* analysis and allows for the situation where participants may have been switched between groups.

RCTs are prone to errors in design from inappropriate randomisation strategies. Each participant should have an equal chance of being allocated to either group. Methods abound regarding random assignment. This can include randomisation in blocks to guarantee equal numbers in groups after, say, 20 recruits. 'Alternate assignment' whereby a patient seen is allocated to one group and the next patient seen is allocated to the second group is not the same as randomisation. Critical appraisal of an RCT requires an assessment of the quality and completeness of the randomisation process. If in doubt over the technique you should consult the statistics books.

The strengths and weaknesses of RCTs are given in Table 6.

Table 6 Strengths and weaknesses of a randomised controlled trial

Strengths	• Considered best scientific evidence of effectiveness • Provides better control over known (and unknown) confounders • Limits bias through double-blinding, where possible • Allows evaluation of a single intervention / drug on an outcome
Weaknesses	• Prone to problems of inappropriate randomisation • Double-blinding, or single-blinding not always possible • For drug trials the assumption that participants do take the medication according to the instructions • Can be expensive • Requires considerable resource through project management • Results may be over-estimated due to rigorous inclusion and exclusion criteria applied. Hence, results may not be replicated in the general population of patients.

The research question determines the study design (Table 7).

Table 7. Research questions and quantitative study designs

Research question	*Study design*
What is the prevalence of asthma in school aged children in Fife?	Cross-sectional
What is the association between barriers to physical activity and socio-economic position in adults aged 40-65?	Cross-sectional
Is there an association between shipyard welding and respiratory symptoms?	Cross-sectional
To what extent is the development of respiratory symptoms related causally to shipyard welding?	Prospective cohort
What is the incidence of laryngeal cancer in former steel workers?	Retrospective cohort
Is laryngeal cancer associated with past exposure to acid mists in steel mills?	Case control
Is maternal obesity a risk factor for stillbirth?	Case control
What is the frequency of occurrence of anaemia in relation to the diagnosis of colorectal cancer and site of tumour?	Retrospective cohort
What is the impact of a primary care-based dermatology nurse intervention on the quality of life of children with atopic eczema?	Randomised controlled trial
Is drug X better than placebo in treating fatigue in patients with multiple sclerosis?	Cross-over (or parallel group RCT)

(6) Hierarchy of the strength of evidence

When reviewing a paper you need to be aware of the hierarchy of evidence. Several classifications exist to rate the levels of evidence. One such scheme is:

I-1 Systematic review and meta-analysis of 2 or more double-blind randomised controlled trials
I-2 One or more large, double-blind RCTs
II-1 One or more well-conducted cohort studies
II-2 One or more well-conducted case-control studies
II- 3 A dramatic, uncontrolled experiment
III Expert committee sitting in review; peer leader opinion
IV Personal experience

The distinction is between evidence-based practice (categories I and II) and practice-based evidence (categories III and IV). The latter is perfectly acceptable in the absence of level evidence I and II.

(7) Check lists for critical appraisal of different study designs

Critical appraisal check lists are available according to different study designs from the Critical Appraisal Skills Programme (CASP). These lists have been adapted for use from advice from the Journal of the American Medical Association. Each checklist is a series of prompts covering important aspects of that particular study design. The appraisal tools are not copyright-free but can be downloaded free for personal use at:

https://casp-uk.net/casp-tools-checklists

- Qualitative
- Case Control
- Cohort
- Diagnostic test
- Economic Evaluation
- Systematic Review
- Randomised controlled trial
- Clinical Prediction Rule Checklist

(8) Bias

Bias is the unequal distribution of error. It is the greatest threat to the validity of any research study and a key aspect to look for when reviewing published papers. There are many sources of bias (Table 8).

Table 8. Principal sources of bias in research studies.

Source	Comment
Design	Any aspect of study design, e.g. faulty sampling, incorrect randomisation, temporal differences in examination of subgroups, inappropriate calibration of instruments, poor statistical analysis with failure to account for confounding, use of wrong statistical tests.
Assumption	Faulty logic of investigator, which can lead to faulty conceptualisation of the research problem, faulty interpretations, and conclusions.
Selection	Faulty selection when the characteristics of the sample differ from those of the wider, target population. All potential subjects should have an equal chance of being chosen e.g. a written invitation not read by illiterate people or those who cannot read English.
Ascertainment	Variation in diagnostic criteria used between or within studies (e.g. criteria to define hypertension). Criteria may change with time.
Response	A major source of bias leading to a systematic error from differences in characteristics between those who accept and those who decline an invitation to take part in the research. It is not always possible to compare the characteristics of responders and non-responders but it should be done where there is a source of independent data.
Measurement	Systematic error from poor calibration regimes, measurement errors, change of instruments between repeated assessments, different instruments used to collect data from different subgroups, data handling procedures, digit preference.
Measurement decay	Error from a change in the measurement process over time due to a change in instrument performance or from change in technique by an observer.

Classification	Categorisation of the results. For example, definition of an ex-smoker (abstinent for one day, one week, one month, six months, one year, ten years?)
Recall	Recall by respondents may be selective or otherwise different between groups with different rates of cognitive decline.
Reporting	Respondents may be apprehensive about being interviewed and give the responses they think the interviewer wants. Respondents may under-report or over-report symptoms depending on any vested interest e.g. occupational surveys of back injury, with denial to avoid being made redundant, or over reporting to get compensation. Bias can arise with postal questionnaires when it may be uncertain who has filled in the questionnaire and if they have had help.
Social desirability	People may wish to present themselves at their best and will respond to questions accordingly.
Acquiescence response set ('yes-saying')	Respondents will more frequently endorse a statement than disagree with its opposite
Observer	Differences in measurement techniques between observers, and within observers over time (measurement decay). Different interviewers may show systematic differences in asking questions and recording responses. Interviewers may ask questions in a manner which encourages respondents to answer in a desired way. Initial training and inter-observer assessments are very important to eliminate differences in techniques. May need to be repeated throughout the study.
Follow-up	Loss of follow-up. Bias due to systematic differences in characteristics between those who return compared to those who decline to attend, or are otherwise lost to follow-up measurements in a cohort study.
Lead time	Failure to follow-up two or more comparison groups at the same time.
Analysis	Inappropriate use of statistical methods, for example, different treatment of outliers, missing data, incorrect tests of significance and neglect of confounders.
Interpretation	Errors in inferences drawn from the statistical analyses, for example over aspects of association versus causation.
Publication	Reports of negative findings are less likely to be selected by editors for publication. Authors may have over-emphasised any positive findings to encourage acceptance of their paper. Publication bias may lead to a researcher believing that his/her contribution is unique and original.

An early example of a biased RCT was a study to compare 2-year survival for radiotherapy versus surgery in patients with operable lung cancer. Patients with lung cancer were randomised to either radiotherapy or surgery. The study findings suggested that mortality experience was significantly better for surgery than for radiotherapy. However, the exclusion criteria included those with *inoperable cancer*. Patients with inoperable cancer were identified if they were in the surgery arm but not if they were in the radiotherapy arm. Hence, there was an inherent bias in favour of surgery.

Exercise: Consider a study with the aim to determine the risk profile for diabetes in the general population. The intention is to employ 4 researchers, working Monday to Friday (9am to 5pm), who will telephone 1000 respondents and ask questions on their age, gender, ethnicity, height, weight, waist and hip size, and family history of diabetes. Telephone numbers will be extracted at random from a local British Telecom telephone directory. What are the potential sources of bias in this study? Use the list in Table 8 above. Afterwards check your thoughts with those expressed in Appendix A (page 22).

(9) General considerations appraising a qualitative paper

The subjective nature of qualitative research represents a challenge to deriving a formal, rigorous checklist. However, some general areas are worth considering. The list below has been compiled from several sources but you may wish to consult the lists from others and develop you own version.

Overview (screening questions):

1. Does the paper describe an important clinical problem?
2. Is there a clearly formulated, focussed research question?
3. Is a qualitative approach appropriate?

Subsidiary considerations

4. Who are the authors?
5. What are their professional backgrounds?
6. Where was the work done?
7. Who funded it?

Methods

8. How were the settings/participants selected? Could there be bias from non-participation in the selection of target participants?
9. Is the method of data collection (interviews, focus groups, participant observation etc) the most appropriate? Has the methodology chosen been justified?
10. Has the number of participants been justified (for example, through data saturation)
11. Are there any ethical concerns over, for example, the need for informed consent, or the disclosure of sensitive information that could result in a conflict of interest for the researcher?

Analysis

12. Who has analysed the transcripts / field notes?
13. Is there a comprehensive description of the analysis process and was it systematic?
14. Has any coding frame been fully described?
15. Is the theoretical framework chosen appropriate (grounded theory, ethnography, phenomenological)?
16. Is there evidence of attempts at triangulation, where the findings are confirmed by, for example, respondent validation?
17. Has the researcher's background and any source of observer bias been acknowledged?
18. Has the researcher's reflexivity been addressed?

Results

19. Are the results credible?
20. Are the results (themes) supported by appropriate quotations from participants?
21. Is there an alternative explanation for the results?
22. Has any contradictory data been presented? Are both positive and negative examples cited?

Conclusions

23 Have any strengths and limitations been acknowledged?
24 Has any potential bias been overlooked?
25 Are the conclusions justified by the results?
26 Are the results clinically important?
27 Are the findings of the study transferable to other clinical settings?

There is some debate as to whether the findings from qualitative research are 'generalisable'. A study researching the lived experiences of a small sample of patients with heart failure may have different findings when carried out in a different setting where patients had access to different services (e.g. heart failure nurses working in the community). Other factors, such as the severity of the heart failure, the age and responsibilities of the patient and presence of any comorbidity are likely to influence the responses to the researcher's questioning. The issue of a study's 'generalisability' may be determined by how the participants were recruited, whether as a convenience sample (likely to be less generalisable) or purposive, where a wide range of participants are deliberately targeted to ensure a comprehensive assessment of the views of that particular patient cohort.

(10) General considerations appraising a quantitative paper

The nature of the appraisal of a quantitative study will depend on the study design which in turn will depend on the research question(s). The appropriate CASP checklist referred to in section 6 above may be downloaded and used. However, the following list includes some general considerations appropriate for all study designs. The list was compiled from several sources. The challenge is to review the lists from others and develop you own version.

First thoughts

1 Who are the authors?
2 What are their professional backgrounds?
3 Where was the work done?
4 Who funded it?
5 Could there be a conflict of interest?
6 Does the paper describe an important clinical problem?
7 Is there a clearly formulated, focussed research question?
8 Are the aims / objectives clearly stated?

Methods

9 Is the study design appropriate for the research question?
10 Is there a clear description of the participants / patients / target population?
11 Is the sample recruited representative of the target population?
12 Are the inclusion / exclusion criteria clearly stated?
13 Is the control group, if any, comparable?
14 Is there a clear description of the treatments?
15 If an intervention is being compared with 'usual care' is a description given of it?
16 Are measurements made concurrently in both intervention and control groups?
17 If patients have been randomised is the technique used appropriate?

18 Are the outcome measures reliable / reproducible / valid / clinically important?
19 Are any details given of calibration and quality control procedures?
20 Have the observations been made blinded?

Data analysis

21 Is there a statement justifying the size of the sample (power calculation)?
22 Is the handling of data described (treatment of outliers, missing data etc)?
23 Is the choice of significance tests (parametric or non-parametric) justified?

Results

24 Tables and graphs: are they clear and self-explanatory?
25 Do numbers referred to in the text conform to data in the tables and graphs?
26 Do the results include measures of central tendency (mean or medians)?
27 Do the results include measures of spread (standard deviations, standard errors of the mean or inter-quartile ranges)?
28 Do the results include confidence intervals?
29 Have any potential confounders been overlooked?
30 Are the results believable?
31 What is the potential for bias?
32 Have any sources of bias been overlooked?

Discussion

33 Have the authors addressed any weaknesses in their study design?
34 Are the conclusions supported by the data?
35 Are the findings transferable to other clinical areas?
36 Are conclusions relevant to the original questions?
37 Could the findings have occurred because of uncontrolled bias, confounding, or chance alone?

(11) Sample paper

Read the sample paper below and critically appraise it using the list above. Can you identify any sources of bias?

<u>Some criticisms are presented in Appendix B but only look at these once you have carried out your own review.</u>

The impact of a work-based fitness program on sickness absence of female employees.

Authors: Drs Charlatan[1], Fraud[1] and Swindler[2]
University Department of Sports Science[1] and Sports Equipment Supplier[2]

Nursing staff employed in the NHS are at increased risk of developing musculo-skeletal injuries and have sickness absence rates which are higher than that for other groups of workers (Peterfield et al, 2011). 'Fitness for duty' is an important prerequisite determined by physical functioning, including joint flexibility, and general physical fitness. However, sharing facilities with men has been cited as a barrier by women who wished to join a workplace fitness program (Parker et al, 2001). Accordingly, a pilot study was undertaken to evaluate the impact of a work-based fitness program on physical functioning and sickness absence amongst women working in a hospital.

A gymnasium was set up to provide facilities within the hospital for women only to use between 7 am and 8 pm each day Monday to Friday and 9 am to 2 pm on Saturdays. Equipment included 2 cycle ergometers, a motorised treadmill, weights and stepping exercisers.

All female employees of the 'X' Hospital NHS Trust were invited by letter to participate in the fitness program. Those who volunteered were given an individual assessment by a physiotherapist working as a 'fitness advisor' for the Trust. Each participant had a 12 lead ECG and measurements taken of resting heart rate, anthropometry, blood pressure and lung capacity. Women with ECG abnormalities (as reviewed by a cardiologist), hypertension, poor lung function, musculo-skeletal or mobility problems were excluded from further testing and were not judged eligible to join the course. For the remainder, lifestyle and current habitual activity were reviewed by questionnaire and physical fitness assessed using a progressive exercise test with a standardised bicycle test protocol (Sullivan et al, 1999). The test estimates the maximal oxygen uptake (VO_2max) from measurements of heart rate and work load during the test procedure. The VO_2max is the 'gold standard' measure of fitness. However, the VO_2max is related to body weight and, to compare different individuals was expressed as estimated maximum oxygen uptake per kg body weight (VO_2max / kg in ml/minute/kg).

In consultation with the physiotherapist each participant had a personal 'fitness plan' drawn up covering aspects of the type and frequency of physical activity to be undertaken and personal targets to be achieved. Training was given in the safe use of the equipment and the choice of exercise undertaken was at the discretion of the participant. The intention was to encourage the participant to exercise in the facilities provided at the hospital at least three times per week for at least 20 minutes on each occasion in keeping with the recommended guidelines of the Health Development Agency (NICE).

The level of exercise required was set to achieve a target heart rate of 75% of maximum using the equation: target heart rate = 0.75 x (220-age). This level was chosen because it should have enabled the participant to feel slightly sweaty and mildly breathless without becoming unduly distressed.

Each participant was asked to keep a daily diary and record their use of the facilities and additional exercise activities outside the hospital.

Volunteers were recruited in February and the program was run over a six-month period (March to August). During this time, the physiotherapist was available weekdays for help and advice. Progress was reviewed at 3 months, when new targets were set, and at the end of the program when the exercise test and physical measures were repeated.

Sickness Absence
At the end of the program the sickness absence record of each participant was reviewed for the 6 months of the program. Control subjects were chosen from the list of all female employees ranked by order of date of birth (youngest first). An age-matched control was then selected for each participant as the next woman listed. The sickness absence record of the control subject was reviewed for the same time period and compared with that for the participant.

Data Analysis
Data were coded onto an Excel database and then transferred into SPSS for statistical analysis. Comparison of means was by t-test and the 5% level ($P<0.05$) accepted as statistically significant.

Results
Forty-three women volunteered for the program of which only one was judged ineligible (hypertension with diastolic blood pressure 106mm Hg). Of the remainder 38 completed the course, two women developed ankle sprains and one woman left employment. The women were aged 19-48 with a mean of 27.9 and standard deviation 4.7 years.

Following completion of the program completed diaries were returned by 29 women. A review of the entries showed that compliance rates with the target activity patterns set by the physiotherapist were reasonable. Of the 29 women 18 (62.07%) achieved all the targets set.

At follow-up there were no changes in mean blood pressure or lung capacity in the participants ($P>0.05$). However, the women had lost, on average, 2.1 kg in weight ($P<0.023$) and the fitness score had improved from 33.3 to 35.2 ml/minute/kg body weight ($P<0.04$). At the same time the resting heart rate had declined from a mean of 78 to 74 beats per minute though this difference just failed to reach statistical significance ($0.10>P>0.05$).

The number of days lost from work because of sickness absence was significantly less in the participants compared with that in the control subjects (Table).

Table: Days lost from work due to sickness in control subjects and participants in a workplace fitness program.

	Participants (n=38)	Controls (n=35)∗
Mean (days)	2.1	3.3
SD (days)	2.7	3.9
difference in means (days)	1.2	
SE (difference)	0.563	
T-value	2.13	
significance	P<0.05	

∗ results of one control woman was discarded because during the trial period she went on long-term sick leave (hysterectomy).

Discussion

This study has shown the benefits of a supervised, personalised workplace fitness program directed at individual women. Benefits were seen in controlling weight and improving fitness scores. Blood pressure did not change though this may have been as a result of excluding those women with higher blood pressures at the outset. However, one important finding was the reduced concurrent sickness absence whereby participants had, on average, 1.2 sick days fewer than control subjects.

The initial set-up costs of a program may be high but this cost may be justified in the long-term when set against the cost of employing 'bank' staff to cover for sick employees.

Conclusion

In conclusion, this study has revealed that the policy of offering workplace fitness programs can have benefits for female staff health. The offer should now be extended to other staff (for example, men) though this would require a formal evaluation of the program before final adoption as Trust policy.

(12) Further reading

A Dictionary of Epidemiology. 6th ed. Porta M (Ed), 2014, Oxford University Press.

How to Read a Paper: the Basics of Evidence-Based Medicine. 6th ed. Greenhalgh, T, 2019, BMJ Books, Blackwell Publishing.

Handbook of Health Research Methods: Investigation, Measurement and Analysis. Bowling A, Ebrahim S (Editors), 2005, Open University Press.

Medical Statistics. An A-Z Companion. Pereira Maxwell F, 2018, Arnold.

Medical Statistics at a Glance. 4th ed. Petrie A, Sabin C, 2019, Blackwell Publishing.

Practical Statistics for Medical Research. Douglas G Altman, 1991, Chapman and Hall.

An Introduction to Medical Statistics. 4th ed. Martin Bland, 2015, Oxford Medical Publications.

Pocket Guide to Critical Appraisal. Crombie IK, 2nd ed 2022, Wiley Blackwell.

Qualitative Inquiry and Research Design: Choosing Among Five Approaches. 4th ed. Creswell JW,, Poth CN. 2017. London: Sage.

Chinn DJ, Cinkotai FF, Lockwood MG, Logan SHM. Airborne dust, its protease content and byssinosis in 'Willowing' mills. Annals of Occupational Hygiene. 1976;**19**:101-8.

Pope C, Ziebland S, Mays N. Qualitative research in health care. Analysing qualitative data. *BMJ* 2000; 320: 114-116.

Appendix A:

Principal sources of bias in the telephone survey to determine the risk profile for diabetes in the general population.

Study design: 4 researchers, working Monday to Friday (9am to 5pm), to telephone 1000 respondents with telephone numbers taken at random from the local BT directory. Questions asked: age, gender, ethnicity, height, weight, waist and hip size, and family history of diabetes.

Source	Comment
Design	Any aspect of study design, e.g. faulty sampling, incorrect randomisation, temporal differences in examination of subgroups, inappropriate calibration of instruments, poor statistically analysis with failure to account for confounding, use of wrong statistical tests.
	The TARGET is the GENERAL POPULATION but: *Not everyone has a telephone* *Use of telephone book – persons registered ex-directory will be missed* *Adolescents and young adults unlikely to be the registered user* *Unlikely to get access to those people who cannot afford a land line 'phone* *Who will be interviewed, the listed name only or anyone in the household when called?*
Selection	Faulty selection when the characteristics of the sample differ from those of the wider, target population. All potential subjects should have an equal chance of being chosen e.g. written invitation not read by illiterate people or those who cannot read English.
	As above, plus *Limiting the calls to 9am – 5pm, Monday to Friday reduces opportunity to catch working people, other than those on night shift! Those at home during these times likely to be mothers with young family, unemployed, disabled, long-term sick etc.* *Reduced opportunity to interview persons with hearing loss* *Reduced opportunity to interview persons whose first language is not English*
Response	A major source of bias leading to a systematic error from differences in characteristics between those who accept and those who decline an invitation to take part in the research.
	Always a source of bias
Measurement	Systematic error from poor calibration regimes, measurement errors, change of instruments between repeated assessments, different instruments used to collect data from different subgroups, data handling procedures, digit preference.
	Bias from reliance from self-report. Respondent bias from lack of knowledge or poor estimation of anthropometry etc. Potential to falsify information which cannot be checked by the interviewer.
Measurement decay	Error from a change in the measurement process over time due to a change in instrument performance or from change in technique by an observer.
	Possible source from change in interview's technique over time
Classification	Categorisation of the results. For example, definition of an ex-smoker (abstinent for one day, one week, one month, six months, one year, ten years?)
	Possible source from recording of ethnicity
Recall	Recall by respondents may be selective or otherwise different between groups with different rates of cognitive decline.
	Respondents may be unaware of their family history of diabetes. Respondent may be adopted, or been estranged from their family or their family members may have died at ages before ever developing diabetes. There is another problem with this type of enquiry. *If a respondent affirms they have a family history of diabetes, then we can interpret this as a confirmed positive response. If a respondent states that, to their knowledge, they do not have a family history of diabetes, or are unaware of such a family history, we cannot be certain that this represents a TRUE absence of a positive family history.*

Reporting	Respondents may be apprehensive about being interviewed and give the responses they think the interviewer wants. Respondents may under-report or over-report symptoms depending on any vested interest.
	Again, a problem relying on self-report
Social desirability	People may wish to present themselves at their best and will respond to questions accordingly.
	Again, a problem relying on self-report
Observer	Differences in measurement techniques between observers, and within observers over time (measurement decay). Different interviewers may show systematic differences in asking questions and recording responses. Interviewers may ask questions in a manner which encourages respondents to answer in a desired way.
	Serious potential source of bias in the way interviewers ask questions and record responses

The table contains only preliminary thoughts. Did you identify any other sources?

It is still possible to carry out this survey which may be considered the most cost-efficient way of collecting the relevant information but all sources of bias must be recognised and the results interpreted with caution accordingly.

Appendix B: Critical Appraisal Exercise

The impact of a work-based fitness program on sickness absence of female employees.

Authors: Drs Charlatan[1], Fraud[1] and Swindler[2]
University Department of Sports Science[1] and Sports Equipment Supplier[2]

First thoughts

1 Who are the authors?

Drs Charlatan, Fraud and Swindler

2 What are their professional backgrounds?

Not specified

3 Where was the work done?

University Department of sports science

4 Who funded it?

Not specified

5 Could there be a conflict of interest?

Possible with involvement of a sports equipment supplier

6 Does the paper describe an important clinical problem?

Yes

7 Is there a clearly formulated, focussed research question?

Yes. Research Question: "the impact of a work-based fitness program on physical functioning and sickness absence amongst women working in a hospital".

8 Are the aims / objectives clearly stated?

Not clearly specified

Methods

9 Is the study design appropriate for the research question?

Study design appropriate - Cohort study / observational / 'before and after' study. Physiological data requires a paired analysis (before and after) with citation of mean and standard deviation of the <u>*change*</u>*. Comparison of sickness absence data requires a paired-analysis (of the difference between a participant and her control) and authors chose to use a between groups analysis.*

10 Is there a clear description of the subjects / patients / target population?

Yes, all women working in the Trust.

11 Is the sample recruited representative of the target population?

*No. **Target group –** the introduction refers to nursing staff (female from the title) but the study recruited 'all female employees' that were working in the hospital Trust.*

***Sampling -** no details of number of women invited initially so no details of response rate to initial letter. Likely to be a highly self-selected group of women already motivated to do more exercise. Unlikely that the sample was representative of the target population.*

Mean age was 27.9 (standard deviation, SD of 4.7) years, hence 95% of the women would be expected to have an age between 18.5 and 37.3 (mean +/- 2 SD), with 2½% younger and 2½% older. Does this range fit with the typical age distribution of women working in the health service? I suspect not. Hence, the <u>sample</u> is unlikely to be representative of the <u>target</u> population (which was ALL female employees).

12 Are the inclusion / exclusion criteria clearly stated?

No "Women with ECG abnormalities (as reviewed by a cardiologist), hypertension, poor lung function, musculo-skeletal or mobility problems were excluded from further testing and were not judged eligible to join the course."

No detail given as to criteria for 'ECG abnormalities', 'hypertension', 'poor lung function', 'musculo-skeletal or mobility problems'. Hence, this study could not be repeated as insufficient detail provide in the methods.

13 Is the control group, if any, comparable?

There was no control group for the physiological test measures.

For the sickness absence outcome measure - "Control subjects were chosen from the list of all female employees ranked by order of date of birth. An age-matched control was then selected for each participant as the next woman listed."

***Control group –** selection of 'age matched controls' from an employee list ordered by date of birth (youngest first) guarantees that controls chosen as the 'woman <u>next listed</u>' will <u>always</u> be older than the participant with whom she is 'matched'. Also, unlikely that a control subject will be comparable to a participant as no account taken of other potential confounders such as full-time/part-time, type of work or grade of post, marital status, disabilities etc. Hence, control group NOT comparable.*

14 Is there a clear description of the treatments?

Yes, for the exercise group.

15 If an intervention is being compared with 'usual care' is a description given of it?

Not relevant

16 Are measurements made concurrently in both intervention and control groups?

Yes, for sickness absence though this was a retrospective review

17 If patients have been randomised is the technique used appropriate?

Not relevant

18 Are the outcome measures reliable / reproducible / valid?

Outcome measures:

'fitness score' - *the measure of interest ('fitness') is estimated VO_2max* *<u>before and after</u> the intervention, but the authors have chosen to divide this by body weight to 'compare different individuals'. However, they are not interested in how different individuals compare with one another, only how a person individually responds to the intervention. Hence the outcome measure should be the VO_2max in each participant.*

'daily diary' – OK

'sickness absence' – OK

19 Are any details given of calibration and quality control procedures?

No

20 Have the observations been made blinded?

No

Data analysis

21 Is there a statement justifying the size of the sample (power calculation)?

No

22 Is the handling of data described (treatment of outliers, missing data etc)?

"Data were coded onto an Excel database and then transferred into SPSS for statistical analysis. Comparison of means was by t-test and the 5% level (P<0.05) accepted as statistically significant."

No description of treatments of missing data or outliers.

23 Is the choice of significance tests (parametric or non-parametric) justified?

No. For the sickness absence data it's likely they have used an *independent between groups t-test to compare sickness absence. The mean of 2.1 (SD of 2.7) days for the participants and 3.3 (SD 3.9) days for controls immediately suggests the distribution is not bell-shaped (Gaussian or normally distributed), hence the t-test is inappropriate. They should have analysed the <u>difference</u> in sickness absence rates between a participant and her control. They could have used a t-test if the distribution of the <u>differences</u> was bell-shaped but, if not, they would need to use a paired, non-parametric test which makes no assumption about the distribution of the data.*

Results

24 Tables and graphs: are they clear and self-explanatory?

No, there are serious problems present.

Outcome measure, 'fitness score' - *the measure of interest ('fitness') is estimated VO_2max <u>before and after</u> the intervention, but the authors have chosen to divide this by body weight to 'compare different individuals'. Dividing the VO_2max by weight guarantees that a reduction in weight (the denominator) from the intervention will result in an increase in the adjusted fitness score without there being any change necessarily in the real measure of interest (which is the VO_2max, the numerator).*

Outcome measure, 'daily diary' - *there is incomplete data capture and no account taken of missing data. The author's state 29 diaries were returned and that compliance rate with the target activity was 62.07% (18/29). But the denominator should be 38, not 29 hence the compliance rate is actually 18/38=47%. There's also this problem where many authors cite percentages to 2 or more decimal places which creates an illusion of accuracy.*

Outcome measure, 'sickness absence' - *it's likely they have used a between groups t-test to compare sickness absence. The mean of 2.1 (SD of 2.7) days for the participants and 3.3 (SD 3.9) days for controls immediately suggests the distribution is not bell-shaped (Gaussian or normally distributed), hence the t-test is inappropriate. They should have analysed the <u>difference</u> in sickness absence rates between a participant and her control. They could have used a t-test if the distribution of the <u>differences</u> was bell-shaped but, if not, they would need to use a paired, non-parametric test which makes no assumption about the distribution of the data.*

25 Do numbers referred to in the text conform to data in the tables and graphs?

No, Results paragraph - numbers in text do not add up, 43 women volunteered, one was judged ineligible, hence 42 enrolled in the programme. 38 completed the course (so 4 drop outs) but details are given for only 3 women who left the programme (2 with ankle strains and one left employment).

26 Do the results include measures of central tendency (mean or medians)?

Medians not given which should have been as the sickness absence data is clearly not distributed as a bell-shape (Normal or Gaussian distribution). Citing the mean for a skewed distribution is unreliable as a measure of central tendency.

27 Do the results include measures of spread (standard deviations, standard errors of the mean or inter-quartile ranges)?

Insufficient detail given for exercise results. The actual measure of interest is the mean and standard deviation of the change in physiological measures of the 38 women but these values are not cited.

28 Do the results include confidence intervals?

No

29 Have any potential confounders been overlooked?

No recognition of bias from age and other attributed in comparison of sickness absence rates.

30 Are the results believable?

Problem with sickness absence results. The study design requires a paired analysis (to analyse the difference between a participant's sickness absence and her control) but it's suggestive from the table (38 participants but only 35 controls) they have used an independent, between-group analysis. The analysis requires an 'intention to treat' approach and the authors should not have rejected the control woman who went on long term sick leave. Matched observations require the same number in each group – why 38 participants but only 35 controls?

Inappropriate citation of the P-values, for example, P=<0.023 should be cited as P=0.023. Changes in mean blood pressure or lung capacity were listed as P>0.05 but the actual P-values should be given. Again, the resting heart rate changed from a mean of 78 to 74 and P is cited as 0.10>P>0.05. The actual value should be given and stating 'this difference just failed to reach statistical significance' is frowned upon by statisticians and constitutes poor practice. The actual measure of interest is the mean and SD of the change in heart rate of the 38 women but these values are not cited.

31 What is the potential for bias?

Plenty.

32 Have any sources of bias been overlooked?

Yes.

No account given of potential biases such as the matching by age of *participants and controls, and the issue of 'inverse causality' whereby women with poor health (and presumably high sickness absence rates) would be less likely to volunteer compared with those with an interest in sports and activity. The volunteers were likely to be relatively fit at the outset, hence a failure to demonstrate change in physical measures such as blood pressure and resting heart rate.*

Discussion

33 Have the authors addressed any weaknesses in their study design?

No

34 Are the conclusions supported by the data?

Conclusions not supported by the study design or data. No acknowledgment of any conflict of interest where one of the authors worked for a sports equipment supplier.

35 Are the findings transferable to other clinical areas?

No, study design and treatment of data fundamentally flawed.

36 Are conclusions relevant to the original questions?

No, treatment of data and analysis flawed

On reading the paper there's a suggestion that the analysis of sickness absence may have been an after-thought as the physiological benefits from participating in the exercise were minimal.

37 Could the findings have occurred because of uncontrolled bias, confounding, or chance alone?

Yes, poorly designed study with flaws in treatment of data, analysis, and conclusions drawn.

Glossary

Tip: search Google for an on-line glossary of research terms not included here

Bias	The unequal distribution of error leading to a deviation from the truth
Blinding	The process by which participants and researchers are made unaware of the treatment received in a clinical trial. Blinding can be 'single' when either the participant or researcher is naive or 'double' when both the participant and researcher are naive to the treatment assigned.
Case control	A study that begins with the identification of patients with a disease (or condition) of interest and a suitable control group without the disease. Cases and controls are 'matched' for important features and compared to measure the relative frequency of occurrence of a characteristic believed to be associated with the disease (or condition) in question.
Causality	The relating of causes to the effects they produce.
Clinical trial	An experiment that involves the administration of a test regime to evaluate its efficacy and safety to participants who are patients
Cohort study	An observational study in which a group or groups of individuals are followed-up with repeated measures over time to determine the relative frequency of occurrence of a disease or condition. The cohort may be studied prospectively or defined in the past and followed-up to the present day (retrospectively).
Confidence interval	A range of values in which the true mean for a population is likely to lie. It usually has a proportion assigned to it (for example 95%) to give it an element of precision.
Confounding	A source of error that occurs when groups being compared differ with regard to an important characteristic related to both the disease in question and the feature under study but which has not been controlled for in the study design. An example is a study comparing a drug with placebo to treat hypertension where one group is significantly older than the other group. Hypertension is age-related and the difference in study outcome (blood pressure) between the drug and placebo may be a consequence of confounding due to the failure to account for the difference in age rather than the effect of the drug.
Critical appraisal	A systematic method of assessing the strengths and weaknesses of a research study by considering issues of validity, accuracy, bias and clinical relevance.
Cross-over study	A design in which study participants are given all treatments

	under investigation but in a sequence with a suitable washout period between treatment periods. Each participant then acts as their own control.
Cross-sectional	An observational study to determine the frequency of a particular disease, characteristic or condition measured in a defined population at one point in time.
Data saturation	Data collection in a qualitative study is continued until the analysis reveals no new themes emerging.
Discourse analysis	The analysis of speech and text to gain an understanding behind the words people use.
Document analysis	Systematic analysis of document contents to answer a research question in a qualitative study
Ethnography	A qualitative research methodology studying people in their natural settings to describe their social interactions and culture. The method is commonly used by anthropologists.
Focus group	A qualitative research method in which participants are questioned by a researcher in a small group allowing interaction between members of the group to elicit views.
Grounded theory	A method of analysis of qualitative data in which the researcher identifies issues that emerge from the data to establish theories that can be tested against further emerging evidence as the analysis progresses.
Hawthorne effect	An effect when participants change their behaviour, consciously or unconsciously, as a result of knowing they are being observed.
Hypothesis	A statement of the relationship between 2 or more study variables. *See Null Hypothesis*
Intention to treat analysis	A method of analysis in a randomised controlled trial whereby all participants are followed-up whether or not they actually received or completed the intervention and their outcome measures are analysed in the group to which they were assigned.
Intervention	A treatment, service or policy intended to improve health status or welfare of an individual, family or community.
Meta-analysis	A statistical technique that pools the results from two or more studies into one overall estimate of the effect of an intervention.
Non-parametric	Statistical method of data analysis that makes no assumptions about the distribution of the data. The method is appropriate when the distribution of the data is skewed (not bell-shaped).

Null Hypothesis, H_O	The statement that assumes there is no difference between two populations being compared, or no relationship or association between two variables in a population. An experiment may be undertaken to see if H_O can be rejected in favour of an alternative hypothesis, H_A.
Parametric methods	Statistical method of data analysis that assumes the distribution of the data is bell-shaped (also called Normal or Gaussian), or approximately so. Examples include the t-test, and Pearson's correlation.
Per-protocol analysis	A method of analysis in a randomised controlled trial whereby participants' outcome measures are analysed according to the treatment they received and not in the group to which they were originally assigned.
Phenomenology	A research methodology which has its roots in philosophy and which focuses on the lived experiences of individuals.
Power	The probability of rejecting the null hypothesis when it is false.
Power calculation	A method of calculating the number of subjects needed for the results of a study to be considered statistically significant.
Qualitative research	A method of studying the meanings people give to their lived experiences, attitudes, expectations and how they make sense of their world. Data may be collected by interview (personal or in a focus group), by participant observation or by reading what they have written. The analysis is non-statistical.
Quantitative research	A method to measure and investigate the relationship between variables. It may involve estimating simple correlations (associations) between measures or investigating causal relationships between one thing (the independent variable) and another (the dependent variable). Results can be expressed in simple descriptive terms or as tests of statistical significance between two or more groups.
Randomised controlled trial	A clinical trial to compare one or more treatments with a control condition. Participants are assigned to a group (treatment or control) by random allocation to minimise bias in the study design.
Semi-structured interview	An interview where the researcher has a set of questions to ask but which can be varied in the order given and where the interviewer can depart from the question set to explore emerging themes.
Structured interview	An interview where the researcher has a set of questions to ask each participant but in which the order and wording is fixed.

Systematic review	A systematic and exhaustive collection of all published work relating to a specific research question. Those papers identified which meet certain pre-defined criteria of quality are subjected to critical review and analysis, usually in a meta-analysis to pool the results of each study to estimate a single, overall effect.
Triangulation	The use of more than one method, theory, data source in a research study to affirm the study results.
Unstructured interview	An interview where the researcher asks participants very general questions without any predetermined plan to allow the participant to shape the interview in whichever way they prefer.