

Systematic review: Bias in imaging studies - the effect of manipulating clinical context, recall bias and reporting intensity

Darren Boone · Steve Halligan · Susan Mallett ·
Stuart A. Taylor · Douglas G. Altman

Received: 23 June 2011 / Revised: 27 August 2011 / Accepted: 12 September 2011 / Published online: 30 September 2011
© European Society of Radiology 2011

Abstract

Objectives To perform a systematic review of diagnostic test accuracy studies which manipulate or investigate the context of interpretation. In particular, those which modify or conceal sample characteristics (e.g. disease prevalence or reporting intensity) or research setting (“laboratory” versus “field”). We also investigated recall bias.

Methods We searched the biomedical literature to March 2010 using 3 complementary strategies. Inclusion criteria were: imaging studies quantifying the effect on diagnosis of modifying the context of observers’ interpretations, varying disease prevalence, concealing sample characteristics, reporting intensity and recall bias.

Results 11247 abstracts were reviewed, 201 full texts examined and 12 ultimately included. There were 5 to 9520 patients and 2 to 129 observers per study. Nine studies investigated clinical review bias of sample level information. Only 3 studies investigated prevalence, 2 of which investigated maximum enrichment well below the levels often used

by researchers. We identified no research specifically directed at concealing disease prevalence. Available research found no evidence of recall bias or “washout” on study results.

Conclusions Several sources of bias central to the design of diagnostic test accuracy studies are poorly researched; the implications for evidence-based-practice remain uncertain. Research is suggested to guide methodological design, particularly in the context of screening.

Key Points

- *Imaging research studies often ignore the possible effect of disease prevalence*
- *It is unclear how the expectation of disease influences radiological interpretation*
- *The potential effect of observer recall bias is poorly researched*
- *Such factors might introduce bias into radiological research methodology*
- *This systematic review attempts to illustrate these points*

Keywords Review, systematic · Bias, observer · Methodology · Imaging, diagnostic · Design, experimental

D. Boone · S. Halligan · S. A. Taylor
Centre for Medical Imaging, University College London,
London, UK

S. Mallett
Department of Primary Care Health Sciences, University of Oxford,
Oxford, UK

D. G. Altman
Centre for Statistics in Medicine, University of Oxford,
Oxford, UK

S. Halligan (✉)
Department of Specialist Radiology, Podium Level 2,
University College Hospital,
235 Euston Road,
London NW1 2BU, UK
e-mail: s.halligan@ucl.ac.uk

Introduction

Studies of diagnostic test accuracy should be designed to minimise bias, a principle that underpins guidance for both reporting [1] and appraising the quality of diagnostic test research [2, 3]. At the same time, study results should ideally be generalisable to everyday clinical practice. Balancing bias against generalisability is not straightforward. For example, to reduce the risk of clinical review bias, it is generally accepted that study observers should be blind to prior investigations [4]. However, concealing information contrasts with daily practice where patients’ clinical history, examination and

prior investigations are known to the observer when formulating a diagnosis. Particularly in the fields of radiology, histopathology and endoscopy, test interpretation involves a significant subjective element that could be influenced by methods which manipulate the clinical context.

In addition to individual patient information, study observers are often unaware of sample characteristics, notably disease prevalence. This issue is potentially important when assessing diagnostic tests intended for screening: In daily practice, observers will expect asymptomatic patients to have low likelihood and lower stage of disease (i.e. more subtle pathology). However, it is unclear how the observer's a priori expectations influence subsequent interpretation, if at all: Some studies have found diminished vigilance when prevalence is low [5] while clustering of abnormal cases in high prevalence situations may also bias interpretation [6]. Nevertheless, studies of diagnostic test accuracy usually increase the prevalence of abnormality to achieve adequate statistical power within a feasible study size [7, 8]. Therefore, results of studies performed in the "laboratory" may not be transferable to lower prevalence, screening populations in "the field."

Other pragmatic issues may also influence generalisability. For example, to complete research within a reasonable timescale, reporting intensity (the number of cases reported within a given timescale) frequently exceeds normal practice and is often exacerbated by the requirement to re-evaluate cases under different conditions (e.g. when comparing MR to CT) [8] or on more than one occasion (e.g. with and without computer aided detection). Moreover, because it is widely believed that prior exposure will influence subsequent interpretation (observer recall bias), it is recommended that consecutive interpretations are separated by a "washout phase" [9]. However, the ideal duration is unknown and there is little evidence that such procedures are effective or necessary.

While these potential "laboratory effects"[10, 11] have been discussed in the methodology literature[6, 11–14], their impact remains unverified. To attempt to quantify their

magnitude, we performed a systematic review of studies where the context of interpretation was manipulated or investigated (i.e. "laboratory" versus "field"). In particular, we wished to investigate the effect of varying sample characteristics, for example, enriching disease prevalence or increasing reporting intensity. Moreover we aimed to explore the effect of concealing sample information (especially prevalence) from observers. We were also interested in studies that addressed "memory effect" due to observer recall bias.

Methods

Data sources and search strategies

D.B. searched the biomedical literature to March 2010 using three complementary search strategies. A primary search identified any existing systematic reviews dealing with our research questions (Table 1). Because our review was not restricted to a specific test, diagnosis or clinical situation (which would facilitate keyword identification), we examined 10 key publications [6, 10, 15–22] known to the authors in the fields of radiology, medical statistics and image perception, that had dealt with case-specific information (Table 2). Relevant keywords/phrases identified from these 10 articles were clinical information, recall bias, intensity, prevalence, prior knowledge and laboratory effect. The MEDLINE database was then searched via PubMed (<http://www.nlm.nih.gov/pubmed>) applying the systematic review filter to each term in turn. "Snowballing," an iterative process for searches of complex material [23], identified potentially relevant publications by reintroducing new key words, repeating the process until no new relevant material emerged.

A secondary search was performed to A) identify indexed literature that shared two or more of the references cited by the 10 key publications and, B) identify all indexed literature citing a key publication (using "related records" and "citation map" searches through Web of Knowledge

Table 1 Primary search strategy: Search for related systematic reviews using six keywords or phrases identified by hand-searching the ten "key publications" described in Table 2

Keyword/phrase queried through Pubmed using the "systematic[sb]" systematic review filter	Total abstracts (including duplicates)	Full text examined for relevance
Report* & intens*	123	1
Recall & bias	71	1
Prevalen*	5142	44
Prior & knowledge	301	2
Lab*& effect*	45	1
Clinical & info*	368	6
Additional relevant references via 'snowballing'		1
Total	6050	56
Articles for data extraction following application of selection criteria		1

Table 2 Secondary search strategy: Details of the 10 “key publications”, the related record search, and the number of publications citing each key publication

Key publication	Number of references cited by key publication	Related record search for publications with ≥ 2 references in common with the key publication	Number of articles citing key publication
Kundel, 1982[21]	2	279	15
Swensson, 1985 [22]	7	567	39
Berbaum,1988a [17]	12	232	45
Berbaum1988b [18]	5	152	42
Berbaum,1989 [16]	8	59	25
Good, 1990 [20]	8	86	37
Samuel, 1995 [15]	10	92	36
Aideyan, 1995 [16]	9	67	16
Egglin, 1996 [6]	16	544	63
Gur, 2008 [10]	5	335	15
Total abstracts reviewed	82	2413	333
Full texts examined	2	27	5
Full texts included	0	4	2

<http://www.isiknowledge.com>). Citations were collated, duplicates eliminated and abstracts reviewed (or titles if abstracts were unavailable) for potential inclusion (Table 2).

Lastly a tertiary search was initiated by retrieving Medical Subject Heading (MeSH) terms from each potentially relevant publication identified by the primary and secondary searches. Terms were ranked in order of frequency and terms likely to be non-discriminatory excluded (e.g. adult, male, female, mammography, CT). Multiple suffixes (e.g. radiology, radiological) were substituted by a truncated heading (e.g. radiol*). Related disciplines (e.g. histopathology, endoscopy) were linked with “OR” operators. Ultimately there were three “modality” terms (endoscop*, radiol* and [cyto* OR histo* OR patho*]) and six “manipulation” terms (prevalen*, attention, Bayes theorem, bias*, observer varia*, and research design), which were paired using the “AND” operator. MEDLINE was searched using these strings using the “diagnosis” option in the “Clinical Queries” filter. Duplicates were excluded and abstracts examined (Table 3). Potentially relevant publications were expanded using the secondary search strategy previously described and any new publication introduced using snowballing [23].

The search strategies were tested: The secondary search identified all 10 key publications. The tertiary search identified all articles from which the MeSH headings had been compiled, and 7 of the 10 key publications.

Inclusion criteria

English language studies to March 2010 inclusive were eligible if they investigated the effect of experimentally modifying the context of observers’ interpretations on diagnosis. In particular, the effects of varying disease

prevalence, blinding to sample characteristics, reporting intensity, and studies investigating recall bias. Studies exploring artificial “laboratory” conditions on outcome were also eligible. However, we excluded studies whose focus was manipulation of case-specific information (e.g. concealment of individual-patient information) since this has been investigated previously by systematic review[4]. Participants were human observers (computer-assisted detection was excluded), making subjective diagnoses based on interpretation of visual data, blind to reference results. Studies were excluded if the number of observers or cases interpreted was unreported. There was no restriction to disease type. We anticipated most studies would be radiological, but subjective interpretation of any medical image (e.g. endoscopy, histopathology) was eligible. Non-medical interpretation was excluded (e.g. airport security X-ray) as were narrative reviews.

Data extraction

D.B. extracted data from the full-text articles consulting S.H. and S.A.T., both experienced in systematic review, if uncertain. Differences of opinion were resolved by consensus. Data were extracted into a data-sheet incorporating measures developed from QUADAS[2] and QAREL[24], with additional fields specific to the review question. We extracted: Author, Journal, imaging modality, topic, number of observers/cases and their characteristics (e.g. professional background and experience), reference standard, case and observer concealment of population characteristics, blinding observers to study participation and purpose, reporting intensity, washout period, prevalence of abnormality and whether this varied, and data clustering (grouping of normal/abnormal cases).

Table 3 Table detailing the Boolean search strings used for the tertiary search strategy and the number of individual abstracts identified by each term, with details of the full texts subsequently examined

'Modality' MeSH term	'Manipulation' MeSH term	Total abstracts (including duplicates)	Full texts retrieved (duplicates removed)	Full text examined for relevance
Endoscopy ¹	& Attention	25	1	0
	& Bayes theorem	6	0	0
	& bias*	84	8	3
	& observer variation	86	3	0
	& prevalen*	64	2	0
	& research design	69	1	1
Radiology ²	& Attention	2	1	1
	& Bayes theorem	0	0	0
	& bias*	708	14	1
	& observer variation	699	36	0
	& prevalen*	89	5	2
	& research design	185	10	0
Pathology ³	& Attention	4	0	0
	& Bayes theorem	21	1	0
	& Bias	96	3	3
	& observer variation	19	10	2
	& prevalen*	131	14	0
	& research design	81	2	0
		2369	111	13
Selection criteria applied				3
Additional references via 'snowballing'				2
Total for data extraction				5

Search String: Endoscopy¹=(endoscop*[MH]); Radiology²=(radiol* [MH]); Pathology³=(cyto* OR histo* OR patho*[MH])

Results

The primary search (Table 1) found 6050 abstracts. 56 full articles were retrieved by D.B.; one was suitable[25]. The secondary search (Table 2) identified 2828 publications with the full text retrieved for 34: ultimately 6 were included [6, 13, 26–29] and 28 rejected because the research focused on case-specific information. The tertiary search (Table 3) identified 74 MeSH terms which were combined into 18 Boolean search strings: These identified 111 potential articles with a further 2 via snowballing; 5 articles were ultimately included [11, 12, 30–32]. Overall, 11247 abstracts were reviewed, 201 full articles retrieved, and 12 ultimately included for systematic review (Table 4).

Description of studies investigating clinical context

Of the 12 identified studies that investigated the effect of manipulating clinical context, 3 focused on varying the prevalence of abnormality [6, 13, 26]. The remaining 9 studies investigated observer performance in different situations with fixed prevalence: 4 compared performance in the laboratory to

daily practice [10, 12, 32]; 3 investigated observer blinding to previous clinical investigations [29–31]; 1 investigated training [27]; 1 investigated varying reporting conditions[25]; 1 investigated recall bias [28]. The 4 studies that investigated interpretation in “the field” used retrospective data obtained from normal clinical practice [10, 12, 25, 32]. One study recruited from an international conference [30]. The remaining 7 used a laboratory environment exclusively.

Study characteristics and settings (Table 4)

The following diagnostic tests were investigated by the 12 included studies: 9 studies were radiological (5 mammographic [10, 12, 25, 28, 29], 3 chest radiology [13, 26, 27], 1 angiographic[6]), 2 endoscopic [30, 32], and 1 histopathological [31]. A single research group contributed 5 studies [10, 13, 26–28].

Study design

All studies used a study design with an independent reference standard excepting a single study of observer

Table 4 Details of the 12 publications included in the systematic review

Publication	Diagnostic test assessed and condition tested	Research focus and relevance to the review	Sample size	Case sample selection	Sample prevalence of abnormality	Observer sample size	Observer qualification and experience	Observer blinding to prevalence of disease	Summary of findings
Gur 1990 [27]	Chest radiography: Lung nodules, interstitial disease and pneumothorax	Laboratory effect; The effect of training observer to use the extent of the ROC scale in observer studies	300	Unclear	Enriched; 80%	4	Board certified, variable experience	Yes	No significant training effect for detecting interstitial disease and pneumothoraces. Accuracy of Lung nodule detection accuracy was affected for two readers and the overall accuracy increased for one reader.
Egglin 1996 [6]	Pulmonary angiography: Pulmonary emboli	Tests prevalence effect, context bias. Effect of deliberate clustering of abnormal cases during observer interpretation of enriched datasets.	24	Unclear	Enriched; 20% or 60%	6	Board certified, variable experience	Yes	Enriching prevalence from 20% to 60% led to an increase in observer sensitivity from 60% to 75%.
Rutter 2000 [12]	Mammography: Breast cancer	Lab vs. field, population blinding, prevalence effect.	1890 in clinic 120 in lab	Consecutive for field cases. Characteristics of laboratory cases unclear	Enriched; 25% in 'lab' cases Population prevalence in 'field' cases	27	Board certification implied	Yes	Mean sensitivity and specificity are both higher in routine practice compared to an artificial research setting.
Meining 2002 [32]	Endoscopic ultrasound: oesophageal and pancreatic cancer	Lab vs. field, effect of blinding. Performance of interpretation in artificial setting both with and without prior information	100	Unclear	Enriched; 100% in 'lab' cases, but not in 'field' cases	2	Board certified, experienced	Yes	Observer performance was reduced in the research setting compared to interpretation in the clinic but this effect was reduced when observers were unblinded to prior information.
Gur, 2003 [13]	Imaging, radiography: Lung nodules, fractures pneumothorax and consolidation	Prevalence effect, blinding to population characteristics. Effect of deliberately enriching prevalence of abnormality	1632	Selected for optimum quality	Enriched, 2 to 28%	14	Board certified, variable experience	Observers instructed to consider the cases as screening tests. Yet prevalence up to 25%	No significant increase in sensitivity when observers report studies in a sample with prevalence enriched up to 28%.
Burnside, 2005 [25]	Mammography: breast cancer	Reporting intensity; Effect of changing clinical reporting environment to high intensity reading	9522	Consecutive	Population risk; 0.05%	5	Board certified, specialist	No; known screening population	Recall rates were 20.1% before and 16.2% after the introduction of high intensity batch reading. Cancer detection rates were not significantly different.
Hardesty, 2005 [28]	Mammography: breast cancer	Memory effect, recall bias. Effect of reading cases which had been previously interpreted in the past and recall of those cases	182	Difficult to interpret cases only (previously incorrectly reported)	5%, enriched compared with screening population	8	Board certified, Experienced 7–20 years	Observers correctly informed the population was enriched	No significant difference in average performance between mammograms observers had interpreted in clinic and those they had not. 7 out of

Table 4 (continued)

Publication	Diagnostic test assessed and condition tested	Research focus and relevance to the review	Sample size	Case sample selection	Sample prevalence of abnormality	Observer sample size	Observer qualification and experience	Observer blinding to prevalence of disease	Summary of findings
Irwig 2006[29]	Mammography, and ultrasound: breast cancer	Blinding: Interpretation bias due to incorrect interpretation of test results in the light of contextual information.	480	Consecutive	Enriched; 50%	2	Board certified, Experienced	Yes	8 observers did not remember previously interpreting any of the mammograms Sensitivity of USS read with mammography was 4.6% higher than without mammography. Comparing combined accuracy of mammography and ultrasound read with and without prior knowledge showed much smaller differences
Bytzer, 2007 [30]	Gastroscopy: ulceration, gastritis, cancer	Effect of providing misleading contextual information. Effect of population blinding and 'study knowledge bias'	5	Attendees at a medical conference	Enriched; 100%	129	Board certified, variable experience	Yes; observers unaware of study participation	Only 23% observers gave the same diagnosis for two identical cases when deliberately misleading contextual information was provided.
Gur 2007[26]	Chest radiography: lung nodules, interstitial disease and pneumothoraces	Prevalence effect, blinding to population characteristics. Effect of deliberately enriching prevalence of abnormality	1632	Selected for optimum technical quality	Enriched; 2 to 28%	14	Board certified, variable experience	Observers instructed to consider the cases as screening investigations yet prevalence up to 28%	Varying prevalence resulted in no significant bias demonstrated in terms of reader accuracy. However, observer confidence that a specific abnormality is truly present is higher in low prevalence (28%) settings
Fandel 2008 [31]	Histopathology: prostate cancer	Lab vs field bias. Interpretation bias due to unavoidable exposure to bias inherent in the interpretation techniques.	178	Selected for optimum technical quality	Enriched; 100%	3	Board certified, specialist	No; two observers involved in study Design	Blinding pathologists to features present on low power in the lab significantly improved accuracy of high power field interpretation
Gur 2008[10]	Mammography: breast Cancer	Lab vs field. Comparison between observer performances when lab interpretations are compared to performance reading the same mammograms in the clinic.	3000	Consecutive	Enriched; 25% in 'lab' cases, population prevalence in 'field' cases	9	Board certified, specialist >3000 read per year. 6 to 32 years experience	Observers instructed to consider the cases as screening investigations yet prevalence up to 28%	Mean sensitivity and specificity were both higher in the clinic compared to a research setting.

agreement [31]. With the exception of one study [31], all observers were blinded to the research hypothesis. Furthermore, one study [30] used observers who were unaware that they were taking part in research. However, despite attempts to overcome “study knowledge bias” [14] (an area of interest to this review) this was not formally quantified by repeating the study with observers who were aware of the study.

Observer and case characteristics (Table 4)

In all studies the observers were medically qualified/board certified with a median of 8 observers per study (inter-quartile range [IQR] 3.5 to 14, range 2 to 129), with 6 studies restricted to observers who were “specialists” [10, 25, 31] or “experienced” [28, 29, 32]; but only 2 studies [10, 28] quantified this. Five studies included less-experienced observers, e.g. residents [6, 13, 26, 27, 30]. In one study, the authors did not detail experience [12]. The median number of cases per study was 300 (IQR 100 to 1761, range 5 to 9520). Case selection criteria were well-defined for 9 (75%) studies. Of these, in 4 studies [10, 12, 25, 29] recruitment was consecutive, 4 [13, 26, 30, 31] selected cases for optimal technical quality, and 1 [28] selected “stress” cases (specifically, cases misinterpreted previously in clinical practice). In all 12 studies technically acceptable material was used, e.g. genuine radiographs, video endoscopy.

Effect of sample disease prevalence (Table 5)

Three articles investigated the effect of varying the prevalence of abnormality on observers’ diagnoses (Table 5). The earliest [6] investigated context bias (to determine if clustering of abnormal cases influenced interpretation of subsequent cases), finding that sensitivity for pulmonary embolus increased significantly (from 60% to 75%) when prevalence was increased from 20% to 60% (7). Two studies by Gur and colleagues [13, 26] increased the prevalence of subtle chest radiographic findings from 2% to 28% in a sample of 3208 cases read by 14 observers of varying experience in a laboratory environment. While no significant effect on observer performance [via receiver operator characteristic (ROC) area under curve (AUC)] was demonstrated [13], reader confidence scores increased at higher prevalence

levels [26]. However, the effects on sensitivity, or indeed the ROC curve itself were not addressed. Furthermore, the maximum prevalence used was 28% but researchers frequently increase prevalence far beyond this level: 6 (50%) studies in this review used prevalence between 50 and 100% [6, 27, 29–32].

Effect of blinding observers to disease prevalence (Table 5)

Of the 12 studies reviewed, 8 (66%) concealed the prevalence of disease. One mammographic study [10], informed observers that the prevalence of abnormality in the sample was enriched (while concealing the exact extent and proportion) but that BiRads ratings should be assigned as if in a screening environment. Of the remaining three studies, observers were told the sample prevalence [28], aware of prevalence because they designed the study [31], or aware of prevalence because the entire study was performed in the clinic [25].

Although 2 studies [13, 26], varied the sample prevalence without informing readers, these studies did not specifically test the effects of revealing the sample prevalence on observers’ interpretation. Hence the effect of blinding readers to the spectrum of abnormality in the study sample remains uncertain.

Effect of reporting intensity (Table 6)

We did not identify any research that specifically manipulated reporting intensity (i.e. burden of interpreting cases) in the laboratory or compared it to daily practice. While a retrospective analysis of mammography in daily practice found that false-positive diagnoses diminished following implementation of high-intensity, batch-reading [25], the change was unquantified. The researchers believed improved performance was due to decreased disruption. Of the remaining 11 studies, 6 detailed setting, observer experience, and case-load enabling an inference of reporting intensity vs. normal practice (Table 6). Observers each read a median of 300 (IQR 100 to 3208) cases at a median rate of 50 (IQR 40 to 50) cases per session. One angiographic study [6] stipulated interpretation within three minutes, which likely exceeded normal practice. Intensity was either unreported or unclear in 5 studies. No article attempted to justify reporting intensity.

Table 5 Articles investigating the effect of manipulating the prevalence of abnormality on studies of diagnostic test accuracy

Publication	Imaging technique	Observers blinded to prevalence of pathology in study sample	Clustering of abnormal cases avoided	Prevalence of abnormality in study sample
Eggin, 1996 [6]	Imaging, angiography	Yes	Deliberate clustering of abnormal cases	60% or 20%
Gur, 2003 [13]	Imaging, chest radiographs	Yes	Yes	2–28%
Gur, 2007 [26]	Imaging, chest radiographs	Yes	Yes	2–28%

Table 6 Estimation of reporting intensity and generalisability to daily practice of “lab” studies

Publication	Total number of cases read per reader	Reporting intensity	Diagnostic test employed in test conditions as per clinical practice	Reporting intensity and environment judged equivalent to daily practice
Gur 1990 [27]	300	50 per session ?interval	Yes	Yes
Eggin 1996 [6]	40	Three minutes per angiogram. Selected images only reviewed.	Selected images only reviewed. No additional views available	No: higher
Rutter 2000 [12]	120	30 per hour every 2 weeks	Yes	Yes
Gur, 2003 [13]	3208	>50 per session, fortnightly over 18 months	Yes	Yes
Gur 2007 [26]	3208	>50 per session, fortnightly over 18 months	Yes	Yes
Gur 2008 [10]	300	20–60 films per session	Yes	Yes

Effect of observer recall bias (Fig. 1)

One article investigated recall bias specifically [28], asking observers to reinterpret mammograms reported by them in clinical practice 14 to 36 months previously. One observer recognised a single mammogram, but subsequently reported it incorrectly. The authors concluded that recall is rare and unlikely to bias studies. The same group [13] tested for 2 week recall via subgroup analysis, finding no effect, but the study was neither designed nor powered for this analysis. 8 (66%) studies included repeated observations of the same cases. One study [30], did not account for recall bias at all, requiring reinterpretation within minutes. The remaining studies incorporated a washout period between observations, with 3 studies using between 2 to 8 weeks and 3 indicating 14 to 36 months,

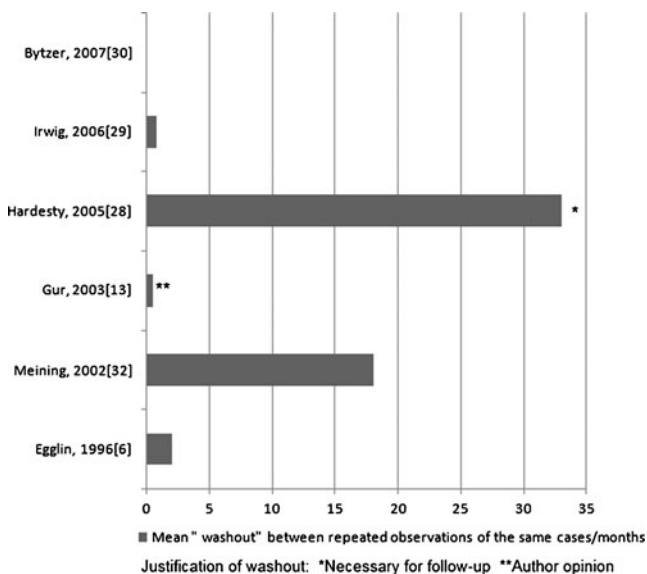


Fig. 1 Duration and scientific justification of the “washout” interval to reduce observer recall bias in studies requiring repeated observations of the same data

and the exact duration unclear in 1 article [Fig. 1]. Moreover, only one article [13] justified the interval and, even then, based this upon anecdotal opinion.

“laboratory” vs “field” settings

All articles considered aspects of generalisability to daily practice, which was the primary focus of 6 articles [Table 4]. Three studies [10, 12, 32] compared “laboratory” interpretation with observers’ prior interpretation of the same cases in clinical practice. Gur [10] and Rutter [12] found higher mean observer sensitivity and specificity in normal clinical practice. However, while Meining et al. also found improved accuracy in the clinical environment, laboratory performance improved significantly when observers had access to clinical information [32].

Irwig [29] questioned whether results from standard tests should be revealed when new diagnostic alternatives are assessed, believing that observers may give undue weight to standard tests with which they are familiar, and so confound the assessment. The authors concluded that such practice is acceptable only when the standard test is both sensitive and specific. One histopathological study examined whether unavoidable initial viewing of low-magnification images may bias subsequent interpretation of high-magnification images [31], arguing that performance would be diminished if studies were restricted to high-power fields. One article [27] explored “checkbox” bias in ROC methodology, concluding that measures encouraging readers to use the full extent of confidence scales might itself introduce bias.

Discussion

We wished to investigate and quantify the effect on diagnostic accuracy results of blinding observers interpreting medical images to sample information, including disease prevalence.

We found that, although manipulation/concealment of individual case information is relatively well-investigated, including a 2004 meta-analysis of 14 studies [4], few researchers have addressed sample information. Our systematic review identified only 12 studies (9 radiological) that investigated generalisability of results from laboratory environments to daily practice and, of these, only 3 focused specifically on prevalence [6, 13, 26], 2 from the same research group. Furthermore, only 2 modalities have been investigated, angiography [6] and chest radiography [13, 26]. The literature base is therefore very insubstantial. We had originally intended to perform a meta-analysis to quantify the effect of the potential biases investigated, but the paucity of available data prevented this.

Enriched prevalence may be an unavoidable aspect of study design, to complete within an acceptable timeframe, with available resources and without undue observer burden. It is important to distinguish between two potential reasons why prevalence might affect sensitivity. Firstly, high prevalence clinical settings are often associated with a more severe disease spectrum, which in itself, will increase sensitivity. Secondly, prevalence may be increased without an increase in disease severity, a situation often encountered in research studies, especially of screening technologies. In this latter situation, it is uncertain how increased prevalence will affect study results. For results to be generalisable we must know the effect, if any, of these enriched study designs on measures of diagnostic test performance, and to what degree and direction. It is widely believed that increasing prevalence raises sensitivity because disease is encountered more frequently than in daily practice [21]; a view supported by Eggin et al. [6]. However, it is only where an increased prevalence is associated with an increase in disease severity that there are theoretical reasons to expect prevalence to affect the ROC curve [33]. It is important to note that although Gur et al. did not demonstrate a significant difference in ROC AUC, despite varying prevalence [13], it does not necessarily follow that a prevalence effect does not exist. Indeed the authors cautioned in a separate editorial [11] that while results obtained in enriched populations should be generalisable to lower prevalence lab-based studies (provided they were analysed using ROC AUC methods) this is not the case for clinical practice. In addition, it is important to consider that while the maximum prevalence was 28%, this level is still well below that often employed by researchers.

Our interest in sample prevalence was precipitated by studies of CT colonography for colorectal cancer screening but we could find no research that addressed the design of these studies. Screening for lung and colorectal cancer by CT, and for breast cancer by mammography, are the subject of considerable research but it is currently impossible to draw evidence-based conclusions regarding the effect of sample prevalence on measures of diagnostic test accuracy.

It is intuitive that observers' prior knowledge of sample prevalence in a study will influence their expectation of disease and we were interested whether this might affect measures of diagnostic accuracy. For example, it is believed that vigilance is reduced if prevalence is low (e.g. screening), because disease is encountered infrequently [34]. Surprisingly, we could identify no research that specifically addressed this issue, either by blinding/unblinding or misleading readers. Most studies concealed prevalence altogether whereas some altered prevalence, but without readers' knowledge. Recall bias (i.e. where interpretation is influenced by recollection of prior interpretations) is a related issue. Many studies incorporated a "washout" phase between consecutive interpretations of identical cases but we could find no research that specifically investigated the impact of varying its duration. It could be argued that the repetitive nature of screening (in terms of material and task) argues for short washout. Indeed, one study concluded recall bias does not exist [28]. We could find no research that specifically addressed the effect of manipulating reporting intensity on measures of diagnostic test performance.

Although anecdotal opinion suggests that observers' performance in an artificial "laboratory" environment (reviewing cases enriched with pathology, far from the pressures of normal daily practice) should exceed that achieved in "the clinic," the available evidence identified by our review [10, 12, 32] suggests the opposite. The fact that clinical information is available in normal practice might help explain this but metaanalysis suggests the effect is small [4]. Another possible explanation is that observers in laboratory studies are aware their assessments will have no clinical consequences; "study knowledge bias" is also likely to influence observer studies but we found no research to substantiate this. Lastly, a substantial reporting burden associated with research studies (often performed at unsocial hours so as not to interfere with normal duties) may explain why accuracy is diminished. This discrepancy between "lab" and "field" performance has important implications, not only for evaluation of diagnostic tests, but also for how radiologists' performance is assessed in isolation. For example, the PERFORMS programme for evaluating mammographic interpretation uses a cancer prevalence of 22% [35] and so may not reflect radiologist performance in clinical practice. Toms et al. suggested a more accurate assessment would be obtained by sporadically introducing abnormal test cases into normal daily reporting [36].

Our review revealed that the existing evidence-base is too insubstantial to guide many aspects of study design. High-quality research is needed to investigate and quantify the biases we investigated. Inevitably, studies specifically designed to answer the questions we posed will be expensive and time-consuming. For example, most studies we identified used observer samples in the single digits and

variance is likely to be high; much larger studies are required. The authors predict that funding would be difficult to achieve for large-scale methodological research specifically designed to quantify these potential biases. However, given that funding agencies have previously provided very substantial support for large-scale studies of screening technologies, the authors suggest that future studies incorporate additional research that aims to estimate bias and generalisability. For example, this could be achieved via sub-studies/parallel/nested studies that incorporate unblinded observers, different contexts, or by varying the duration of washout period for different groups of observers. Such an approach would combine large-scale diagnostic test accuracy studies with methodological research for relatively little additional cost.

Our review does have limitations. In particular, relevant research may have been missed because of a lack of search terms specific to our review question. For example, many papers will discuss potential bias but few will test this as a primary outcome. Aware of this, we used multiple search strategies and snowballing to maximise studies retrieved. Even so, the total body of relevant literature we identified was rather small and was heterogeneous in the issues addressed.

In summary, several issues central to the design of studies of diagnostic test accuracy have not been well-researched and there is an insufficient evidence-base to guide many aspects of study design. High quality research is needed to address potential bias resulting from observers' knowledge of prevalence and the effects of recall bias across several imaging technologies and diseases, most notably for studies of screening.

Acknowledgements This article represents independent research commissioned by the National Institute for Health (NIHR) Research under its Programme Grants for Applied Research funding scheme (RP-PG-0407-10338). This work was undertaken at University College London Hospital (UCLH) and University College London (UCL), which receive a proportion of funding from the NIHR Comprehensive Biomedical Research Centre funding scheme. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

References

- Bossuyt PM, Reitsma JB, Bruns DE et al (2003) Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. *Radiology* 226:24–28
- Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J (2003) The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol* 3:25
- Lucas NP, Macaskill P, Irwig L, Bogduk N (2010) The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). *J Clin Epidemiol* 63(8):854–861
- Loy CT, Irwig L (2004) Accuracy of diagnostic tests read with and without clinical information: a systematic review. *JAMA* 292(13):1602–1609
- Wolfe JM, Horowitz TS, Kenner NM (2005) Cognitive psychology: rare items often missed in visual searches. *Nature* 435(7041):439–440
- Egglin TKP, Feinstein AR (1996) Context bias—A problem in diagnostic radiology. *Jama-Journal of the American Medical Association* 276(21):1752–1755
- Wagner RF, Beiden SV, Campbell G, Metz CE, Sacks WM (2002) Assessment of medical imaging and computer-assist systems: lessons from recent experience. *Acad Radiol* 9(11):1264–1277
- Bossuyt PM, Irwig L, Craig J, Glasziou P (2006) Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ* 332(7549):1089–1092
- Metz CE (1989) Some practical issues of experimental design and data analysis in radiological ROC studies. *Invest Radiol* 24(3):234–245
- Gur D, Bandos AI, Cohen CS et al (2008) The "Laboratory" effect: Comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology* 249(1):47–53
- Gur D, Rockette HE, Warfel T, Lacomis JM, Fuhrman CR (2003) From the laboratory to the clinic: The "prevalence effect". *Acad Radiol* 10(11):1324–1326
- Rutter CM, Taplin S (2000) Assessing mammographers' accuracy. A comparison of clinical and test performance. *J Clin Epidemiol* 53(5):443–450
- Gur D, Rockette HE, Armfield DR et al (2003) Prevalence effect in a laboratory environment. *Radiology* 228(1):10–14
- Gur D (2004) Imaging technology and practice assessments: diagnostic performance, clinical relevance, and generalizability in a changing environment. *Radiology* 233(2):309–312
- Samuel S, Kundel HL, Nodine CF, Toto LC (1995) Mechanism of satisfaction of search: eye position recordings in the reading of chest radiographs. *Radiology* 194(3):895–902
- Aideyan UO, Berbaum K, Smith WL (1995) Influence of prior radiologic information on the interpretation of radiographic examinations. *Acad Radiol* 2(3):205–208
- Berbaum KS, Elkhoury GY, Franken EA, Kathol M, Montgomery WJ, Hesson W (1988) Impact of clinical history on fracture detection with radiography. *Radiology* 168(2):507–511
- Berbaum KS, Franken EA, Dorfman DD, Barloon TJ (1988) Influence of clinical history upon detection of nodules and other lesions. *Investig Radiol* 23(1):48–55
- Berbaum KS, Franken EA, Elkhoury GY (1989) Impact of clinical history on radiographic detection of fractures—a comparison of radiologists and orthopedists. *Am J Roentgenol* 153(6):1221–1224
- Good BC, Cooperstein LA, DeMarino GB et al (1990) Does knowledge of the clinical history affect the accuracy of chest radiograph interpretation? *Am J Roentgenol* 154(4):709–712
- Kundel HL (1982) Disease prevalence and radiological decision-making. *Investig Radiol* 17(1):107–109
- Swensson RG, Hessel SJ, Herman PG (1985) The value of searching films without specific preconceptions. *Investig Radiol* 20(1):100–107
- Greenhalgh T, Peacock R (2005) Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. *BMJ* 331(7524):1064–1065
- Lucas NP, Macaskill P, Irwig L, Bogduk N The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). *J Clin Epidemiol*
- Burnside ES, Park JM, Fine JP, Sisney GA (2005) The use of batch reading to improve the performance of screening mammography. *Am J Roentgenol* 185(3):790–796
- Gur D, Bandos AI, Fuhrman CR, Klym AH, King JL, Rockette HE (2007) The prevalence effect in a laboratory environment: Changing the confidence ratings. *Acad Radiol* 14(1):49–53

27. Gur D, Rockette HE, Good WF et al (1990) Effect of observer instruction on ROC study of chest images. *Invest Radiol* 25(3):230–234
28. Hardesty LA, Ganott MA, Hakim CM, Cohen CS, Clearfield RJ, Gur D (2005) "Memory effect" in observer performance studies of mammograms. *Acad Radiol* 12(3):286–290
29. Irwig L, Macaskill P, Walter SD, Houssami N (2006) New methods give better estimates of changes in diagnostic accuracy when prior information is provided. *J Clin Epidemiol* 59(3):299–307
30. Bytzer P (2007) Information bias in endoscopic assessment. *Am J Gastroenterol* 102(8):1585–1587
31. Fandel TM, Pfnur M, Schafer SC et al (2008) Do we truly see what we think we see? The role of cognitive bias in pathological interpretation. *J Pathol* 216(2):193–200
32. Meining A, Dittler HJ, Wolf A et al (2002) You get what you expect? A critical appraisal of imaging methodology in endosonographic cancer staging. *Gut* 50(5):599–603
33. Metz CE (2006) Receiver Operating Characteristic Analysis: A Tool for the Quantitative Evaluation of Observer Performance and Imaging Systems. *J Am Coll Radiol* 3(6):413–422
34. Rich AN, Kunar MA, Van Wert MJ, Hidalgo-Sotelo B, Horowitz TS, Wolfe JM (2008) Why do we miss rare targets? Exploring the boundaries of the low prevalence effect. *J Vis* 8(15):11–17
35. Esserman L, Cowley H, Eberle C et al (2002) Improving the Accuracy of Mammography: Volume and Outcome Relationships. *Journal of the National Cancer Institute* 94(5):369–375
36. Toms AP (2010) The war on terror and radiological error? *Clin Radiol* 65(8):666–668