

Keywords: breast cancer; neoadjuvant chemotherapy; magnetic resonance imaging; tumour response; monitoring

Meta-analysis of agreement between MRI and pathologic breast tumour size after neoadjuvant chemotherapy

M L Marinovich^{*1}, P Macaskill¹, L Irwig¹, F Sardanelli², G von Minckwitz³, E Mamounas⁴, M Brennan¹, S Ciatto^{5,*} and N Houssami¹

¹Screening and Test Evaluation Program (STEP), Sydney School of Public Health, The University of Sydney, A27, Edward Ford Building, Camperdown, NSW 2006, Australia; ²Dipartimento di Scienze Biomediche per la Salute, Università degli Studi di Milano, Unità di Radiologia, IRCCS Policlinico San Donato, Piazza E. Malan 2, San Donato Milanese, Milano, Italy; ³German Breast Group and Universitäts-Frauenklinik Frankfurt, Martin-Behaim-Strasse 12, 63263 Neu-Isenburg, Germany; ⁴MD Anderson Cancer Center Orlando, 1400 South Orange Avenue, MP 700, Orlando, FL 32806, USA and ⁵UO Senologia Clinica e Screening Mammografico, Dipartimento di Radiodiagnostica, APSS, Viale Verona, 30123 Trento, Italy

Background: Magnetic resonance imaging (MRI) has been proposed to guide breast cancer surgery by measuring residual tumour after neoadjuvant chemotherapy. This study-level meta-analysis examines MRI's agreement with pathology, compares MRI with alternative tests and investigates consistency between different measures of agreement.

Methods: A systematic literature search was undertaken. Mean differences (MDs) in tumour size between MRI or comparator tests and pathology were pooled by assuming a fixed effect. Limits of agreement (LOA) were estimated from a pooled variance by assuming equal variance of the differences across studies.

Results: Data were extracted from 19 studies (958 patients). The pooled MD between MRI and pathology from six studies was 0.1 cm (95% LOA: –4.2 to 4.4 cm). Similar overestimation for MRI (MD: 0.1 cm) and ultrasound (US) (MD: 0.1 cm) was observed, with comparable LOA (two studies). Overestimation was lower for MRI (MD: 0.1 cm) than mammography (MD: 0.4 cm; two studies). Overestimation by MRI (MD: 0.1 cm) was smaller than underestimation by clinical examination (MD: –0.3 cm). The LOA for mammography and clinical examination were wider than that for MRI. Percentage agreement between MRI and pathology was greater than that of comparator tests (six studies). The range of Pearson's/Spearman's correlations was wide (0.21–0.92; 16 studies). Inconsistencies between MDs, percentage agreement and correlations were common.

Conclusion: Magnetic resonance imaging appears to slightly overestimate pathologic size, but measurement errors may be large enough to be clinically significant. Comparable performance by US was observed, but agreement with pathology was poorer for mammography and clinical examination. Percentage agreement can provide supplementary information to MDs and LOA, but Pearson's/Spearman's correlation does not provide evidence of agreement and should be avoided. Further comparisons of MRI and other tests using the recommended methods are warranted.

Magnetic resonance imaging (MRI) has been proposed to have a role in guiding breast cancer surgical extent by measuring the size of the residual tumour after neoadjuvant chemotherapy (NAC),

and has been shown to have good sensitivity for detecting residual disease in that setting (Marinovich *et al*, 2013). Given that current guidelines for response evaluation recommend assessment of the

*Correspondence: ML Marinovich; E-mail: luke.marinovich@sydney.edu.au

*Deceased.

Received 18 June 2013; revised 21 July 2013; accepted 23 July 2013; published online 20 August 2013

© 2013 Cancer Research UK. All rights reserved 0007–0920/13

largest tumour diameter (Eisenhauer *et al*, 2009), estimation of the largest diameter by MRI may guide decisions about whether subsequent mastectomy or breast conserving surgery (BCS) should be attempted, as well as assist in the planning of resection volume to achieve clear surgical margins in BCS. Underestimation of tumour size may therefore lead to involved surgical margins and repeat surgery; overestimation may lead to overly radical surgery (including mastectomy when BCS may have been possible) and poorer cosmetic and psychosocial outcomes (Irwig and Bennetts, 1997).

The assessment of tumour size before surgery is subject to a number of potential errors (Padhani and Husband, 2000). Reactive inflammation, fibrosis or necrosis in response to NAC may present as areas of enhancement on MRI images, which may be difficult to distinguish from residual tumour (Yeh *et al*, 2005; Belli *et al*, 2006). Regression of the tumour as multiple, scattered tumour deposits may also make assessment of the longest diameter problematic, with different approaches to measurement that either include (Rosen *et al*, 2003; Wright *et al*, 2010) or exclude (Cheung *et al*, 2003; Bollet *et al*, 2007) intervening normal tissue. Ductal carcinoma *in situ* (DCIS) may not be well visualised (Berg *et al*, 2010) or, alternatively, may be indistinguishable from invasive cancer (Partridge *et al*, 2002). Imaging artefacts may also introduce errors in tumour size estimation. For example, the placement of markers in or around the tumour may produce areas of increased signal intensity, which are difficult to distinguish from residual foci, or areas of low signal, which may contribute to size underestimation. Underestimation may also occur owing to partial volume effects (Lobbes *et al*, 2012). Furthermore, the inherently pliable nature of breast tissue means that tumour dimensions may vary, depending on patient positioning (Tucker, 2012).

In this systematic review and study-level meta-analysis, we investigate agreement in the measurement of residual tumour size by MRI and pathology (the reference standard) after NAC for breast cancer, as assessed by mean differences (MDs) and 95% limits of agreement (LOA) (Bland and Altman, 1986). We also compare the agreement between pathology and alternative tests which have been used to measure residual tumour before surgery (ultrasound (US), clinical examination and mammography). The consistency of results from different methods to assess agreement is investigated, and recommendations are made about methods for future studies.

MATERIALS AND METHODS

Identification of studies. A systematic search of the biomedical literature up to February 2011 was undertaken to identify studies assessing the accuracy of MRI after NAC in measuring the size of residual tumour. MEDLINE and EMBASE were searched via EMBASE.com; PREMEDLINE, Database of Abstracts of Reviews of Effects, Health Technology Assessment (CLHTA) and the Cochrane databases were searched via Ovid. Search terms were selected to link MRI with breast cancer and response to NAC. Keywords and medical subject headings included 'breast cancer', 'nuclear magnetic resonance imaging', 'MRI', 'neoadjuvant' and 'response'. The full search strategy has been reported previously (Marinovich *et al*, 2012, 2013). Reference lists were also searched and content experts consulted to identify additional studies.

Review of studies and eligibility criteria. All abstracts were screened for eligibility by one author (LM), and a sample of 10% was assessed independently by a second author (NH) to ensure consistent application of the eligibility criteria. Eligible studies were required to have enrolled a minimum of 15 patients with newly diagnosed breast cancer undergoing NAC, with MRI and at least one other test (US, mammography and clinical examination)

undertaken after NAC to assess the size of residual tumour before surgery. Pathologically measured tumour size based on surgical excision was the reference standard, but studies were not excluded if alternative reference standards were used in a minority of patients.

Potentially eligible citations were reviewed in full (LM or NH). The screening and inclusion process is summarised in Supplementary Information Resource 1 (PRISMA flowchart).

Data extraction. Data relating to tumour size assessment, study design, patient characteristics, tumours, treatment, technical details of MRI, comparator tests and the reference standard were extracted independently by two authors (LM, and either SC, MB or FS). Quality appraisal was undertaken using the Quality Assessment of Diagnostic Accuracy Studies checklist (version 1, modified for this clinical setting; Whiting *et al*, 2003, 2006). Disagreements were resolved by discussion and consensus, with arbitration by a third author (NH) when required.

Measures of agreement. Bland and Altman (1986) describe appropriate methods to assess agreement between two continuous measures and highlight the inadequacy of the Pearson's correlation coefficient when used for this purpose. Unlike methods such as intraclass correlation (ICC), the Pearson's correlation coefficient measures the degree to which there is a linear, but not necessarily 1:1 relationship. Hence, it is possible for a high Pearson's correlation to be observed when there is poor agreement between two measures (e.g., when tests systematically under- or overestimate pathologic size). Spearman's rank correlation is similarly problematic. A commonly reported alternative approach involves calculating the percentage of cases for which there is 'agreement' between measures within a chosen 'margin of error'. This approach also has limitations, as the chosen margin of error may be somewhat arbitrary, and tendencies for one measure to under- or overestimate the other within that margin may be obscured.

The approach recommended by Bland and Altman (1986) comprises a scatterplot of the differences between the measures (the vertical axis) against their mean (horizontal axis). If the differences are normally distributed and are independent from the underlying size of the measurements, agreement may be quantified by the MD and associated 95% LOA. Hence, MDs and LOA were extracted from studies reporting these outcomes. When LOA were not presented, data were extracted from which the LOA could be derived (e.g., s.d. of the difference or root mean square error). Despite their limitations, percentage agreement within a margin of error (and associated percentages of under/overestimation) and correlation coefficients were also extracted to provide a descriptive summary of these measures.

Statistical analysis. MDs between tumour size measurements by MRI or comparator tests and pathology were pooled by the inverse variance method by assuming a fixed effect using RevMan 5.2 (The Nordic Cochrane Centre (Copenhagen), The Cochrane Collaboration, 2012) (http://ims.cochrane.org/sites/ims.cochrane.org/files/uploads/documents/revman/RevMan_5.2_User_Guide.pdf). The Cochrane Q statistic was used to assess whether statistically significant heterogeneity was present (significant at $P < 0.10$), and the extent of heterogeneity was quantified by the I^2 statistic (Higgins *et al*, 2003). To estimate the 95% LOA for a pooled MD, a pooled variance was computed under the assumption that the variance of the differences was equal across studies. The pooled variance was calculated as the weighted average of these within-study variances, weighted by the corresponding degrees of freedom for each study (i.e., an extension of the approach used for a two sample Student's *t*-test (Woodward, 1999)).

RESULTS

Study characteristics. A total of 2108 citations were identified. Nineteen studies were eligible for inclusion in the systematic review (Weatherall *et al*, 2001; Balu-Maestro *et al*, 2002; Partridge *et al*, 2002; Rosen *et al*, 2003; Bodini *et al*, 2004; Chen *et al*, 2004; Londero *et al*, 2004; Julius *et al*, 2005; Montemurro *et al*, Yeh *et al*, 2005; Akazawa *et al*, 2006; Bollet *et al*, 2007; Segara *et al*, 2007; Bhattacharyya *et al*, 2008; Moon *et al*, 2009; Prati *et al*, 2009; Nakahara *et al*, 2010; Wright *et al*, 2010; Guarneri *et al*, 2011), reporting data on 958 patients undergoing MRI and/or comparator tests; MRI data were reported for 953 patients. Studies enrolled patients between 1998 and 2007 (median mid-point of recruitment 2002), and included a median of 38 patients with MRI data (range 12–195). Characteristics of included studies are summarised in Table 1. Study quality appraisal is summarised in Supplementary Information Resource 2.

MRI details. Technical characteristics of MRI are summarised in Supplementary Information Resource 3. The majority of studies used DCE-MRI (84.2%) with a 1.5-T magnet (73.7%). Dedicated bilateral breast coils were used in all studies in which the coil type was reported. All studies providing detail on contrast employed gadolinium-based materials, most commonly gadopentetate dimeglumine (68.4%), typically at the standard dosage of 0.1 mmol per kg body weight (68.4%).

Reference standard. Pathology from surgical excision was the reference standard for all patients in all but one study (Bhattacharyya *et al*, 2008), where the absence of residual tumour (pathologic complete response, pCR) in two patients was verified by localisation biopsy, representing 0.2% of patients included in all studies. Study-specific rates of pCR ranged between 0.0% and 28.6%, with a median 14.3% (Table 1).

Mean differences between MRI and pathology. Six studies (Partridge *et al*, 2002; Akazawa *et al*, 2006; Segara *et al*, 2007; Prati *et al*, 2009; Wright *et al*, 2010; Guarneri *et al*, 2011) reported MDs and LOA between MRI and pathology (Supplementary Information Resource 4). All studies measured the longest tumour diameter, except for a study by Akazawa *et al* (2006) that measured the diameter along the plane connecting the nipple and the tumour centre. This study is therefore presented descriptively, but has been excluded from pooled analyses.

Meta-analysis of MDs between MRI and pathologic tumour measurement (Figure 1) showed a tendency for MRI to slightly overestimate pathologic tumour size, with a pooled MD of 0.1 cm (95% CI: -0.1–0.3 cm). There was no evidence of heterogeneity ($I^2 = 0\%$). Pooled LOA indicated that 95% of pathologic measurements fall between -4.2 cm and +4.4 cm of the MRI measurement.

Within-study comparisons of MRI versus US, clinical examination and mammography are presented in Supplementary Information Resource 4. For all but a single study showing similar, small tendencies for overestimation by MRI (0.16 cm) and US (0.06 cm) (Guarneri *et al*, 2011), the absolute values of MDs within studies were lower for MRI than that for the alternative tests. Pooled MDs and 95% LOA are summarised in Table 2 and Figures 2–4. There was no evidence of heterogeneity for MRI in any of the analyses, or for US (all $I^2 = 0\%$). Pooled results from two studies (Segara *et al*, 2007; Guarneri *et al*, 2011) showed similar small overestimation of pathologic tumour size by MRI and US (MDs of 0.1 cm for both tests), with comparable LOA. Pooled MDs and LOA from two studies (Prati *et al*, 2009; Wright *et al*, 2010) were larger for mammography (0.4, 95% LOA - 7.1 to 8.0 cm) than for MRI (0.1 cm, 95% LOA - 6.0 to 6.3 cm), with moderate heterogeneity in MDs for mammography ($I^2 = 39\%$). Pooled

estimates for MRI and clinical examination across four studies (Partridge *et al*, 2002; Segara *et al*, 2007; Prati *et al*, 2009; Wright *et al*, 2010) resulted in substantial heterogeneity for the latter test ($Q = 20.59$, $df = 3$, $P = 0.0001$; $I^2 = 85\%$); three studies reported that clinical examination underestimated pathologic tumour size, and one study reported the reverse. Pooled MDs showed larger underestimation with wider LOA for clinical examination (-0.3 cm, 95% LOA: -5.3 to 4.7 cm) relative to MRI overestimation (0.1 cm, 95% LOA: -4.5 to 4.6 cm).

Percentage agreement. Eight studies (Balu-Maestro *et al*, 2002; Rosen *et al*, 2003; Julius *et al*, 2005; Yeh *et al*, 2005; Akazawa *et al*, 2006; Segara *et al*, 2007; Nakahara *et al*, 2010; Guarneri *et al*, 2011) reported percentage agreement between tumour size measured by MRI and pathology within a variety of margins of error based on absolute size (± 0 , 0.5, 1, 2 and 3 cm) or a percentage of the pathologic measurement (± 30 and 50%; Supplementary Information Resource 4). One study did not report the margin of error used to calculate agreement (Balu-Maestro *et al*, 2002), and two studies reported percentage agreement between MRI and pathology but not the associated percentages of MRI under/overestimation (Julius *et al*, 2005; Akazawa *et al*, 2006).

Studies reporting percentage agreement (plus under/overestimation) for MRI, US and clinical examination by an absolute margin of error are summarised in Figure 5 (no studies reported these data for mammography). As would be expected, percentage agreement between all tests and pathology was observed to be higher for wider margins of error (e.g., ~20% for exact agreement between MRI and pathologic measurements (Segara *et al*, 2007; Guarneri *et al*, 2011) vs 92% for ± 3 cm (Nakahara *et al*, 2010)). With the exception of one study showing a tendency for overestimation (Rosen *et al*, 2003), MRI appeared equally likely to overestimate and underestimate pathologic tumour size across all absolute margins of error. For US and clinical examination, a tendency towards underestimation can be observed in Figure 5, but the majority of estimates showing that bias were contributed by a single study (Segara *et al*, 2007).

Percentage agreement estimates for MRI based on any margin of error were compared with those of alternative tests in six studies (Supplementary Information Resource 4). All six studies compared MRI and US (Balu-Maestro *et al*, 2002; Julius *et al*, 2005; Yeh *et al*, 2005; Akazawa *et al*, 2006; Segara *et al*, 2007; Guarneri *et al*, 2011); MRI was compared with clinical examination in four studies (Balu-Maestro *et al*, 2002; Yeh *et al*, 2005; Akazawa *et al*, 2006; Segara *et al*, 2007) and with mammography in three studies (Balu-Maestro *et al*, 2002; Julius *et al*, 2005; Yeh *et al*, 2005). For all but one study and across the range of reported margins of error, percentage agreement estimates for MRI were higher than those for the comparator tests. In the one exception to this pattern of results, a study reporting multiple margins of error (Segara *et al*, 2007) found higher percentage agreement for MRI than for US at margins of ± 0 and ± 1 cm, but percentage agreement at ± 2 cm was slightly higher for US (92%) than that for MRI (88%). In one other study (Guarneri *et al*, 2011), the difference in percentage agreement favouring MRI over US was relatively small (20% vs 15% at ± 0 cm; 54% vs 51% at ± 0.5 cm; and 71% vs 68% at ± 1 cm).

Correlation coefficients. Sixteen studies (Weatherall *et al*, 2001; Partridge *et al*, 2002; Rosen *et al*, 2003; Bodini *et al*, 2004; Chen *et al*, 2004; Londero *et al*, 2004; Montemurro *et al*, 2005; Akazawa *et al*, 2006; Bollet *et al*, 2007; Segara *et al*, 2007; Bhattacharyya *et al*, 2008; Moon *et al*, 2009; Prati *et al*, 2009; Nakahara *et al*, 2010; Wright *et al*, 2010; Guarneri *et al*, 2011) reported correlations between MRI and pathologic tumour size, and similar correlations for at least one alternative test, either by the Pearson's ($N = 9$) or Spearman's ($N = 5$) method (in two studies (Weatherall *et al*, 2001; Partridge *et al*, 2002), the method was not specified). The range of

Table 1. Summary of cohort, tumour, treatment and reference standard characteristics of included studies

Variable	Number providing data			IQR	Range
	Studies	Patients	Median estimate		
Cohort characteristics					
N (MRI)	19	953	38	21–60	12–195
Recruitment mid-point (year)	12	680	2002	2001–2005	1998–2007
Age, mean (or median) (years)	16	834	48	45–49	42–56
Menopausal status (%)^a					
Pre	5	254	60.4	59.3–68.8	55.3–75.4
Peri/post	5	118	39.6	31.2–40.7	24.6–44.7
Tumour size^a					
Clinical size, mean (or median) (cm)	9	343	4.9	4.7–6.2	4.3–8.2
T stage (%)^a					
T1	9	50	2.1	0.0–2.6	0.0–50.7
T2	9	323	48.2	10.0–72.9	0.0–84.9
T3	9	166	27.1	12.3–47.9	7.2–68.9
T4	9	93	13.2	2.7–30.0	0.0–43.8
Tx	9	1	0.0	0.0–0.0	0.0–0.5
Stage (%)^a					
I	7	2	0.0	0.0–0.0	0.0–6.2
II	6	202	81.4	62.5–86.4	47.6–86.7
III	6	55	18.6	13.6–31.2	13.3–52.4
IV	8	0	0.0	0.0–0.0	0.0–0.0
Histology (%)^a					
IDC	15	552	82.2	71.2–90.0	48.6–96.5
ILC or IDC/ILC	15	79	10.0	5.1–18.8	0.0–26.0
Other	15	31	3.5	0.0–10.2	0.0–16.1
Unknown or NR	15	16	0.0	0.0–0.0	0.0–15.9
Nodal status (%)^a					
Positive	6	316	62.0	45.8–71.1	38.4–93.8
Negative	6	128	36.5	28.9–54.2	6.2–61.6
Unknown or NR	6	1	0.0	0.0–0.0	0.0–3.1
Grade (%)^a					
I	5	23	13.3	6.2–18.8	0.0–22.0
II	5	61	37.5	22.2–43.8	15.3–49.2
III	5	95	43.8	25.4–51.1	25.0–78.0
Unknown or NR	5	16	6.8	3.4–13.3	0.0–25.0
ER (%)^a					
Positive	8	333	63.6	55.9–67.6	40.6–75.0
Negative	7	240	33.9	32.2–48.2	25.0–59.4
Unknown or NR	7	4	0.0	0.0–3.4	0.0–3.4
PR (%)^a					
Positive	6	155	39.2	34.9–47.5	6.8–68.8
Negative	6	247	53.5	31.2–63.8	27.1–65.1
Unknown or NR	6	41	0.0	0.0–3.4	0.0–66.1
HER2 (%)^a					
Positive	8	199	29.7	19.6–39.2	12.5–73.9
Negative	8	373	70.3	58.3–78.7	26.1–87.5
Unknown or NR	8	5	0.0	0.0–1.7	0.0–5.1
NAC regimen (%)^a					
Anthracycline-based	19	316	9.7	0.0–82.7	0.0–100.0
Anthracycline-taxane-based	19	437	20.0	0.0–87.2	0.0–100.0
Other	19	210	1.7	0.0–10.5	0.0–100.0

Table 1. (Continued)

Variable	Number providing data			Median estimate	IQR	Range
	Studies	Patients				
Studies using Trastuzumab with NAC^a						
Trastuzumab used (%)	5 ^b	80		5.6	4.7–42.4	2.1–57.4
Trastuzumab not used (%)	5 ^b	376		94.4	57.6–95.3	42.6–97.9
Type of surgery (%)^a						
BCS	13	281		37.3	23.8–58.1	6.0–100.0
Mastectomy	13	281		62.7	41.9–76.2	0.0–94.0
No surgery	13	2 ^c		0.0	0.0–0.0	0.0–6.2
Type of reference standard (%)^a						
Pathology	19	951		100.0	100.0–100.0	93.8–100.0
Other	19	2 ^c		0.0	0.0–0.0	0.0–6.2
Time from MRI to surgery						
Days, mean (or median/estimate)	8	255		22.0	14–28	7–28
Prevalence of pCR (%)^a						
pCR	19	957		14.3	8.3–18.8	0.0–28.6

Abbreviations: BCS = breast conserving surgery; DCIS = ductal carcinoma *in situ*; ER = oestrogen receptor; HER2 = human epidermal growth factor receptor 2; IDC = invasive ductal carcinoma; ILC = invasive lobular carcinoma; IQR = interquartile range; MRI = magnetic resonance imaging; NAC = neoadjuvant chemotherapy; NR = not reported; pCR = pathologic complete response; PR = progesterone receptor.

^aCalculation of values based on total number of patients enrolled, a minority of whom may not have undergone MRI or were excluded from the analysis for other reasons.

^bUsed in six studies, but figures based on five studies where the proportion of patients receiving Trastuzumab is reported.

^cLocalisation biopsy showed the absence of residual tumour (i.e., pathologic measurement of 0.0 cm).

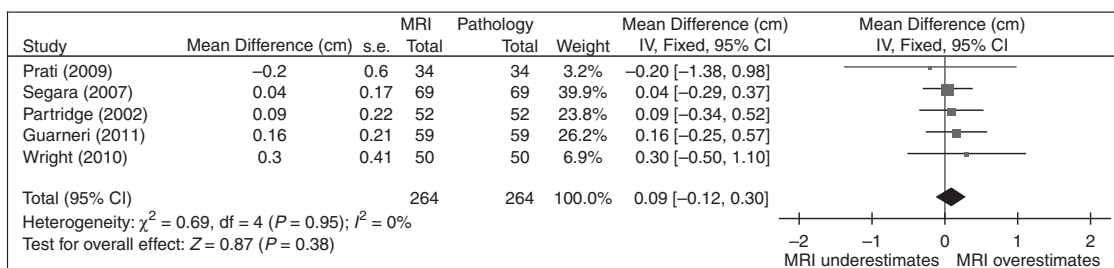


Figure 1. Forest plot of mean difference (cm) between MRI and pathologic size (all studies).

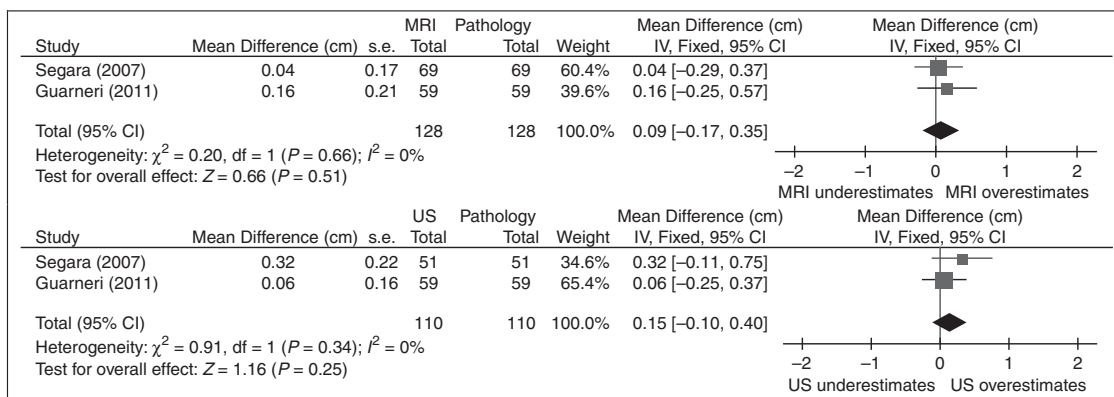


Figure 2. Forest plots of mean difference (cm) between MRI or US and pathologic size (comparative studies).

correlation coefficients was wide (0.21–0.92), with a median value of 0.70 (Supplementary Information Resource 4). Coefficients between 0.20 and 0.39 were reported in two studies, 0.40–0.59 in four studies, 0.60–0.79 in six studies, and 0.80 and above in four studies. One study reported ICC between MRI and pathology (0.48), in addition to Spearman’s rank coefficients (Bollet *et al*, 2007).

Six studies reported correlations with pathology of MRI and mammography (Weatherall *et al*, 2001; Bodini *et al*, 2004; Londero *et al*, 2004; Bollet *et al*, 2007; Prati *et al*, 2009; Wright *et al*, 2010), all of which reported consistently higher correlation coefficients for MRI. However, of the 10 studies that reported correlations with pathology of MRI and clinical examination (Weatherall *et al*, 2001;

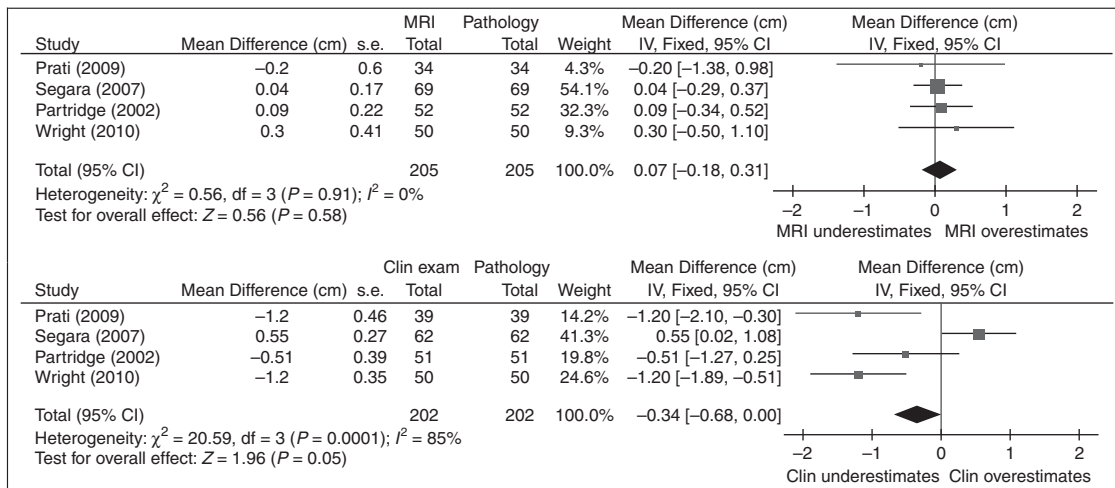


Figure 3. Forest plots of mean difference (cm) between MRI or clinical examination and pathologic size (comparative studies).

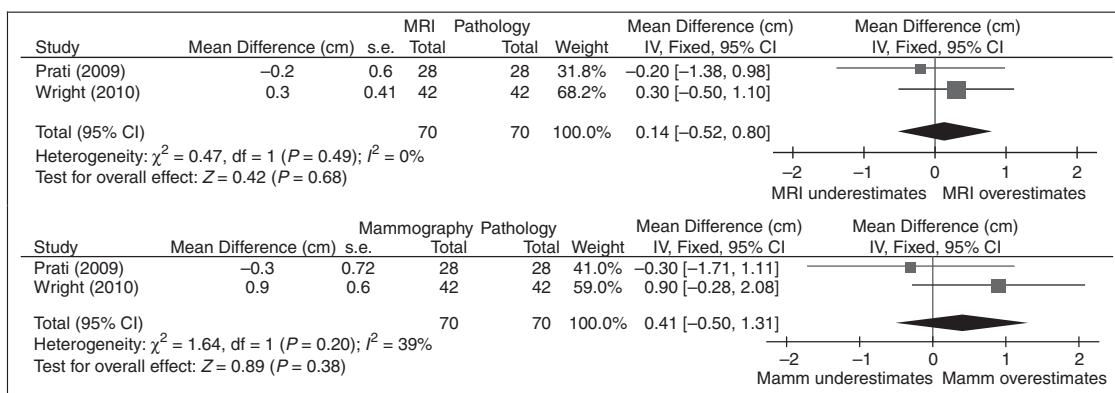


Figure 4. Forest plots of mean difference (cm) between MRI or mammography and pathologic size (comparative studies).

Table 2. Pooled MD and LOA (cm) restricted to studies comparing the respective tests (fixed effects)

	N (studies)	MD (95% CI) (cm)	I ²	LOA (cm)
MRI	2	0.1 (-0.2, 0.3)	0%	-2.9, 3.0
US	2	0.1 (-0.1, 0.4)	0%	-2.6, 2.9
MRI	4	0.1 (-0.2, 0.3)	0%	-4.5, 4.6
Clinical exam	4	-0.3 (-0.7, 0.0)	85%	-5.3, 4.7
MRI	2	0.1 (-0.5, 0.8)	0%	-6.0, 6.3
Mammography	2	0.4 (-0.5, 1.3)	39%	-7.1, 8.0

Abbreviations: CI = confidence interval; LOA = limits of agreement; MD = mean difference; MRI = magnetic resonance imaging; US = ultrasound.

Partridge *et al*, 2002; Rosen *et al*, 2003; Bodini *et al*, 2004; Chen *et al*, 2004; Akazawa *et al*, 2006; Bollet *et al*, 2007; Segara *et al*, 2007; Prati *et al*, 2009; Wright *et al*, 2010), two found correlations favouring the latter test (Prati *et al*, 2009; Wright *et al*, 2010). Similarly, two (Nakahara *et al*, 2010; Guarneri *et al*, 2011) of 11 studies that presented correlations for MRI and US with pathology (Weatherall *et al*, 2001; Bodini *et al*, 2004; Londero *et al*, 2004; Montemurro *et al*, 2005; Akazawa *et al*, 2006; Bollet *et al*, 2007; Segara *et al*, 2007; Bhattacharyya *et al*, 2008; Moon *et al*, 2009; Nakahara *et al*, 2010; Guarneri *et al*, 2011) reported higher correlations for US.

Within-study comparisons of different methods. Six studies (Partridge *et al*, 2002; Akazawa *et al*, 2006; Segara *et al*, 2007; Prati *et al*, 2009; Nakahara *et al*, 2010; Wright *et al*, 2010; Guarneri *et al*, 2011) compared the performance of MRI and other tests by more than one method. In four of those, different methods produced results that could potentially lead to inconsistent conclusions regarding agreement, depending on which measure is considered. In two (Prati *et al*, 2009; Wright *et al*, 2010) of six studies that presented both MDs and correlations, the absolute values of the MD was lower for MRI (≤ 0.3 cm) than for clinical examination (1.2 cm), but a higher correlation was observed between clinical examination and pathologic size. The 95% LOA for MRI were wider than for clinical examination, reflecting the lower correlation for MRI. Similarly, in two of three studies presenting MDs and percentage agreement, the methods suggest opposing conclusions. Guarneri *et al* (2011) found a larger MD and wider LOA for MRI compared with US, but slightly higher percentage agreement, whereas Segara *et al* (2007) reported the reverse (for agreement within 2 cm only). In addition, the slightly higher percentage agreement for MRI than US reported by Guarneri *et al* (2011) contrasts with a lower correlation coefficient, and vice versa for Segara *et al* (2007) (for agreement within 2 cm only).

DISCUSSION

In the neoadjuvant setting, accurate information on the extent of residual malignancy assists in guiding surgical management of

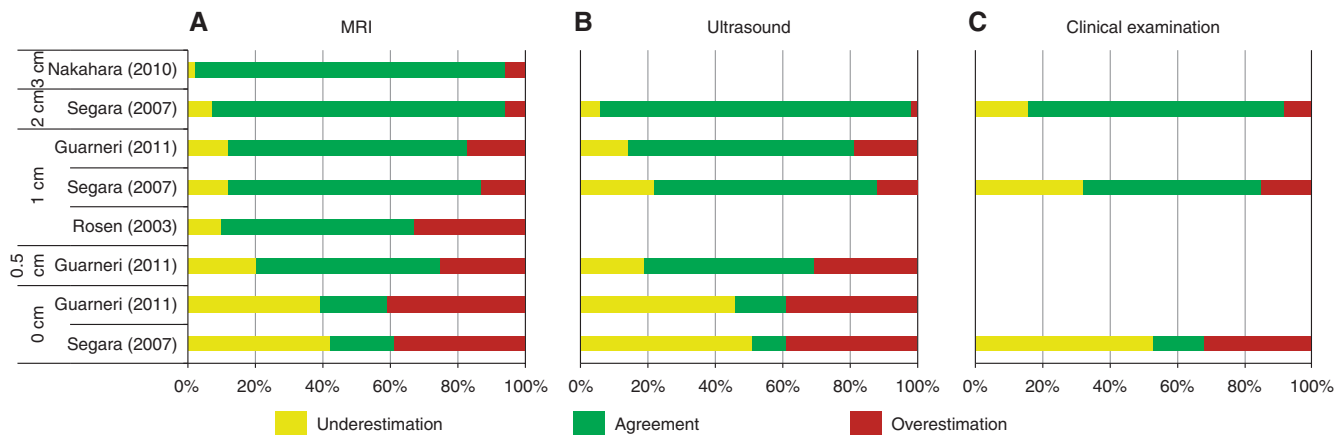


Figure 5. Percentage agreement, underestimation and overestimation for (A) MRI, (B) US and (C) clinical examination by margin of error (cm).

breast cancer. We pooled estimates of the MD between residual tumour size measured by MRI and pathology from six studies, and found that on average, MRI had a tendency to slightly overestimate pathologic size after NAC (MD of 0.1 cm; Figure 1). However, the pooled 95% LOA around this estimate suggest that pathologic tumour measurements may lie between -4.2 cm and $+4.4$ of the MRI measurement, indicating that substantial disagreement may exist. Measurement errors within this range may be of clinical importance in terms of their implications for the choice of treatment approach.

Our analysis of the relative performance of MRI and alternative tests focused on studies directly comparing the tests against pathology (Bossuyt and Leeflang, 2008). Although only two studies reported MDs with pathologic measurements for both MRI and US, pooled estimates suggested that the tests had a similar tendency to overestimate pathologic size, with comparable LOA. The tendency to overestimate pathologic size was greater for mammography than MRI (two studies). Although significant heterogeneity was present in clinical examination findings, three of four studies reported the same direction of effect (underestimation) for this test. Pooled MDs showed clinical examination’s bias towards underestimation to be greater than MRI’s bias for overestimation, and within all four studies the absolute values of MDs were larger for clinical examination. Compared with MRI, wider LOA were observed for both clinical examination and mammography, suggesting that those tests had greater variability in terms of agreement with pathologic measurements. The LOA for all of the alternative tests were large enough to be of potential clinical significance.

Previous summaries of the literature about MRI’s accuracy in measuring residual tumour size have quoted correlations between MRI and pathology, and the percentage of cases in which MRI agrees with, underestimates, or overestimates pathologic measurements. Overall, correlations were considered to be ‘good’ (Lobbes *et al*, 2013), and the statistical significance of those correlations was emphasised (McLaughlin and Hylton, 2011). The methodological limitations of that approach are well documented (Bland and Altman, 1986, 1990). The variable overestimation and underestimation described in those overviews has led others to attach caveats about inaccurate measurement to conclusions about the value of MRI in measuring residual tumour size (Sardanelli *et al*, 2010; McLaughlin and Hylton, 2011). This inconsistency reflects an evidence base which is extensive but disparate in terms of the methods used to assess agreement, and highlights uncertainty about drawing meaningful conclusions from the literature.

Pearson’s and Spearman’s rank correlation coefficients were the most commonly reported statistics in our review (in contrast to MDs and LOA, the more appropriate statistics, yet the least reported). These correlation coefficients, which do not measure

agreement (Bland and Altman, 1986), varied widely and were commonly inconsistent with more appropriate measures reported in the same study. Intraclass correlation, which does assess the degree to which a 1:1 relationship between measurements exists, was presented for MRI and pathology in just one study and was not reported for comparator tests (Bollet *et al*, 2007). The ICC may be an adjunct to the analyses recommended by Bland and Altman (1986), but this statistic alone is also limited in the extent to which it assesses agreement, as it is dependent on the range of observed values and does not separate systematic from random error (Bland and Altman, 1990).

The percentage of MRI measurements which ‘agree’ with pathology within a ‘margin of error’ may provide useful information to supplement MDs and LOA. However, the studies in our review varied considerably in the tolerated discrepancy between measures which was used to define ‘agreement’, reflecting the somewhat arbitrary nature of an ‘acceptable’ error. Furthermore, studies differed in the methods of calculating that discrepancy (i.e., absolute or relative differences), and accompanying percentages of under- or overestimation by MRI were not universally reported. This lack of consistency between studies renders the body of evidence difficult to interpret; future studies can facilitate comparability by reporting agreement, under- and overestimation for multiple margins of error, starting with exact agreement and increasing at 1 cm increments. In contrast to our pooled analysis of MDs showing that MRI has a tendency to slightly overestimate pathologic size, studies describing an absolute margin of error suggested that MRI was equally likely to under- and overestimate the pathologic measurement, highlighting that this method may obscure small measurement biases.

Studies of the agreement between imaging and pathologic size have inherent limitations. Although pathology is considered to be the ‘gold standard’, a variety of potential errors in pathologic measurement have been identified (Lagios, 2005; Provencher *et al*, 2012; Tucker, 2012), meaning that discrepancies with pathology may occur even when residual tumour size is accurately assessed before surgery. For example, pathologic diameters are likely to be overestimated when measured from a combination of tumour fragments, or excised and re-excised specimens (Lagios, 2005). There may also be errors in orientating intact specimens so that tumour diameters on imaging and pathology are measured in the same plane (Provencher *et al*, 2012), particularly if three-dimensional imaging data are unavailable to the pathologist (Weatherall *et al*, 2001; Tucker, 2012); this could result in pathologic measurements underestimating the longest diameter for irregularly shaped tumours (Lagios, 2005). There also exists the possibility that the process of removal, preparation or measurement of the pathologic specimen may shrink, expand or otherwise

distort tumour dimensions (Pritt and Weaver, 2005; Pritt *et al*, 2005; Behjatnia *et al*, 2010; Provencher *et al*, 2012). Furthermore, the inclusion or exclusion of residual DCIS in pathologic measurements has the potential to affect estimates of agreement. Pooled MDs between pathology and MRI or alternative tests (and the associated LOA) must therefore be interpreted with awareness of these issues. However, if errors in the pathologic measurement are random and do not favour MRI over the comparators (or vice versa), these estimates allow for valid comparisons (Glasziou *et al*, 2008). Although this assumption may be reasonable when MRI, comparator tests and pathology are undertaken in the same patients, four (Partridge *et al*, 2002; Segara *et al*, 2007; Prati *et al*, 2009; Wright *et al*, 2010) of six studies reporting MDs excluded patients from one (or more) testing group(s), with discrepancies ranging from a single patient (2%) to up to 26% of patients with MRI data being excluded from analyses of comparator tests (Supplementary Information Resource 4).

Furthermore, differences in test performance may be observed if tumour size is estimated better (or more poorly) in patients selected to (or excluded from) a particular testing group. Authors should be encouraged to present data which allows agreement to be assessed for patients unique to particular analyses vs those common to all testing groups. In addition, these issues also highlight the importance of study authors clearly describing the characteristics of patients excluded from particular analyses. The presentation of important study design characteristics in included studies was generally suboptimal, but in particular, reporting of study withdrawals or exclusions (when they did occur) was poor (Supplementary Information Resource 2).

An important consideration in the interpretation of pooled MD and LOA estimates is that they may be misleading if the difference between tests is systematically related to underlying tumour size, or if the differences are not normally distributed (Bland and Altman, 1986). Plots of the differences by their mean allow for any underlying relationships to be assessed, but were presented in only half of the studies reporting MDs (Partridge *et al*, 2002; Segara *et al*, 2007; Wright *et al*, 2010). Examination of the plots presented in these studies suggests the possibility that the difference in pathology and MRI (or alternative tests) may be greater for larger tumour sizes. Careful attention should be given to graphical presentation of the data before calculating MDs, and data transformation should be considered when systematic relationships exist (Bland and Altman, 1986).

A possible limitation of our analysis is that many studies were not recent, and consequently newer neoadjuvant treatments, including taxanes and trastuzumab, were used in only a minority of patients (Table 1). Agreement between MRI and pathology may vary because of different patterns of tumour regression between taxane-based and non-taxane-based NAC; contrary to previous findings suggesting underestimation when taxanes are used (Denis *et al*, 2004), MDs in studies that used predominantly taxane-based NAC (Wright *et al*, 2010; Guarneri *et al*, 2011) suggest overestimation by MRI relative to studies using non-taxane-based regimens (Segara *et al*, 2007; Prati *et al*, 2009; Supplementary Information Resource 4). Increased rates of pCR owing to modern regimens may also potentially affect MD and LOA estimates, but examination of this issue was not possible owing to the small number of studies reporting those outcomes.

In summary, our meta-analysis is the first to explore and summarise the evidence on agreement between MRI and pathologic tumour measurements after NAC, and to highlight methodological issues which, to date, have precluded conclusions being drawn from the literature. Our work suggests a tendency for MRI to slightly overestimate pathologic tumour size measurements, but LOA are large enough to be of potential clinical importance. Few studies compared MDs between tests and pathology, but the performance of US appeared to be comparable

to that of MRI; poorer agreement was observed for mammography and clinical examination. Although a large number of studies have addressed these questions, most studies have reported Pearson's or Spearman's correlation coefficients. Those measures are inappropriate for assessing agreement, and have contributed to uncertainty about MRI's potential role. Further studies are warranted, and adopt the Bland–Altman approach to assessing MRI's agreement with pathology, and which also assess the agreement with pathology of alternative tests; in addition, we have recommended methods of data presentation to assess the validity of comparisons between tests. Percentages of agreement and associated under/overestimation have limitations, but may provide useful data to supplement Bland–Altman analyses. Similarly, ICCs may also supplement these analyses, but Pearson's and Spearman's correlations should be avoided.

ACKNOWLEDGEMENTS

This work was partly funded by the National Health and Medical Research Council (NHMRC) programme grant 633003 to the Screening and Test Evaluation Program; MLM was supported by an NHMRC post-graduate scholarship.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

- Akazawa K, Tamaki Y, Taguchi T, Tanji Y, Miyoshi Y, Kim SJ, Ueda S, Yanagisawa T, Sato Y, Tamura S, Noguchi S (2006) Preoperative evaluation of residual tumor extent by three-dimensional magnetic resonance imaging in breast cancer patients treated with neoadjuvant chemotherapy. *Breast J* 12(2): 130–137.
- Balu-Maestro C, Chapellier C, Bleuse A, Chanalet I, Chauvel C, Largillier R (2002) Imaging in evaluation of response to neoadjuvant breast cancer treatment benefits of MRI. *Res Treat* 72(2): 145–152.
- Behjatnia B, Sim J, Bassett LW, Moatamed NA, Apple SK (2010) Does size matter? Comparison study between MRI, gross, and microscopic tumor sizes in breast cancer in lumpectomy specimens. *Int J Clin Exp Pathol* 3(3): 303–309.
- Belli P, Costantini M, Malaspina C, Magistrelli A, LaTorre G, Bonomo L (2006) MRI accuracy in residual disease evaluation in breast cancer patients treated with neoadjuvant chemotherapy. *Clin Radiol* 61(11): 946–953.
- Berg WA, Madsen KS, Schilling K, Tartar M, Pisano ED, Larsen LH, Narayanan D, Ozonoff A, Miller JP, Kalinyak JE (2010) Breast cancer: comparative effectiveness of positron emission mammography and MR imaging in presurgical planning for the ipsilateral breast. *Radiology* 258(1): 59–72.
- Bhattacharyya M, Ryan D, Carpenter R, Vinnicombe S, Gallagher CJ (2008) Using MRI to plan breast-conserving surgery following neoadjuvant chemotherapy for early breast cancer. *Br J Cancer* 98(2): 289–293.
- Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 327(8476): 307–310.
- Bland JM, Altman DG (1990) A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Comput Biol Med* 20(5): 337–340.
- Bodini M, Berruti A, Bottini A, Allevi G, Fiorentino C, Brizzi MP, Bersiga A, Generali D, Volpi D, Marini U, Aguggini S, Tampellini M, Alquati P, Olivetti L, Dogliotti L (2004) Magnetic resonance imaging in comparison to clinical palpation in assessing the response of breast cancer to epirubicin primary chemotherapy. *Res Treat* 85(3): 211–218.
- Bollet MA, Thibault F, Bouillon K, Meunier M, Sigal-Zafrani B, Savignoni A, Dieras V, Nos C, Salmon R, Fourquet A (2007) Role of dynamic magnetic resonance imaging in the evaluation of tumor response to preoperative concurrent radiochemotherapy for large breast cancers: a prospective phase II study. *Int J Radiat Oncol Biol Phys* 69(1): 13–18.

- Bossuyt PM, Leeflang MM (2008) Developing Criteria For Including Studies. In *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*. The Cochrane Collaboration. Available at <http://srdta.cochrane.org/sites/srdta.cochrane.org/files/uploads/Chapter06-Including-Studies%20%28September-2008%29.pdf>.
- Chen X, Moore MO, Lehman CD, Mankoff DA, Lawton TJ, Peacock S, Schubert EK, Livingston RB (2004) Combined use of MRI and PET to monitor response and assess residual disease for locally advanced breast cancer treated with neoadjuvant chemotherapy. *Acad Radiol* **11**(10): 1115–1124.
- Cheung YC, Chen SC, Su MY, See LC, Hsueh S, Chang HK, Lin YC, Tsai CS (2003) Monitoring the size and response of locally advanced breast cancers to neoadjuvant chemotherapy (weekly paclitaxel and epirubicin) with serial enhanced MRI. *Res Treat* **78**(1): 51–58.
- Denis F, Desbiez-Bourcier AV, Chapiron C, Arbion F, Body G, Brunereau L (2004) Contrast enhanced magnetic resonance imaging underestimates residual disease following neoadjuvant docetaxel based chemotherapy for breast cancer. *Eur J Surg Oncol* **30**(10): 1069–1076.
- Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, Dancey J, Arbuck S, Gwyther S, Mooney M, Rubinstein L, Shankar L, Dodd L, Kaplan R, Lacombe D, Verweij J (2009) New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer* **45**(2): 228–247.
- Glasziou P, Irwig L, Deeks JJ (2008) When should a new test become the current reference standard? *Ann Intern Med* **149**(11): 816–821.
- Guarneri V, Pecchi A, Piacentini F, Barbieri E, Dieci MV, Ficarra G, Tazzioli G, Frassoldati A, Battista R, Canossi B, Mauri C, D'Amico R, Conte P, Torricelli P (2011) Magnetic resonance imaging and ultrasonography in predicting infiltrating residual disease after preoperative chemotherapy in stage II-III breast cancer. *Ann Surg Oncol* **18**(8): 2150–2157.
- Higgins JPT, Thompson SG, Deeks JJ, Altman DG (2003) Measuring inconsistency in meta-analyses. *Br Med J* **327**(7414): 557–560.
- Irwig L, Bennetts A (1997) Quality of life after breast conservation or mastectomy: A systematic review. *Aust New Zealand J Surg* **67**(11): 750–754.
- Julius T, Kemp SEG, Kneeshaw PJ, Chaturvedi A, Drew PJ, Turnbull LW (2005) MRI and conservative treatment of locally advanced breast cancer. *Eur J Surg Oncol* **31**(10): 1129–1134.
- Lagios MD (2005) Problems in the assessment of tumor size: an elusive grail in current practice. *Semin Breast Dis* **8**(1): 24–30.
- Lobbes M, Prevos R, Smidt M (2012) Response monitoring of breast cancer patients receiving neoadjuvant chemotherapy using breast MRI—a review of current knowledge. *Cancer Ther Res* **1**(34): doi:10.7243/2049-7962-1-34.
- Lobbes MBI, Prevos R, Smidt M, Tjan-Heijnen VCG, Goethem M, Schipper R, Beets-Tan RG, Wildberger JE (2013) The role of magnetic resonance imaging in assessing residual disease and pathologic complete response in breast cancer patients receiving neoadjuvant chemotherapy: a systematic review. *Insights Imaging* **4**: 163–175.
- Londero V, Bazzocchi M, Del Frate C, Puglisi F, Di Loreto C, Francescutti G, Zuiani C (2004) Locally advanced breast cancer: comparison of mammography, sonography and MR imaging in evaluation of residual disease in women receiving neoadjuvant chemotherapy. *Eur Radiol* **14**(8): 1371–1379.
- Marinovich ML, Houssami N, Macaskill P, Sardanelli F, Irwig L, Mamounas EP, Von Minckwitz G, Brennan ME, Ciatto S (2013) Meta-analysis of magnetic resonance imaging in detecting residual breast cancer after neoadjuvant therapy. *J Natl Cancer Inst* **105**(5): 321–333.
- Marinovich ML, Sardanelli F, Ciatto S, Mamounas E, Brennan M, Macaskill P, Irwig L, Von Minckwitz G, Houssami N (2012) Early prediction of pathologic response to neoadjuvant therapy in breast cancer: systematic review of the accuracy of MRI. *Breast* **21**: 669–677.
- Mclaughlin R, Hylton N (2011) MRI in breast cancer therapy monitoring. *NMR Biomed* **24**(6): 712–720.
- Montemurro F, Martincich L, De Rosa G, Cirillo S, Marra V, Biglia N, Gatti M, Sismondi P, Aglietta M, Regge D (2005) Dynamic contrast-enhanced MRI and sonography in patients receiving primary chemotherapy for breast cancer. *Eur Radiol* **15**(6): 1224–1233.
- Moon HG, Han W, Lee JW, Ko E, Kim EK, Yu JH, Kang SY, Moon WK, Cho N, Park IA, Oh DY, Han SW, Im SA, Noh DY (2009) Age and HER2 expression status affect MRI accuracy in predicting residual tumor extent after neo-adjuvant systemic treatment. *Ann Oncol* **20**(4): 636–641.
- Nakahara H, Yasuda Y, Machida E, Maeda Y, Furusawa H, Komaki K, Funagayama M, Nakahara M, Tamura S, Akiyama F (2010) MR and US imaging for breast cancer patients who underwent conservation surgery after neoadjuvant chemotherapy: comparison of triple negative breast cancer and other intrinsic subtypes. *Breast Cancer* **18**(3): 152–160.
- Padhani AR, Husband JE (2000) Are current tumour response criteria relevant for the 21st century? *Br J Radiol* **73**(874): 1031–1033.
- Partridge SC, Gibbs JE, Lu Y, Esserman LJ, Sudilovsky D, Hylton NM (2002) Accuracy of MR imaging for revealing residual breast cancer in patients who have undergone neoadjuvant chemotherapy. *Am J Roentgenol* **179**(5): 1193–1199.
- Prati R, Minami CA, Gornbein JA, Debruhl N, Chung D, Chang HR (2009) Accuracy of clinical evaluation of locally advanced breast cancer in patients receiving neoadjuvant chemotherapy. *Cancer* **115**(6): 1194–1202.
- Pritt B, Tessitore JJ, Weaver DL, Blaszyk H (2005) The effect of tissue fixation and processing on breast cancer size. *Hum Pathol* **36**(7): 756–760.
- Pritt B, Weaver DL (2005) Accurate determination of breast cancer size: The role of histopathology and imaging. *Curr Diagn Pathol* **11**(6): 435–442.
- Provencher L, Diorio C, Hogue JC, Doyle C, Jacob S (2012) Does breast cancer tumor size really matter that much? *Breast* **21**(5): 682–685.
- Rosen EL, Blackwell KL, Baker JA, Soo MS, Bentley RC, Yu D, Samulski TV, Dewhirst MW (2003) Accuracy of MRI in the detection of residual breast cancer after neoadjuvant chemotherapy. *Am J Roentgenol* **181**(5): 1275–1282.
- Sardanelli F, Boetes C, Borisch B, Decker T, Federico M, Gilbert FJ, Helbich T, Heywang-Kobrunner SH, Kaiser WA, Kerin MJ, Mansel RE, Marotti L, Martincich L, Mauriac L, Meijers-Heijboer H, Orecchia R, Panizza P, Ponti A, Purushotham AD, Regitnig P, Del Turco MR, Thibault F, Wilson R (2010) Magnetic resonance imaging of the breast: Recommendations from the EUSOMA working group. *Eur J Cancer* **46**(8): 1296–1316.
- Segara D, Krop IE, Garber JE, Winer E, Harris L, Bellon JR, Birdwell R, Lester S, Lipsitz S, Iglehart JD, Golshan M (2007) Does MRI predict pathologic tumor response in women with breast cancer undergoing preoperative chemotherapy? *J Surg Oncol* **96**(6): 474–480.
- Tucker FL (2012) Imaging-assisted large-format breast pathology: program rationale and development in a nonprofit health system in the United States. *Int J Breast Cancer* **1**(1): <http://www.hindawi.com/journals/ijbc/2012/171792/cta/>.
- Weatherall PT, Evans GF, Metzger GJ, Saborrian MH, Leitch AM (2001) MRI vs. histologic measurement of breast cancer following chemotherapy: comparison with x-ray mammography and palpation. *J Magn Reson Imaging* **13**(6): 868–875.
- Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, 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**: 1–13.
- Whiting PF, Weswood ME, Rutjes AWS, Reitsma JB, Bossuyt PNM, Kleijnen J (2006) Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. *BMC Med Res Methodol* **6**: 9.
- Woodward M (1999) Basic analytical procedures. In *Epidemiology: Study Design and Data Analysis*. pp 31–105. Chapman & Hall/CRC: Boca Raton.
- Wright FC, Zubovits J, Gardner S, Fitzgerald B, Clemons M, Quan ML, Causer P (2010) Optimal assessment of residual disease after neo-adjuvant therapy for locally advanced and inflammatory breast cancer - clinical examination, mammography, or magnetic resonance imaging? *J Surg Oncol* **101**(7): 604–610.
- Yeh E, Slanetz P, Kopans DB, Rafferty E, Georgian-Smith D, Moy L, Halpern E, Moore R, Kuter I, Taghian A (2005) Prospective comparison of mammography, sonography, and MRI in patients undergoing neoadjuvant chemotherapy for palpable breast cancer. *Am J Roentgenol* **184**(3): 868–877.

This work is published under the standard license to publish agreement. After 12 months the work will become freely available and the license terms will switch to a Creative Commons Attribution-NonCommercial-Share Alike 3.0 Unported License.

Supplementary Information accompanies this paper on British Journal of Cancer website (<http://www.nature.com/bjc>)