Central pathology review with two-stage quality assurance for pathological response after neoadjuvant chemotherapy in the ARTemis Trial

Jeremy St John Thomas¹, Elena Provenzano^{2,3,4}, Louise Hiller⁵, Janet Dunn⁵, Clare Blenkinsop⁵, Louise Grybowicz⁶, Anne-Laure Vallier⁶, Ioannis Gounaris⁷, Jean Abraham^{2,3,8}, Luke Hughes-Davies^{8,9}, Karen McAdam^{9,10}, Stephen Chan¹¹, Rizvana Ahmad¹², Tamas Hickish^{13,14}, Stephen Houston¹⁵, Daniel Rea¹⁶, Carlos Caldas^{2,3,8,17}, John MS Bartlett¹⁸, David Allan Cameron¹⁹, Richard Laurence Hayward²⁰ and Helena Margaret Earl^{2,3,8} for the ARTemis Investigators Group

¹Department of Pathology, Western General Hospital, Edinburgh, UK; ²NIHR Cambridge Biomedical Research Centre, Addenbrookes Hospital, Cambridge, UK; ³Cambridge Breast Cancer Research Unit, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK; ⁴Department of Histopathology, Addenbrookes Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK; ⁵Warwick Clinical Trials Unit, University of Warwick, Coventry, UK; ⁶Cambridge Clinical Trials Unit – Cancer Theme, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK; ⁷The Queen Elizabeth Hospital King's Lynn NHS Foundation Trust, King's Lynn, UK; ⁸University of Cambridge, Department of Oncology, Addenbrooke's Hospital, Cambridge, UK; ⁹Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK; ¹⁰Peterborough City Hospital, Edith Cavell Campus, Peterborough, UK; ¹¹Nottingham University Hospital (City Campus), Nottingham, UK; ¹²West Middlesex University Hospital, Isleworth, UK; ¹³Poole Hospital NHS Foundation Trust, Poole, UK; ¹⁴Bournemouth University, Poole, UK; ¹⁵The Royal Surrey County Hospital NHS Foundation Trust, Guildford, UK; ¹⁶Institute for Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK; ¹⁷Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Cambridge, UK; ¹⁸Ontario Institute for Cancer Research, MaRS Centre, Toronto, ON, Canada; ¹⁹University of Edinburgh Cancer Research Centre, Western General Hospital, Edinburgh, UK and ²⁰Department of Medical Oncology, NHS Lothian, Western General Hospital, Edinburgh, UK

The ARTemis Trial tested standard neoadjuvant chemotherapy ± bevacizumab in the treatment of HER2-negative early breast cancer. We compare data from central pathology review with report review and also the reporting behavior of the two central pathologists. Eight hundred women with HER2-negative early invasive breast cancer were recruited. Response to chemotherapy was assessed from local pathology reports for pathological complete response in breast and axillary lymph nodes. Sections from the original core biopsy and surgical excision were centrally reviewed by one of two trial pathologists blinded to the local pathology reports. Pathologists recorded response to chemotherapy descriptively and also calculated residual cancer burden. 10% of cases were double-reported to compare the central pathologists' reporting behavior. Full sample retrieval was obtained for 681 of the 781 patients (87%) who underwent surgery within the trial and were evaluable for pathological complete response. Four hundred and eighty-three (71%) were assessed by JSJT, and 198 (29%) were assessed by EP. Residual cancer burden calculations were possible in 587/681 (86%) of the centrally reviewed patients, as 94/681 (14%) had positive sentinel nodes removed before neoadjuvant chemotherapy invalidating residual cancer burden scoring. Good concordance was found between the two pathologists for residual cancer burden classes within the 65-patient quality assurance exercise (kappa 0.63 (95% CI: 0.57–0.69)). Similar results were obtained for the between-treatment arm comparison both from the report review and the central pathology review. For pathological complete response, report review was as good as central pathology review but for minimal residual

Correspondence: Dr J St J Thomas, MB, BS, Department of Pathology, NHS Lothian, Western General Hospital, Edinburgh EH4 2XU, UK. E-mail: Jeremy.thomas@nhslothian.scot.nhs.uk or jstjthomas@gmail.com

Received 2 December 2016; revised 15 February 2017; accepted 19 February 2017; published online 26 May 2017

disease, report review overestimated the extent of residual disease. In the ARTEMIS Trial central pathology review added little in the determination of pathological complete response but had a role in evaluating low levels of residual disease. Calculation of residual cancer burden was a simple and reproducible method of quantifying response to neoadjuvant chemotherapy as demonstrated by performance comparison of the two pathologists. *Modern Pathology* (2017) **30**, 1069–1077; doi:10.1038/modpathol.2017.30; published online 26 May 2017

The ARTEMIS trial is an open-label, randomized, phase 3 trial assessing the efficacy of neoadjuvant bevacizumab added to docetaxel followed by fluorouracil, epirubicin, and cyclophosphamide, for women with HER2-negative early breast cancer. Its primary end point was pathological complete response, defined as the absence of invasive disease in the breast and axillary lymph nodes. Initially, the two randomized arms of the trial were compared in terms of rates of pathological complete response as determined by a two-reader blinded review of local pathology reports by the Chief Investigators.¹ In addition, a central pathology review and a large-scale two-stage pathology quality assurance exercise was undertaken. Thereby, the accuracy of this commonly used primary end point in neoadjuvant chemotherapy breast cancer trials was assessed and compared with the two-reader report review, which until now has been the standard used by this group.² In addition, the reliability of central specimen review has been investigated by independent double-reading of residual cancer burden categories carried out by the two central pathologists in a subset of cases. This allows us to report on the comparison between assessment of local pathology reporting and central pathology review of original diagnostic material and also the reporting behavior of the two reviewing pathologists. Although central pathological review has been carried out in studies reporting major center results,³ as far as the authors are aware this is the first report of central pathology review of pathological complete response with residual cancer burden scoring and class definition, carried out as part of a multicenter large randomized phase 3 trial.

Materials and methods

Between May 2009 and January 2013, the ARTEMIS trial recruited 800 women ≥ 18 years old with newly diagnosed HER2-negative early invasive breast cancer (radiological tumor size > 20 mm, with or without axillary involvement). Patients with inflammatory cancer, T4 tumors with direct extension to the chest wall or skin, and ipsilateral supraclavicular lymph node involvement were also eligible with any size of primary tumor. Full eligibility criteria details have been described in detail elsewhere.¹ Patients were randomized from 66 UK sites and assigned, via a central computerized minimization procedure, to three cycles of docetaxel (100 mg/m^2 once every 21 days) followed by three

cycles of fluorouracil (500 mg/m^2), epirubicin (100 mg/m^2), and cyclophosphamide (500 mg/m^2) once every 21 days (docetaxel-fluorouracil/epirubicin/cyclophosphamide), without or with four cycles of bevacizumab (15 mg/kg) (bevacizumab+docetaxel-fluorouracil/epirubicin/cyclophosphamide). A total 781 patients (98% of the randomized 800) underwent surgery following their neoadjuvant treatment and could be assessed, via local pathology reports, for the primary end point of absence of invasive breast cancer in the breast and axillary lymph nodes.

Methods

Diagnostic and surgical excision histopathology slides were requested from the relevant participating sites for all the 781 evaluable patients. All retrieved cases underwent central independent review, blinded to the local histopathology report, including any block descriptions, by an experienced breast histopathologist with a special interest in neoadjuvant clinical trials (JSJT and EP) between June 2011 and March 2016. The reviewing pathologist was not the same as the pathologist who had previously assessed the slides locally, and/or would have access to the histopathology results at their hospital. Any missing slides or additional relevant operations (eg, sentinel lymph node biopsy) were re-requested as necessary. The variables recorded were maximum invasive tumor size in two dimensions, whole tumor size (including ductal carcinoma *in situ*) in two dimensions, posttreatment tumor grade, presence of lymphovascular invasion, presence and nature of *in situ* disease, percentage tumor cellularity, percentage cellularity that is *in situ* disease, total number of lymph nodes, number of positive lymph nodes, and size of largest nodal metastasis.

In addition to assessing the validity of the findings from the blinded review of local pathology reports, an inter-pathologist reproducibility exercise was also undertaken. For this, a randomly chosen 10% of patients had samples reviewed by both pathologists for determination of levels of agreement between central review findings. To simplify this exercise, variables were restricted to those required to calculate the residual cancer burden score: invasive size (length and width), percent tumor cellularity, percent of tumor that is DCIS, size of largest nodal metastasis, and number of positive nodes.⁴ The 10% sample was randomly chosen, while ensuring a representative residual cancer burden class split, as

recorded by the first pathologist's reviews. This approach was written into the pre-planned agreement decided by the ARTEMIS Trial Management Group. Ten percent of ARTEMIS cases for co-review was the opportunistic sample that was deemed a manageable workload for the two ARTEMIS pathologists.

The results of the central pathology review were also compared with the outcome results as determined by the central review of the original histopathology reports from the source laboratory. In particular, we compared rates of pathological complete response and minimal residual disease as determined by formal assessment of the residual cancer burden and by interpretation of the histopathology. Original pathology reports were reviewed by the two Chief Investigators (LH and HME) according to residual cancer burden class definitions

laid out in the Trial Protocol. The primary end point of the trial was pathological complete response and that was the main focus of report review by the Chief Investigators. Detailed information in the pathology report was not always available to enable estimates of degrees of partial response by the Chief Investigators who were given guidance by the Trial Pathologists as follows: pathological complete response—no residual invasive carcinoma in the breast or lymph nodes; residual cancer burden 1/minimal residual disease—residual tumor < 5 mm; residual cancer burden 2—< 50% tumor cellularity; residual cancer burden 3—no appreciable response. The local pathologists were not given any formal reporting guidelines specifically for this trial.

Statistical Methods

Agreement between the two pathologists' residual cancer burden classes, and also between central review and local reports in determination of pathological complete response, was undertaken using the kappa statistic. Agreement between the two pathologists in terms of residual cancer burden scores and its six components were scrutinized using Bland–Altman plots and assessed using overall concordance correlation coefficient.⁵ Comparison of patient characteristics between groups was undertaken using chi-squared tests with continuity correction where appropriate. Logistic regression was used to assess the effect of randomized treatment arm on pathological complete response rates, after adjustment for stratification factors.

Results

A total of 22 916 slides from 727 patients were reviewed. Full sample retrieval was obtained for 681 (87%) of the 781 ARTEMIS patients who underwent surgery within the trial and were evaluable for the primary end point of pathological complete response. A total 483/681 patients (71%) were assessed by JSJT and 198/681 patients (29%) by EP. The maximum number of slides per patient was 164; median 29 slides. A total 94/681 patients (14%) had a positive pre-chemotherapy sentinel lymph node biopsy (SLNB) thus invalidating the calculation of a residual cancer burden score at surgery. Residual cancer burden scores and classes were thus calculated on the remaining 587 patients (75% of the 781). Patient characteristics of the 587 patients with assessable residual cancer burden appeared representative of the trial sample as a whole (Table 1).

Inter-Rater Reproducibility of Pathologists

Sixty-five patients were double reviewed by JSJT and EP. The 65 patients were representative of the 587 sample as a whole in terms of patient characteristics

Table 1 Patient characteristics

	Full trial population (n = 800) N (%)	Central pathology review sample (n = 587) N (%)	Inter-rater reliability sample (n = 65) N (%)
<i>Randomized treatment:</i>			
Bevacizumab +D-FEC	399 (50%)	290 (49%)	29 (45%)
D-FEC	401 (50%)	297 (51%)	36 (55%)
<i>Age</i>			
≤ 50 years old	543 (68%)	393 (67%)	44 (68%)
> 50 years old	257 (32%)	194 (33%)	21 (32%)
<i>Estrogen receptor status</i>			
Negative (Allred score 0–2)	248 (31%)	194 (33%)	18 (28%)
Weakly positive (Allred score 3–5)	75 (9%)	60 (10%)	10 (15%)
Strongly positive (Allred score 6–8)	477 (60%)	333 (57%)	37 (57%)
<i>Tumor size</i>			
≤ 50 mm	635 (79%)	472 (80%)	54 (83%)
> 50 mm	165 (21%)	115 (20%)	11 (17%)
<i>Clinical involvement of axillary nodes</i>			
No	383 (48%)	288 (49%)	38 (58%)
Yes	417 (52%)	299 (51%)	27 (42%)
<i>Inflammatory or locally advanced disease or both</i>			
No	651 (81%)	484 (82%)	51 (78%)
Yes	149 (19%)	103 (18%)	14 (22%)
<i>Residual cancer burden class^a</i>			
0	—	121 (21%)	14 (21%)
1	—	90 (15%)	9 (14%)
2	—	290 (49%)	33 (51%)
3	—	86 (15%)	9 (14%)

Abbreviation: D-FEC, Docetaxel-Fluorouracil/Epirubicin/Cyclophosphamide.

^aResidual cancer burden class is only known for those patients included in the central pathology review sample.

Table 2 Residual cancer burden classes for the 65 patients, by the two pathologists

Pathologist 2	Pathologist 1				Total
	Residual cancer burden class 0	Residual cancer burden class 1	Residual cancer burden class 2	Residual cancer burden class 3	
Residual cancer burden class 0	13	—	—	—	13
Residual cancer burden class 1	1	5	3	—	9
Residual cancer burden class 2	—	4	27	3	34
Residual cancer burden class 3	—	—	2	7	9
Total	14	9	32	10	65

(Table 1) and the random sampling technique determined that they were also representative in terms of residual cancer burden class as recorded by the first pathologist's central review.

Residual cancer burden class. The two pathologists showed very similar reporting profiles for residual cancer burden class (observed frequencies of residual cancer burden 0:1:2:3 being 14:9:32:10 for pathologist 1 and 13:9:34:9 for pathologist 2; Table 2). In 52/65 (80%) of patients, there was agreement on residual cancer burden class, and in 13/65 (20%) where there was disagreement none were more than by one residual cancer burden class. A good level of agreement was observed over all residual cancer burden classes (kappa 0.70 (95% CI: 0.55–0.84); Figure 1). No differences were found between patient groups where JSJT and EP agreed on residual cancer burden class ($n=52$) or disagreed ($n=13$) in terms of randomized treatment arm or stratification variables (age, ER status, tumor size, clinical involvement of axillary nodes, locally advanced/inflammatory disease data not shown).

Residual cancer burden score. For the 13 patients where there was disagreement in residual cancer burden class, the majority of disagreements were due to the two pathologists' residual cancer burden scores falling just either side of the published residual cancer burden score cut-points of 1.36 and 3.28 (Figure 2). There was good overall concordance in residual cancer burden score (concordance correlation coefficient 0.75 (95% CI: 0.40–0.91)), with the average discrepancy in residual cancer burden score being of the magnitude 0.245 (IQR: 0.135–0.501, range: 0.085–1.840).

Components of the residual cancer burden score. Focusing on the 13 patients where the two pathologists differed in residual cancer burden class assignment, the greatest inter-rater variability was in the assessment of percentage of ductal carcinoma *in situ* within the tumor (concordance correlation coefficient –0.04 (95% CI: –0.30 to 0.21)) and, to a lesser extent, in the assessment of invasive size (concordance correlation coefficient 0.20 (95% CI: –0.12 to 0.47) for width and concordance correlation coefficient 0.35 (95% CI: –0.11 to 0.68) for length)

Figure 1 Level of agreement across two pathologists' rating of residual cancer burden class. One rectangle is depicted for each level of residual cancer burden class, their height and width based on the row and column cumulative totals. Thus for the residual cancer burden 0 rectangle of pathologist 1 vs pathologist 2 comparison, all patient's categorized as residual cancer burden 0 by either pathologist are included. The boundaries of the rectangles along both axes represent the number of patients that were categorized as residual cancer burden 0 for each pathologist. Dark squares within the rectangles represent exact agreement between residual cancer burden classes from the two pathologists (ie, both rating as residual cancer burden 0), and are of size based on the cell frequencies and located according to the cumulative totals of the previous levels. Light rectangles represent partial agreement, where the residual cancer burden class from one pathologist is 1 different of the residual cancer burden class from the other pathologist (ie, residual cancer burden 0 by one pathologist but residual cancer burden 1 by the other pathologist). White areas within the rectangle reflect disagreement by more than one level (ie, residual cancer burden 0 by one pathologist and residual cancer burden 2 or 3 by the other pathologist).

and percent of tumor cellularity (concordance correlation coefficient 0.30 (95% CI: –0.05 to 0.59)). The strongest agreement was observed in identification of number of positive nodes (concordance correlation coefficient 0.95 (95% CI: 0.85–0.98)) followed by size of the largest nodal metastasis (concordance correlation coefficient 0.74 (95% CI: 0.37–0.91)).

Sources of discrepancy. Seven cases where there was a disagreement in residual cancer burden class due to substantial differences in size measurement, cellularity or nodal status were reviewed again with

Figure 2 (a and b) Inter-rater reliability of pathologists' residual cancer burden scores, where there is disagreement in residual cancer burden class ($n = 13$ patients). (a) Pathologist 1 vs Pathologist 2 residual cancer burden scores. (b) Average of the two pathologists' residual cancer burden scores.

joint discussion by the two pathologists. Sources of discrepancy included interpretation of multiple tumor foci as one lesion or multiple lesions, measurement of lesion size from single slides or estimating total number of slides, inclusion of pre-treatment sentinel lymph node metastases in the residual cancer burden calculation, errors in measurement, and interpretation of degenerate cells in posttreatment lymph nodes as metastasis or not. The different weightings of the elements of the residual cancer burden equation reduce the effect of the variance among those component scores.

Central Review of Pathology Specimens vs Review of Local Pathology Reports: Inter-Method Reliability

Both methods determined similar levels of pathological complete response in the 587 patients where both assessment results were available; 121 (21%) with residual cancer burden class 0 from central pathology review and 119 (20%) reported as pathological complete response from local pathology report

(Table 3). A good level of agreement was observed between the two methods' findings when grouped as the three levels of residual cancer burden 0 (pathological complete response) vs residual cancer burden 1 (minimal residual disease) vs residual cancer burden 2/3 (moderate/extensive disease; kappa 0.63 (95% CI: 0.57–0.69); Figure 3). However, for six patients, the level of disagreement was by more than one class (one patient with pathological complete response from the report review but residual cancer burden class 2 from specimen review, and five patients with moderate/extensive disease from the report review but with residual cancer burden class 0 from the specimen review.)

Slides for five of the six cases were available for second review by one of the pathologists (EP). Sources of discrepancy included not receiving all the tumor slides for review (two cases) and interpretation of residual tumor as ductal carcinoma *in situ* or invasive disease (one case). In one case, the second review agreed with the histopathology report (residual tumor) rather than the central review

Table 3 Levels of residual cancer at surgery, from the two assessment methods for the 781 patients

Local pathology report review	Central specimen review					Total
	Residual cancer burden class 0	Residual cancer burden class 1	Residual cancer burden class 2	Residual cancer burden class 3	Not available	
Pathological complete response ^a	109	9	1	0	34	153
Minimal residual disease	7	25	29	0	16	77
Moderate/extensive residual disease	5	56	260	86	144	551
Total	121	90	290	86	194	781

^aPathological complete response in all breast tumors and absence of disease in all removed axillary lymph nodes (ypT0/Tis ypN0). Entries in bold indicate agreement.

Figure 3 Level of agreement across the two methods of review: One rectangle is depicted for each level of pathologic response, their height and width based on the row and column cumulative totals. Thus for the residual cancer burden 0/pathCR rectangle of method comparison, all patients categorized as residual cancer burden 0 by pathologist or pathological complete response by report review are included. The boundaries of the rectangles along both the axes represent the number of patients that were categorized as residual cancer burden 0/pathological complete response. Dark squares within the rectangles represent exact agreement between methods (ie, pathologist rating as residual cancer burden 0 and report review as pathological complete response), and are of size based on the cell frequencies and located according to the cumulative totals of the previous levels. Light rectangles represent partial agreement, where the conclusion from the pathologist is one group different to that from the report review (ie, residual cancer burden 0 by pathologist but minimal residual disease by report review). White areas within the rectangle reflect disagreement by more than one level (ie, residual cancer burden 0 by pathologist and moderate/extensive residual disease by report review).

(pathological complete response). In another case called pathological complete response on central review, the discrepancy appears to be due to inconsistency in the original report in calling tumor cells in the node viable and nonviable; both central reviewers thought this represented an area of necrosis.

ARTEMIS Primary End Point Results

The ARTEMIS trial's primary end point was previously reported using the local pathology report reviews on

781 patients and showed significantly more bevacizumab+docetaxel-fluorouracil/epirubicin/cyclophosphamide patients achieving a pathological complete response compared with docetaxel-fluorouracil/epirubicin/cyclophosphamide patients: 22% (95% CI: 18–27) of 388 bevacizumab+docetaxel-fluorouracil/epirubicin/cyclophosphamide patients compared with 17% (95% CI: 13–21) of 393 docetaxel-fluorouracil/epirubicin/cyclophosphamide patients (adjusted $P=0.03$; Table 4A).¹ Using the residual cancer burden classes from the central pathology specimen review, the results remained the same: 25% (95% CI: 20–30) of 290 bevacizumab+docetaxel-fluorouracil/epirubicin/cyclophosphamide patients achieved a residual cancer burden 0 compared with 16% (95% CI: 12–21) of 297 docetaxel-fluorouracil/epirubicin/cyclophosphamide patients (adjusted $P=0.02$; Table 4B).

Likewise previously, using local pathology report reviews, pathological complete response rates had been found to differ significantly across both ER status (ER negative 38% [95% CI: 32–45], weakly positive 41% [29–53], strongly positive 7% [5–9]; $P<0.0001$), and tumor grade (grade 1/2 7% [4–11], grade 3 29% [25–34]; $P<0.0001$).¹ Using the central pathology specimen review, similar results were found for rates of residual cancer burden 0; ER negative 39% [95% CI: 32–46], weakly positive 35% [23–48], strongly positive 7% [5–11] ($P<0.0001$) and grade 1/2 7% [4–12], grade 3 31% [26–37] ($P<0.0001$).

Discussion

This review focused on the presence or absence of pathological complete response in the excision specimen including the presence of residual ductal carcinoma *in situ*. Local pathologists were not given reporting proformas or guidelines for the assessment of response, which have been shown to aid concordance between pathologists in clinical trials.⁶ Because the reviewing pathologists were assessing the original sections in the overwhelming majority of cases, analytical issues do not impinge on this central review although differences in practice among different local laboratories would necessitate caution in drawing any comparison between centers.

Table 4A Treatment arm comparison using local pathology report review data (*n* = 781 patients)

<i>Local pathology report review</i>	<i>D → FEC^a</i> % (95% CI) (<i>n</i> = 66/393)	<i>Bevacizumab + D → FEC</i> % (95% CI) (<i>n</i> = 87/388)	<i>p^b</i>
Pathological complete response in all breast tumors and absence of disease in all removed axillary lymph nodes (ypT0/Tis ypN0)	17% (13–21%)	22% (18–27%)	0.03
ER negative (Allred 0–2) (<i>n</i> = 241)	31% (23–40)	45% (36–55)	
ER weak positive (Allred 3–5) (<i>n</i> = 74)	30% (16–47)	51% (34–68)	
ER strong positive (Allred 6–8) (<i>n</i> = 466)	7% (4–11)	6% (3–10)	
Grade 1/2 (<i>n</i> = 293) ^c	8% (5–14)	6% (3–11)	
Grade 3 (<i>n</i> = 403)	23% (17–29)	36% (29–43)	

^aD-FEC, Docetaxel-Fluorouracil/Epirubicin/Cyclophosphamide.^bAdjusted for the five stratification variables.^cTumor grade of each patient's largest breast tumor at baseline.**Table 4B** Treatment arm comparison using central pathology specimen review data (*n* = 587 patients)

<i>Central pathology review</i>	<i>D → FEC^a</i> % (95% CI) (<i>n</i> = 49/297)	<i>Bevacizumab + D → FEC</i> % (95% CI) (<i>n</i> = 72/290)	<i>p^b</i>
Residual cancer burden 0	16% (12–21%)	25% (20–30%)	0.02
ER negative (Allred 0–2) (<i>n</i> = 194)	33% (24–44%)	45% (35–55%)	
ER weak positive (Allred 3–5) (<i>n</i> = 60)	22% (9–42%)	45% (28–64%)	
ER strong positive (Allred 6–8) (<i>n</i> = 333)	6% (3–11%)	8% (4–14%)	
Grade 1/2 (<i>n</i> = 216) ^c	9% (4–16%)	6% (2–12%)	
Grade 3 (<i>n</i> = 313)	23% (17–31%)	39% (32–47%)	

^aD-FEC, Docetaxel-Fluorouracil/Epirubicin/Cyclophosphamide.^bAdjusted for the five stratification variables.^cTumor grade of each patient's largest breast tumor at baseline.

In this review, the pathologists were blinded to the macroscopic description and therefore had to reconstruct the tumor bed dimensions from the slides as best as they could. With hindsight given the differences in reporting practice in this area, access to reports would have been of benefit in some cases but not in others. Normally, a pathologist records a block map to aid reconstruction of the tumor area when viewing the slides, and this was highlighted as being of particular importance for accurate assessment of response in the recent BIG-NABCG working group recommendations (<https://www.mdanderson.org/education-and-research/resources-for-professionals/clinical-tools-and-resources/clinical-calculators/calculators-rcb-pathology-protocol2.pdf> (Accessed 19 August 2016)). In some cases, the tumor bed was present on megaslices and this made assessment much easier. The assessment of tumor bed size is often not straightforward following neoadjuvant chemotherapy because the tumor is poorly defined macroscopically, and it can be difficult to determine the tumor boundaries histologically. Tumor cellularity can be very heterogeneous, and is also difficult to assess in spite of the availability of online guidance tools (<https://www.mdanderson.org/education-and-research/resources-for-professionals/clinical-tools-and-resources/clinical-calculators/calculators-rcb-pathology-protocol2.pdf> (Accessed 19 August 2016))⁷ and there is

inconsistency among pathologists in these assessments.⁸ Agreement about pathological complete response should, however, be good and will only usually cause difficulty if small residues of tumor cells are overlooked, or if there is difficulty in interpreting *in situ* from invasive disease. Although in this study the best level of agreement in the reviewing pathologists' cross-over study was of numbers of lymph nodes, this also is not always easy to determine without the macroscopic description. In some cases, the local pathologist had written on the slide to state the number of nodes present. Our data are similar to those recorded in a recent audit where a residual cancer burden score concordance of 0.25 points was reported for most cases with a kappa value of 0.714 between residual cancer burden classes. In this study, the reviewing pathologist had access to a block map/description.⁹ The concordance between the two pathologists is better than recorded in a recent review of consistency of reporting of residual cancer burden and replicates the finding that the reporting of the lymph node component of the score is more reproducible.⁸ Given the limitations of this study detailed above with lack of access to source reports and block descriptions, the residual cancer burden is shown to be a very robust system for quantifying residual disease in the clinical trial context.

The central pathology review was immensely labor-intensive. The maximum number of slides submitted for a single case was 164 (median 29 per case). At best, it was only possible to review three or four cases per hour. Not only was it time consuming for the pathologists, but it also placed a burden on local pathology departments retrieving slides and a considerable logistic burden for the Trials Office. One must question whether the exercise was worth the effort given that there was no clinically significant numerical change in the end results. However, one cannot generalize about central pathology review. In some trials central review is used at the outset to confirm eligibility whether this be Her-2 or ER status for example or a particular tumor type, eg, triple negative breast cancer. In the ALTTO Trial, both Her-2 and ER status were changed following central re-testing in 5–15% of cases.¹⁰ An important distinction must be drawn here between re-testing, potentially using different reagents and conditions, and the review of original diagnostic material. In ARTEMIS, we accepted a patient's eligibility as reported but reviewed critically the end point, which was very specifically pathological.

Review of pathology reports by the two Chief Investigators was made more difficult by a lack of standardization of how local reports were written—not all units use easy-to-read synoptic reports. Moreover, the majority of standardized reports are designed for the adjuvant setting without specific fields for the additional variables that need to be recorded post neoadjuvant therapy, such as tumor cellularity and fibrosis in lymph nodes. We accept that review of lesser degrees of tumor response to neoadjuvant chemotherapy by report review was imperfect particularly for residual cancer burden class 2 (an intermediate level of response) and this is reflected by our concordance data. However, the primary end point of this clinical trial was pathological complete response and we have shown that this is reported reliably by local pathologists. Standardization of routine reporting in clinical practice for neoadjuvant cases has been addressed recently by an international working group which should make this easier when designing future clinical trials.¹¹ It is possible that should such standardization be adopted, a measure of response such as residual cancer burden could be calculated locally. Also, although there is some evidence that pathologists are better at assessing chemotherapy response by reading pathology reports than are practicing clinicians,¹² this was not borne out by our data. It was evident on review of some of the cases where there were discrepancies between report review and central review that this was due to missing slides. In this trial, the two pathologists were involved in the resolution of disputed minimal residual disease on report review and were probably helpful in that area but minimal residual disease was not an end point of the trial. Furthermore, one must urge caution in trying to make direct comparisons between residual

cancer burden and more descriptive approaches to assessment of response to neoadjuvant chemotherapy particularly in equating residual cancer burden Class 1 with minimal residual disease. Residual cancer burden 1 is strongly dependent on low tumor cellularity while minimal residual disease as determined by the report review, where there was often no information on comparison of pre- and posttreatment tumor cellularity, was heavily influenced by residual tumor size and these are not always the same.

The literature on central pathology review of clinical trials is limited. The NSABP requires central pathology review for its randomized clinical trials and central reviewers are trained to operate with 90% concordance on pathological features compared with 65% concordance between local and central reporting in the NSABP B-18 trial for example.¹³ Recently reported central review of bone marrow fibrosis showed a concordance of 58% between central and local reporting, whereas a central panel of three reviewing pathologists achieved consistency of 88% for all three pathologists and 98% for two.¹⁴ However we were unable to find any reports of central pathology review in the context of neoadjuvant breast cancer trials.

We chose residual cancer burden as our method of measuring chemotherapy response primarily because it gave a numerical score that proved particularly convenient when it came to the cross-over study between the two pathologists. Its principal shortcoming is the lack of comparison with the baseline core biopsy, but from a clinical point of view, the tumor burden following chemotherapy is a sensible feature to measure and has been shown to correlate well with outcomes at 10 years follow-up.¹⁵ One of the important aspects of ARTEMIS is the future program of translational research and that has required sections from core biopsies, excised tumors and nodes to be marked up for future tissue sampling. A pathologist would certainly be required to support that aspect of a future trial. The central review process described here also provides great confidence in the recorded ARTEMIS end points, thus supporting subsequent translational work aimed at understanding the determinants of individual tumor response and the correlations of that with long-term outcomes. Recently, it has been shown that combining residual cancer burden with Ki67 measurements further increases the predictive power of this tool.¹⁶ Also there is a growing interest in either post-neoadjuvant studies, or allowing patients to enter other studies post neoadjuvant chemotherapy, which means that where low volumes of residual disease are permitted in such studies, perhaps caution is needed about relying on local reporting—whereas for pathological complete response or bulk residual disease one can rely on local reporting.

In conclusion, central pathology review of the ARTEMIS trial has allowed a direct comparison with

report review and has shown that when the primary end point of the trial, pathological complete response, is compared, the two methods are equally effective. Central pathology review has a place in the assessment of minimal residual disease but if that is not an agreed pre-specified trial end point there is little extra value in doing this. Learning from the experience of ARTEMIS, future neoadjuvant clinical trials could be improved by training in the routine calculation of residual cancer burden. Also, standardized routine reporting using report templates would greatly assist in report review.¹⁷ Such training might provide more robust reporting of residual cancer burden classes, facilitating future clinical management, when current and planned trials of adjuvant treatment in patients not achieving a pathological complete response to neoadjuvant therapy come to fruition.

Acknowledgments

The trial was funded by Cancer Research UK (CRUK/08/037 for trial coordination), Roche (provision of free study drug bevacizumab and unrestricted educational grant for trial coordination), and Sanofi–Aventis (unrestricted educational grant for trial coordination). We acknowledge the 116 investigators and their teams from 66 participating UK centers who entered patients into the ARTEMIS trial. Our gratitude also goes to the 800 women who kindly participated in our study. We thank the data and safety monitoring committee members: I Craig Henderson (University of California, San Francisco, CA, USA), Luca Gianni (Ospedale San Raffaele, Milan, Italy), Mark F Brady (Gynecologic Oncology Group Statistical & Data Center, Roswell Park Cancer Institute, Buffalo, NY, USA), and Xavier Pivot (Centre Hospitalier Regional Universitaire et Institut Regional Federatif de Cancerologie, Besancon, France). We thank the trials unit staff: phase 3 coordination at Warwick Clinical Trials Unit, Coventry, UK; translational coordination at University of Cambridge, Addenbrooke's Hospital, Cambridge, UK and University of Edinburgh Department of Oncology (BioBank); and statistical analysis at Warwick Clinical Trials Unit, UK.

Disclosure/conflict of interest

The authors declare no conflict of interest.

References

- 1 Earl HM, Hiller L, Dunn JA, *et al*. Efficacy of neoadjuvant bevacizumab added to docetaxel followed by fluorouracil, epirubicin and cyclophosphamide, for women with HER2 negative early breast cancer (ARTEMIS): an open-label randomised phase 3 trial. *Lancet Oncol* 2015;16:656–666.

- 2 Earl HM, Vallier A-L, Hiller L, *et al*. Effects of the addition of gemcitabine, and paclitaxel-first sequencing, in neoadjuvant sequential epirubicin, cyclophosphamide, and paclitaxel for women with high-risk early breast cancer (Neo-tAnGo): an open-label, 2 × 2 factorial randomised phase 3 trial. *Lancet Oncol* 2014;15:201–212.
- 3 Chavez-Mac Gregor M, Hubbard R, Meric-Bernstam F, *et al*. Residual cancer burden (RCB) in breast cancer patients treated with taxane- and anthracycline-based neoadjuvant chemotherapy: the effect of race. *J Clin Oncol* 2010;28(No 15 suppl):607.
- 4 Symmans WF, Peintinger F, Hatzis C, *et al*. Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy. *J Clin Oncol* 2007;25:4414–4422.
- 5 Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989;45:255–268.
- 6 Fisher ER, Constantino J. Quality assurance of pathology in clinical trials: The National Surgical Adjuvant Breast and Bowel Project Experience. *Cancer* 1994;74:2638–2641.
- 7 Residual Cancer Burden Calculator. Available from <http://www3.mdanderson.org/app/medcalc/index.cfm?pagename=jsonconvert3> (Accessed 8 September 2016).
- 8 Peintinger F, Sinn B, Hatzis C, *et al*. Reproducibility of residual cancer burden for prognostic assessment of breast cancer after neo-adjuvant chemotherapy. *Mod Pathol* 2015;28:913–920.
- 9 Naidoo K, Parham DM, Pinder SE. An audit of residual cancer burden reproducibility in a UK context. *Histopathology* 2017;70:217–222.
- 10 McCullough AE, Dell'Orto P, Reinholz MM, *et al*. Central Pathology review of Her-2 and ER in early breast cancer: an ALTTO Trial [BIG 2-06/NCCTG NO63D (Alliance)] ring study. *Br Cancer Res Treat* 2014;143:485–492.
- 11 Provenzano E, Bossuyt V, Viale G, *et al*. Standardization of pathological evaluation and reporting of post-neoadjuvant specimens in clinical trials of breast cancer: recommendations from an international working group. *Mod Pathol* 2015;28:1185–1201.
- 12 Provenzano E, Vallier A-L, Champ R, *et al*. A central review of histopathology reports after breast cancer neoadjuvant chemotherapy in the neo-tango trial. *Br J Cancer* 2013;108:866–872.
- 13 Fisher B, Brown A, Mamounas E, *et al*. Effect of preoperative chemotherapy on local-regional disease in women with operable breast cancer: findings from National Surgical Adjuvant Breast and Bowel Project B-18. *J Clin Oncol* 1997;15:2483–2493.
- 14 Pozdnyakova O, Rodig S, Bhandarkar S, *et al*. The importance of central pathology review in international trials: a comparison of local versus central bone marrow reticulin grading. *Leukemia* 2015;29:241–244.
- 15 Symmans WF, Wei C, Gould R, *et al*. Long-term prognostic risk after neoadjuvant chemotherapy associated with residual cancer burden and breast cancer subtype. *J Clin Oncol* 2017;35:1049–1060.
- 16 Sheri A, Smith IE, Johnston SR, *et al*. Residual proliferative cancer burden to predict long term outcome following neoadjuvant chemotherapy. *Ann Oncol* 2015;26:75–80.
- 17 Bossuyt V, Symmans WF. Standardizing of pathology in patients receiving neoadjuvant chemotherapy. *Ann Surg Oncol* 2016;23:3153.