Value of Three-Dimensional Ultrasound in Differentiating Malignant From Benign Breast Tumors

A Systematic Review and Meta-Analysis

Lian Bin, MM,* Yang Huihui, MM,† Yang Weiping, MB,‡ Wei Changyuan, MD,* Qin Qinghong, MD,* and Meng Weiyu, MB§

Abstract: Currently, the aim of differentiating malignant from benign breast tumors and diagnosing early breast cancer is more accurately fulfilled with the addition of ultrasound (US) scans. Some studies have shown that threedimensional (3D) imaging opens a new diagnostic window because of a better and more detailed impression of the spatial arrangement of focal breast masses. Therefore, we carried out this study to assess the efficacy of 3D-US diagnostic techniques for the detection of breast cancer using systemic analyses. We searched the MEDLINE, Pubmed, EMBASE, and Cochrane Library for using 3D-US for the diagnosis of breast cancer. The terms used were “three-dimensional ultrasound,” “breast neoplasms,” “sensitivity and specificity,” and “accuracy”. Patients were adults. The quality of the studies was checked according to Quality Assessment of Diagnostic Accuracy Studies criteria. We calculated the diagnostic accuracy rates and developed a meta-analysis to synthesize the results. Twelve studies of diagnostic tests were selected, which were of moderate to high quality. The pooled sensitivity was 0.923 (95% confidence interval [CI], 0.896–0.945), and the pooled specificity was 0.872 (95% CI, 0.849–0.893). The meta-analysis detected a threshold effect and heterogeneity between studies. The pooled positive likelihood ratio was 6.965 (95% CI, 5.242–9.255), the negative likelihood ratio was 0.106 (95% CI, 0.079–0.142), and the diagnostic odds ratio was 84.239 (95% CI, 52.237–135.84). This meta-analysis indicates that 3D-US has acceptable diagnostic accuracy rates in differentiating malignant from benign breast tumors.

Key Words: breast cancer, three-dimensional imaging, ultrasound, meta-analysis

As a common malignancy of women, breast cancer's incidence has increased year by year. Today, the aim of diagnosing early breast cancers is more accurately fulfilled with the addition of ultrasound (US) scans. As a tool for diagnosing breast cancer, US's noninvasive, high sensitivity, and good reproducibility show a superiority that cannot be replaced. In manual US, a breast lesion has to be immediately characterized during the examination. The addition of US screening can be especially useful in patients with dense breasts on mammography.

However, a weakness of this method is the compromised possibility of a second evaluation on hard copies due to the lack of standardization of sonographic documentation. It has been reported that combining mammography and sonomammography for women with dense breasts or other high-risk factors results in a higher sensitivity to breast cancer diagnosis. Previous studies have shown that 3D imaging opens a new diagnostic window because of a better and more detailed impression of the spatial arrangement of focal breast masses.

The aim of our study was to determine, by meta-analysis, the diagnostic performance of 3D-US in patients with breast lesions and provide evidence-based medicine for clinical evidence.

MATERIALS AND METHODS

Search Strategy

We electronically searched The Cochrane Library (issue 2017), Pubmed (until December 2017), MEDLINE (until December 2017), and EMBASE (until December 2017) according to the Cochrane Search Strategy for the identification of trials, using the words “three-dimensional,” “ultrasound,” “breast neoplasms,” “sensitivity and specificity,” and “accuracy.” We also searched the China Biological Medicine Database (CBM-disc, 1979 to 2017), VIP Chinese Journals Database (1968 to 2017), and China National Knowledge Infrastructure Whole Article Database (CNKI, 1994 to 2017). The list of articles was supplemented with extensive cross-checking of the reference lists of all retrieved articles.

Study Eligibility

Eligibility criteria are as follows: (1) examined 3D-US for diagnosis of breast nodules or mass lesions; (2) enrolled at least 30 participants with breast nodules or masses; (3) cytology or...
pathological examination was regarded as the golden diagnostic standard; (4) data from each study were organized systematically and purged to obtain the true positive (TP), false positive (FP), false negative (FN), and true negative (TN) values (Table 2); and (5) no language restriction was used.

Exclusion criteria are as follows: (1) review articles, letters, comments, case reports, unpublished articles, and articles that did not include raw data were not selected. (2) Studies that contained the same sample.

Data Extraction

Two investigators independently evaluated potential studies, and a checklist was used to determine final eligibility. Disagreements about inclusion or exclusion were resolved by consensus. For eligible studies, we collected the following clinical features: (1) study authors, (2) year of publication, (3) age, (4) number of subjects, (5) model of the equipment, (6) sampling method, (7) benign and malignant lesions evaluated, (8) inclusion and exclusion criteria, (9) number of breast lesions, (10) study design (prospective, retrospective, or unclear), and (11) technical characteristics of the 3D-US diagnostic parameters considered.

Study Quality

Quality assessment was performed according to an assessment system including 14 items proposed by the Quality Assessment of Diagnostic Accuracy Studies (QUADAS), disagreement was resolved by consensus. The 14 items, phrased as questions, are scored as “yes,” “no,” or “unclear.” The quality assessment score can range from 0 (observed minimum) to 14 (observed maximum). A more detailed description of each item, together with a guide on how to score each item, is provided in the article by Whiting et al. 4

Statistical Analysis

The meta-analysis was performed using MetaDisc 1.4 (for Windows; XI Cochrane Colloquium; Barcelona, Spain) and Stata 14.0 and 9.0 (the Midas module in the Stata software, version r.14.StataCorp) software. We calculated the pooled sensitivity and specificity, positive and negative likelihood ratios, and the diagnostic odds ratio (OR) of 3D-US, along with the respective 95% confidence intervals (CIs), using a bivariate meta-analysis model. We constructed a hierarchical summary receiver operating characteristic (SROC) curve plotting sensitivity versus specificity, and calculated the area under the curve. These data were pooled using a fixed-effects model (Mantel-Haenszel method). We compared these findings with results obtained using a random-effects model (DerSimonian and Laird method). Cochran Q was used to determine the probability coefficients and the OR. A χ² test for heterogeneity was reported for each clinical feature. P value of 0.05 or less was considered to denote statistically significant heterogeneity.

TABLE 1. Literature and Design Information

| Authors          | Year | Country | Median Age | No. Cases/No. Lesions | Prospective/Retrospective | Blinded | Model            | Sampling Method | Score |
|------------------|------|---------|------------|-----------------------|---------------------------|---------|------------------|-----------------|-------|
| Blohmer et al 8  | 1996 | Germany | unclear    | 50/50                 | Prospective               | Single  | Combi 530 unit   | Unclear         | 11    |
| Rotten 9         | 1999 | France  | 47         | 186/186               | Prospective               | Single  | Combi 530 unit   | Random          | 12    |
| Chen et al 10    | 2003 | China   | unclear    | 161/161               | retrospective             | Single  | Voluson 530      | Random          | 11    |
| Kotsianos et al 11 | 2005 | Germany | unclear    | 60/60                 | Prospective               | Single  | Single           | Single          | 10    |
| Choa et al 12    | 2005 | Korea   | 44         | 202/202               | Retrospective             | Double  | logiq 9          | Unclear         | 13    |
| Chen et al 13    | 2005 | China   | unclear    | 161/161               | Retrospective             | Single  | Voluson 530      | Random          | 11    |
| Kotsianos et al 14 | 2009 | China   | 52         | 97/107                | Prospective               | Double  | logiq 9          | Unclear         | 13    |
| Kalmantis et al 15 | 2009 | Greece  | 45         | 125/153               | Retrospective             | Single  | Voluson 730      | Unclear         | 12    |
| Chang et al 16   | 2011 | Korea   | 47         | 67/70                 | Retrospective             | Single  | Voluson 68       | Continuous      | 13    |
| Chang et al 17   | 2012 | China   | unclear    | 113/113               | Retrospective             | Single  | Voluson 730      | Continuous      | 12    |
| Lai et al 18     | 2013 | China   | 50         | 82/82                 | Retrospective             | Single  | Voluson E8       | Continuous      | 13    |
| Huang et al 19   | 2013 | China   | unclear    | 107/113               | Retrospective             | Single  | Voluson 730      | Random          | 12    |

TABLE 2. Documents on the Parameters of Diagnostic Tests

| Authors          | TP | FP | FN | TN | Sen (%) | Spe (%) | LR+ | LR− | Acc(%) |
|------------------|----|----|----|----|---------|---------|-----|-----|-------|
| Blohmer JU       | 19 | 4  | 0  | 27 | 100.00  | 87.10   | 6.93| 0.03| 92.00 |
| Rotten D         | 53 | 8  | 5  | 128| 91.38   | 94.12   | 15.53| 0.09| 93.30 |
| Chen DR          | 50 | 9  | 4  | 98 | 92.59   | 91.59   | 11.01| 0.08| 91.93 |
| D. Kotsianos (2005) | 37 | 6  | 1  | 16 | 97.37   | 72.73   | 3.57 | 0.04| 88.33 |
| Choo KR          | 31 | 33 | 0  | 138| 100.00  | 80.70   | 5.05 | 0.02| 83.66 |
| Chen WM          | 48 | 7  | 6  | 100| 88.89   | 93.46   | 13.59| 0.12| 91.93 |
| D. Kotsianos(2009) | 38 | 5  | 1  | 63 | 97.44   | 92.65   | 13.25| 0.03| 94.39 |
| K. Kalmantis      | 49 | 10 | 4  | 62 | 92.45   | 86.11   | 6.66 | 0.09| 88.80 |
| Changa JM        | 17 | 9  | 1  | 37 | 94.44   | 80.43   | 4.83 | 0.07| 84.38 |
| Chang YC         | 45 | 8  | 4  | 56 | 91.84   | 87.50   | 7.35 | 0.09| 89.38 |
| Lai YC           | 35 | 6  | 6  | 35 | 85.37   | 85.37   | 5.83 | 0.17| 85.37 |
| Huang YH         | 46 | 13 | 7  | 47 | 87.04   | 77.97   | 4.01 | 0.17| 82.30 |

Sen, sensitivity; Spe, specificity; Acc, accuracy; LR+, positive likelihood ratio; LR−, negative likelihood ratio.
Statistical heterogeneity was evaluated using the $I^2$ statistic, which assesses the appropriateness of pooling individual study results. Heterogeneity was specifically examined using $I^2$, where $I^2$ values of 50% or more indicate a substantial level of heterogeneity. When heterogeneity was found, we attempted to determine potential reasons for it by examining the individual study and subgroup characteristics. First, we evaluated the presence of a threshold effect on the accuracy of the 3D-US with the Spearman correlation coefficient between the logits of sensitivity and specificity. Second, we included it as covariates in univariate meta-regression analysis (inverse variance weighted). We also analyzed the effects of other covariates on diagnostic OR (DOR) (ie, publication year, age, number of subjects, model of the equipment, sampling method, quality of the study, continent of study origin, number of breast lesions, and study design). The relative DOR (RDOR) was calculated according to standard methods to analyze the change in diagnostic precision in the study per unit increase in the covariate.

Because publication bias is of concern for meta-analyses of diagnostic studies, we tested for the potential presence of this bias using funnel plots. When there is no bias, the graphical distribution is a symmetrical inverted funnel. If the funnel plot is asymmetrical or incomplete, there may be publication bias. We then applied the Deek’s test to detect funnel plot asymmetry: if the $P$ value is less than 0.05, there was funnel plot asymmetry, suggesting potential publication bias; if the $P$ value is greater than 0.05, there was no significant publication bias. An asymmetric funnel plot would suggest that additional small studies may have been conducted but not published because of unfavorable results.

## RESULTS

### Characteristics of the Included Studies

The electronic searches and hand searches identified a final 12 studies10-19 (Table 1). Five studies8,9,11,14,15 were performed in Europe. Seven studies10,12,13,16-19 were performed in Asia. All studies were case series (8 retrospective, 4 prospective). There were a total of 1411 patients and 1432 primary lesions, of which 925 were benign lesions, and 507 were malignant lesions (Table 2).

### Quality Assessment

Eleven studies8-10,12-19 were clearly described; the selected literature used biopsy as the reference standard; and 2 studies12,14 used a rigorous experimental design, and was double-blind, and therefore were of relatively high quality (Table 3).

### Data Synthesis

Figure 1A-E shows the forest plot of sensitivity and specificity for 3D-US in the diagnosis of breast cancer. The pooled sensitivity was 0.923 (95% CI, 0.896–0.945), and the pooled specificity was 0.872 (95% CI, 0.849–0.893). The positive likelihood ratio (PLR) was 6.965 (95% CI, 5.242–9.255), negative likelihood ratio (NLR) was 0.120 (95% CI, 0.079–0.142), and DOR was 84.239 (95% CI, 52.237–135.84). $\chi^2$ values of sensitivity, specificity, PLR, NLR, and DOR were 16.87 ($P = 0.001 < 0.1$), 28.44 ($P = 0.003 < 0.1$), 10.96 ($P = 0.447 > 0.1$), and 14.07 ($P = 0.229 > 0.1$), respectively. Specificity, PLR had a significant heterogeneity between studies.

The analysis was based on the SROC curve. Sensitivity and specificity for the single test threshold identified for each study were used to plot an SROC curve. A random-effects model was used to calculate the average sensitivity, specificity, and other measures across studies.20 A graph of the SROC curve for the 3D-US showing TP rates versus FP rates from individual studies is shown in Figure 2. Our data showed that the SROC curve is positioned near the desirable upper left corner of the graph, and that the maximum joint sensitivity and specificity

### Statistical Heterogeneity

The analysis was based on the SROC curve. Sensitivity and specificity for the single test threshold identified for each study were used to plot an SROC curve. A random-effects model was used to calculate the average sensitivity, specificity, and other measures across studies. A graph of the SROC curve for the 3D-US showing TP rates versus FP rates from individual studies is shown in Figure 2. Our data showed that the SROC curve is positioned near the desirable upper left corner of the graph, and that the maximum joint sensitivity and specificity...
(ie, the Q value) was 0.9008; meanwhile, the area under the curve was 0.9575, indicating a high level of overall accuracy.

The Deeks' funnel plot in Figure 3 shows that the studies were distributed symmetrically with a P value of 0.82. These results did not indicate a potential for publication bias.

Meta Regression and Subgroup Analyses

Meta-regression analysis was then performed to explore other potential sources of heterogeneity. It included publish year (1996–2005 vs 2009–2013), continent of study origin (Asian vs European), predesign (prospective vs retrospective), sampling method (random/continues vs unclear), and quality of the study (QUADAS scores 10–11 vs 12–13). Finally, none of the variables had significant heterogeneity (P > 0.05), so subgroup analysis was not necessary (Table 4.)

**DISCUSSION**

The SROC plot shows the sensitivity and specificity of 3D-US for the diagnosis of breast lesions. The width of the circles is proportional to the number of patients in each study. The square is the summary value for sensitivity and specificity. The DOR is a single indicator of test accuracy that combines the data from sensitivity and specificity into a single number. The value of a DOR ranges from 0 to infinity, with higher values indicating better discriminatory test performance (ie, higher accuracy). In the present meta-analysis, we found that

**FIGURE 1.** A–E, Forest maps of 3D-US in the diagnosis of benign and malignant breast lesions, showing sensitivity, specificity, DOR, positive likelihood ratio, and negative likelihood ratio.

---

© 2019 The Author(s). Published by Wolters Kluwer Health, Inc.
the mean DOR was 84.239, indicating a high level of overall accuracy. The PLR and NLR were also presented to show diagnostic accuracy because the SROC curve and DOR are not easy to interpret and use in clinical practice, and ratios are considered to be more clinically meaningful. Likelihood ratios greater than 10 or less than 0.1 generate large and often conclusive shifts from pretest to posttest probability (indicating high accuracy). In our first classification, a PLR value of 6.965 suggests that patients with various grades of cancers have an approximately 7-fold higher chance of receiving a positive 3D-US result compared with patients with benign breast lesions. This probability would not be considered high enough to begin surgical treatment or other therapy. However, NLR was found to be 0.106 in our current meta-analysis. If the 3D-US result was negative, the probability that this patient has breast cancer is approximately 11%.

Funnel plots graphically show publication bias. An asymmetric funnel plot would suggest that additional small studies may have been conducted but not published because of unfavorable results.

Bias was considered in this meta-analysis. To avoid selection bias, we used not only the MEDLINE and EMBASE databases but also ScienceDirect, Springerlink, and Scopus.

To minimize bias in the selection of studies and in data extraction, reviewers who were blinded to the journal, author, institution, and date of publication, independently selected articles on the basis of inclusion criteria. In addition, scores were assigned to study design characteristics and examination results by using a standardized form that was based on the QUADAS tool (Table 3). The QUADAS tool is an evidence-based quality assessment tool developed for use in systematic reviews of studies of diagnostic accuracy.4

However, some limitations still exist in this meta-analysis. A potential limitation of any meta-analysis is the possibility of publication bias because studies with optimistic results may be more readily published than studies with unfavorable results, and studies with a large sample size may be published more readily than studies with a small sample size. We attempted to examine publication bias by evaluating whether the size of the studies was associated with the results for diagnostic accuracy. No association was found between sample size and diagnostic accuracy.

Inclusion bias should also be considered. The studies were from 5 countries and regions, the study sample was large. Compared with a single study with high reliability, these studies

**TABLE 4. Sensitivity, Specificity for Each Subgroup**

| Parameter       | Category          | Studies (n) | Sensitivity | Specificity | $\chi^2$ | P   | I²  |
|-----------------|-------------------|------------|-------------|-------------|---------|-----|-----|
| Year            |                   | 6          | 0.94 [0.91–0.98] | 0.89 [0.85–0.93] | 2.95    | 0.23| 32  |
|                 | 2009–2013         | 6          | 0.91 [0.87–0.95] | 0.86 [0.80–0.91] |         |     |     |
| Continent       | Asian             | 7          | 0.91 [0.87–0.94] | 0.86 [0.82–0.91] | 3.78    | 0.15| 47  |
|                 | European          | 5          | 0.95 [0.92–0.98] | 0.89 [0.84–0.94] |         |     |     |
| Predesign       | Prospective       | 4          | 0.96 [0.92–0.99] | 0.87 [0.80–0.94] | 4.88    | 0.09| 59  |
|                 | Retrospective     | 8          | 0.91 [0.88–0.94] | 0.88 [0.84–0.92] |         |     |     |
| Sampling method | Random/continues  | 8          | 0.91 [0.88–0.94] | 0.88 [0.84–0.92] | 4.35    | 0.11| 54  |
|                 | Unclear           | 4          | 0.96 [0.93–0.99] | 0.87 [0.80–0.94] |         |     |     |
| Score           | 10–11             | 4          | 0.93 [0.90–0.97] | 0.89 [0.84–0.95] | 0.99    | 0.61| 0   |
|                 | 12–13             | 8          | 0.92 [0.89–0.95] | 0.87 [0.82–0.91] |         |     |     |
showed good clinical practical value. However, a large time span between studies (a maximum of 18 years) does not exclude equipment replacement, perhaps this is the cause of heterogeneity in specificity. Current clinical practices use integrated imaging approaches for the diagnosis of breast lesions. However, there is a lack of in-depth studies of decision making systems for breast cancer diagnosis, and the effect of cost remains an open question.

In conclusion, our meta-analysis shows that 3D-US has high sensitivity (92.3%) and specificity (87.2%) in differentiating malignant from benign breast tumors. As a noninvasive method, 3D-US can be used for mammography and ultrasound examinations, but high-quality prospective studies are required to further confirm that this technology has improved the clinical value of breast ultrasound imaging.

ACKNOWLEDGMENTS

The authors would like to thank the Guangxi Natural Science Foundation (2015GXNSFAA414003) of China for the financial support.

REFERENCES

1. Hooley RJ, Andrejeva L, Scoutt LM. Breast cancer screening and problem solving using mammography, ultrasound, and magnetic resonance imaging [J]. Ultrasound Q. 2011;27(1):23–47.
2. Berlin JA. Does blinding of readers affect the results of meta-analyses? University of Pennsylvania Meta-analysis Blinding Study Group [J]. Lancet. 1997;350:185–186.
3. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method [J]. Med Decis Making. 1993;13:313–321.
4. Whiting P, Rutjes AW, Reitsma JB, et al. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol. 2003;3:25.
5. Higgins JP, Thompson SG, Deeks J, et al. Measuring inconsistency in meta-analyses. BMJ. 2003;327:557–560.
6. Glas AS, Lijmer JG, Prins MH, et al. The diagnostic odds ratio: a single indicator of test performance. J Clin Epidemiol. 2003;56:1129–1135.
7. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics. 1994;50:1088–1101.
8. Blohmer JU, Heinrich G, Paepke S, et al. Three-dimensional ultrasound study (3-D sonography) of the female breast [J]. Geburtshilfe Frauenheilkd. 1996;56(4):161–165.
9. Retten D, Levaillant JM, Zerat L. Analysis of normal breast tissue and of solid breast masses using three-dimensional ultrasound mammography [J]. Ultrasound Obstet Gynecol. 1999;14:114–124.
10. Chen DR, Chen WM, Moon WK. Computer-aided diagnosis for 3-dimensional breast ultrasonography [J]. Arch Surg. 2003;38:296–302.
11. Kotsianos D, Fischer T, Hiltawsky K, et al. 3D ultrasound (3D US) in the diagnosis of focal breast lesions [J]. Radiologe. 2005;45:237–244.
12. Choa KR, Lee C, Pisanod ED, et al. A comparative study of 2D and 3D ultrasonography for evaluation of solid breast masses [J]. Eur J Radiol. 2005;54:365–370.
13. Chen WM, Chang CS, Moon WK, et al. 3-D ultrasound texture classification using run difference matrix [J]. Ultrasound Med Biol. 2005;31(6):763–770.
14. Kotsianos D, Hiltawsky KM, Wirth S, et al. Analysis of 107 breast lesions with automated 3D ultrasound and comparison with mammography and manual ultrasound [J]. Eur J Radiol. 2009;71:109–115.
15. Kalmantis K, Dimitrakakis C, Koupis C, et al. The contribution of three-dimensional power Doppler imaging in the preoperative assessment of breast tumors: a preliminary report [J]. Obstetrics and Gynecology International. 2009;1–8.
16. Chang YC, Hwang VH, Huang CS, et al. Vascular morphology and tortuosity analysis of breast tumor inside and outside contour by 3-D power Doppler ultrasound [J]. Ultrasound Med Biol. 2012;38(11):1859–1869.
17. Lai YC, Huang YS, Wang DW, et al. Computer-aided diagnosis for 3-D power Doppler breast ultrasound [J]. Ultrasound Med Biol. 2013;39(4):555–567.
18. Huang YH, Chen JH, Chang YC, et al. Diagnosis of solid breast tumors using vessel analysis in three-dimensional power Doppler ultrasound images [J]. J Digit Imaging. 2013;26(4):731–739.
19. Dinnes J, Deeks J, Kirby J. A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. Health Technol Assess. 2005;9:1–113.