Supplementary File 4: Narrative synthesis and additional tables for Chapter 2, Clinical utility of Oncotpye DX

In this review, clinical utility relates to the impact of the prospective use of the test on patient outcomes such as survival and recurrence. The ideal study design would be an RCT where patients are randomised to treatment guided by the test or treatment according to usual practice. Additional study designs for clinical utility are observational cohorts (either prospective or retrospective) where patients received the test prospectively in clinical practice, and data are available for both the test results and clinical outcomes. These observational designs are at higher risk of bias from confounding.

Five data sets reported across nine published references[1-9] and one AIC manuscript[10] reported evidence relating to the clinical utility of Oncotype DX and met the inclusion criteria for the review. One further study[9, 11, 12] did not meet the inclusion criteria for the review in that the follow up was less than 5 years (for outcome BCSS). We have presented data relating to this study as it was the only identified study presenting subgroup analyses for micrometastases and by race, both of which were subgroups specified in the NICE scope[13] and for which there are very limited data.

Study design and chemotherapy rates: Oncotype DX clinical utility

Study characteristics are presented in Table 1. Two studies had a prospective trial design.[1-4] Only one study, the Trial Assigning Individualized Options for Treatment (TAILORx),[1] randomises patients to treatment guided by the test or treatment according to usual practice. This study aims to assess the clinical utility of Oncotype DX. Women with RS<11 were assigned to endocrine therapy alone, while women with RS 11-25 were randomised to either endocrine therapy plus chemotherapy or endocrine therapy alone. As of July 2017, this study had only reported results for the low-risk (RS<11) group (n=1626). Data for this group are effectively prospective observational data.

The West German Study Group Plan B (WSG Plan B)[2-4, 14] trial (n=3198) is also a prospective RCT, but does not aim to assess the clinical utility of Oncotype DX, as it randomises patients with RS≥12 to two different sorts of chemotherapy. However, a translational research aim was to assess the risk of recurrence in patients with RS <12 who were not treated with adjuvant chemotherapy. This group is again effectively a prospective observational cohort.

Three studies had an observational design and were retrospective analyses of routinely collected data at three centres or areas: MD Anderson Cancer Centre in the USA (n=1030),[5] Clalit Health Services[7, 8] in Israel (n=1594 LNmic-LN3; n=627 LN0-LNmic; additional analyses[10] were provided to the EAG as Academic in Confidence data but cannot be reported here), and the Memorial Sloan Kettering

Centre in the USA (n=1406).[6] In all cases, treatment was given according to routine clinical practice, including the Oncotype DX RS, which resulted in differing levels of chemotherapy being prescribed per risk group and per study. Chemotherapy ranged from 1%[8] to 12%[6] in low RS groups (RS <18), from 26%[8] to 43%[5] in the intermediate-risk group (RS 18-30) 89%[8] to 90%[5] in the high-risk group.

The study that did not meet the inclusion criteria (due to insufficient follow-up length) was of a similar design to the other retrospective analyses, and was based on the prospectively maintained SEER (Surveillance, Epidemiology, and End Results) database and Genomic Health's clinical laboratory database.[11, 12] Chemotherapy rates for low (RS <18), intermediate (RS 18-30) and high (RS>30) risk patients were 7%, 34% and 69% in lymph node negative patients, respectively, and somewhat higher at 23%, 47% and 75% in lymph node positive patients, respectively.

Patients: Oncotype DX clinical utility

Prospective trials: Both trials[1-4] recruited HR+, HER2- patients, but TAILORx recruited LN0 patients with tumours sized 1.1 to 5cm (or 0.6 to 1.0cm in intermediate or high-risk tumours), whilst WSG Plan B recruited clinically high-risk (pT1-T4c; LN+ (or LN0 with a risk factor (CpT2, grade 2/3, high uPA/PAI-1, <35 years, or HR-negative))) patients with 0 to three positive lymph nodes.

Observational studies: All three data sets[5-8] recruited ER+, HER2- patients and only recruited patients who had had an Oncotype DX test. It was not always clear how (or even whether) patients were selected for the test, and how this may have affected the patient spectrum. The MD Anderson study recruited only Stage 1 patients,[5] the Memorial Sloan Kettering study recruited Stage 1 and 2 patients[6] and the Clalit Health Services study did not restrict by stage of disease.[7] The MD Anderson and Memorial Sloan Kettering studies recruited only patients with no or micro lymph node metastases (LN0-LNmic).[5, 6] The Clalit Health Services reported two subgroups across two publications:[7, 8] patients with LN0-LNmic[8] and patients with micro metastases or between one and three lymph node metastases (LNmic – LN3).[7, 8] Additional analyses[10] were provided to the EAG as Academic in Confidence data but cannot be reported here

The study that did not meet the inclusion criteria (SEER database)[11, 12] recruited patients with LN0 to LN3, and subgrouped patients according to age (40-85 years), lymph node status (LN0, LNmic-LN3, LNmic alone) and race (black, white, other).

Quality assessment: Oncotype DX clinical utility

The highest level of evidence for clinical utility is an RCT of treatment guided by the test versus treatment guided according to usual practice. Assessment with the Cochrane risk of bias tool for RCTs indicates all studies are of poor quality to meet this aim (Table 2).

Results: Oncotype DX clinical utility

Data relating to the clinical utility of Oncotype DX are presented in Table 9 of the main report. Whilst all studies report data relating to recurrence or survival, differences in cut off points (RS<11, <12 and <18), patient populations (clinically high-risk, LN0, LN+), treatment regimens (some patients had chemotherapy in some studies) and outcome measures (DRFS, DFS, DRFI, BCSS, OS) precluded a meaningful meta-analysis.

Whilst two studies use RCT datasets, neither presents data for the test versus usual practice. As such, the evidence base is exclusively single-armed in nature and cannot address the question of whether the test can improve patient outcomes compared to usual practice. It can, however, reveal something about the ability of the test to identify a group at very low risk of recurrence who could avoid chemotherapy. Data relating to risk in intermediate and high-risk categories are, without a no-test comparator arm, difficult to interpret in the context of clinical utility. The results presented here are therefore divided into two subsections:

- Outcomes in low-risk patients: Assessing the ability of the test to identify a group of patients at low-risk of recurrence who can avoid chemotherapy
- Outcomes in intermediate- and high-risk patients treated according to clinical practice: Observational data relating to clinical outcomes in these patients.

A further section relating to protocol-defined subgroups then follows:

- Outcomes in protocol-defined subgroups.

Outcomes in low-risk patients

DRFS: The TAILORx trial[1] and the MD Anderson observational study[5] reported 5-year DRFS in low-risk patients. DRFS appears very low for patients with RS<11 (99.3%)[1] but somewhat higher when the cut point is increased to RS<18 (95.9%)[5] even though this study included only Stage 1 patients.

DRFI: The Clalit Health[7, 10] and the Memorial Sloan Kettering[6] observational studies reported DR rates at 5 years, which have been converted into 5-year DRFI (proportion free of distant recurrence, not including death, at 5 years) for ease of comparison with other outcomes.

In both studies,[6-8] a proportion of patients received chemotherapy in all risk groups (Table 9 of the main report). In the LN0-LNmic group, 5-year DRFI in the low-risk group (RS<18) was similar in both studies, at 99.5% (95% CI: 98.4, 99.8)[8] and 99.6% (95% CI: NR)[6] respectively, although chemotherapy rates were somewhat different at 1% and 12%, respectively. In the LNmic-LN3 group,

reported for the Clalit Health study only, DRFI in the low-risk group (RS<18) was lower at 96.8% (95% CI: NR).[7]

For LN0-LNmic patients, a lower cut point for low-risk patients (RS<11) was reported in the Memorial Sloan Kettering study[6] and the proportion of patients free from distant recurrence at 5 years was higher compared to RS<18, at 99.9% (95% CI: NR). For LNmic-LN3 patients, the lower cut point of RS<11 surprisingly resulted in a DRFI of 95.1% (95% CI: NR), which was slightly lower than for RS<18 (96.8%; 95% CI: NR).[7]

IDFS: The WSG Plan B study[2-4] reported 5-year IDFS, at cut points RS<12 for low-risk, as 94.2% (95% CI: 91.2, 97.3). TAILORx[1] reported IDFS for low-risk (RS<11) patients as 93.8% (95% CI: 92.4, 94.9%).

BCSS/OS: OS was reported in the TAILORx study,[1] and BCSS (converted from breast cancer death rates) was reported in the Clalit Health study for both subgroups (LN0-mic and LNmic-LN3)[7, 8] and for the SEER registry.[11, 12] OS was reported in the WSG Plan B study,[2-4] but follow up was less than 5 years and the data were not extracted. 5-year OS in TAILORx[1] was 98.0% (95% CI: 97.1, 98.6%) for patients with RS<11. In the Clalit Health study, LN0-1mic with RS<18, BCSS was 99.9% (95% CI: 99.0, 100.0%).[8] For the LNmic-LN3 subgroup of the Clalit Health study,[7] BCSS was 98% in RS<11 patients and 99.1% in RS<18 patients.

Outcomes in intermediate and high-risk patients

DRFS: The MD Anderson study[5] also reported 5-year DRFS for the high-risk group. This was 76.4% (95% CI: 59.2, 87.1%). The difference between risk groups was statistically significant in an unadjusted analysis (p<0.0001) and non-significant in a multivariable analysis (p=0.083 for high vs. low; p=0.066 for intermediate vs. low).

DRFI: Data on intermediate and high-risk groups were reported in the Clalit Health study for both LN0-1mic[8] and for the LNmic-LN3[7] groups. DRFI decreased with increasing risk group in both subgroups but formal statistical comparisons were not reported. The LNmic-LN3 subgroup had lower 5-year DRFI in all risk groups (DRFI RS<18: 96.8%; RS18-30: 93.4%; RS>30: 83.6%) compared with LN0-LNmic (DRFI RS<18: 99.5%; RS18-30: 98.8%; RS>30: 93.1%) Data using the 11-25 cut offs were not reported for LN0-LNmic, but resulted in different DRFI in LNmic-LN3 patients(DRFI RS<11: 95.1%;RS11-25: 96.1%; RS>25: 86.8%).

IDFS: The WSG Plan B study[2-4] reported 5-year IDFS, at cut points 12-25 for intermediate-risk and >25 for high-risk. These were 94% and 84% respectively, with p<0.001 between groups

(multivariable p=0.001). TailorX[1] reported IDFS for low-risk (RS<11) patients as 93.8% (95% CI: 92.4, 94.9%).

BCSS/OS: OS was not reported for the intermediate- and high-risk groups in TAILORx.[1] 5-year BCSS for intermediate- and high-risk groups in the LN0-1mic group of the Clalit Health Services study[8] were 98.5% (95% CI: 97.1, 99.2%) and 90.6% (95% CI: 84.5, 94.4%) respectively (p<0.00)1 between risk groups, and 97.4% (95% CI: NR) and 86.9% (95% CI: NR) in the LNmic-LN3 subgroups (p-value not reported) of the Clalit Health Services study.[7]

Outcomes in protocol-defined subgroups

Micrometastases: The NICE scope lists micrometastases as a subgroup of interest to the assessment. Only one study that met the inclusion criteria for the review reported data for patients with micrometastases separately (Clalit Health Services),[7] and as such an additional study (SEER database)[11, 12] that followed up patients for <5 years and reported actuarial 5 year BCSS was included.

In the Clalit Health Services LNmic-LN3 analysis,[7] 5-year DRFI was generally higher in the LNmic group compared to the LN1mic LN-3 group, for example, for low-risk patients (RS<18) DRFI was 99.3% (95% CI: NR) and 96.8% (95% CI: NR) respectively. However, BCSS was very similar in each group at 99.3% (95% CI: NR) and 99.1% (95% CI: NR), respectively.

The SEER registry data[9, 11, 12] reported subgroups of LN0 (ages 40-84 years), LN1-LN3 (all ages) and LNmic (all ages). Actuarial 5 year BCSS for low-risk patients (RS<18) were similar at 99.6% (95% CI: 99.4%, 99.7%), 98.9% (95% CI: 97.4, 99.6%) and 99.4% (95% CI: 97.4, 99.9%), respectively (though data for micrometastases is from a later publication with more patients).[9] Data were also similar across subgroups within the intermediate group (LN0 98.6, LN+ 97.7) and high-risk group (LN0 95.6, LN+ 85.7). There was a statistically significant difference between groups for LN0 (p<0.001, unadjusted and multivariable) and LN+ patients (p<0.001 for unadjusted; not reported for multivariable; Table 9 of the main report).

Race: The NICE scope lists race as a subgroup of interest to the assessment. Only the SEER registry data[11, 12] (which followed up patients for <5 years and reported actuarial 5 year BCSS) reported an analysis by race, whereby patients were categorised as white, black or other. Data were reported for LN0 and LN1-3 patients separately, and showed generally similar rates across race categories, within risk categories (Table 9 of the main report).

Table 1: Clinical utility studies: Oncotype DX

Reference; N	Cohorts	Country	Study design	Details of test	Cut-offs	Population	Nodal status	Endo / chemo
Sparano 2015[1] LN0, N=1626	TAILORx	USA	Prospective cohort (within an RCT)	FFPE Genomic Health	RS<11 pts only	100% HR+ 100% HER2- 70% postmeno 100% female Tumour size 1.1 to 5cm, or 0.6-1.0cm with inter/high grade, indicated for CT[a]	LN0	100% ET 100% CT
Le Du 2015[5] N=1030	MD Anderson	USA	Retrospective cohort study	NR	11-25	100% ER+ 100% HER2- 64% postmeno 100% female Stage I disease Had O-DX test	LN0/LNmic	98% ET 27% CT Treated according to usual practice with O-DX
Nitz 2017 [3, 4, 14] N=2642	WSG PlanB	Germany	Prospective cohort (within an RCT)	NR Genomic Health	12-25	100% HR+ 100% HER2- Pre/post meno 100% female High clinical risk [d]	LN0-3 LN0 58.8% LN1-3 41.2%	Treated according to RS: RS<12 endo only RS≥12, chemo + endo[e]
Stemmer 2016[7] Stemmer 2016[8] 1)LN0-1mic, N=1594[8] 2)LN1mic – LN3, N=627[7]	Clalit Health Services	Israel	Retrospective cohort study	NR	11-25 18-30	100% ER+ 100% HER2- Meno NR Had O-DX test	1) LN0 90% LNmic 10% 2)LNmic-LN3	Treated according to usual practice with O-DX test 1) % ET NR 20% CT 2) % ET NR 27% CT

Wen 2017[6] N=1406	Memorial Sloan Kettering	USA	Retrospective cohort study	NR	RS <18 pts only Cut point RS 11	100% HR+ 100% HER2- 64% postmeno 99.9% female All pts tumour >0.5cm routinely tested and some <0.5cm RS<18 only	LN0-mic	Treated according to usual practice with O-DX test 97% ET 12% CT
Petkov 2016[15] Roberts 2016[12] 1) LN0, all ages N=40,134 2) LNmic-LN3, all ages, N =4,691	SEER registry	USA	Retrospective cohort study	NR Genomic health	18-30	100% HR+ 100% HER2- 40-85 years old Unclear if only those with O-DX test	1) LN0 2)LNmic-LN3	Treated according to usual practice with O-DX test 1) ET NR 23% CT 2) ET NR 35% CT

N, number of patient; CT, chemotherapy; ET, endocrine therapy; FFPE, formalin fixed paraffin embedded; HR+, hormone receptor positive; HER2, human epidermal growth factor receptor; ER+, oestrogen receptor positive; LN, lymph node; RS, Oncotype DX recurrence; mic, micrometastases; NR, not reported

[a] indicated for CT by NCCN guidelines; [d] HER2-negativity; pT1-T4c; LN+ [or LN0 with a risk factor (CpT2, grade 2/3, high uPA/PAI-1, <35 years, or HR-negative)][14] [e] patients were treated according to Oncotype DX score, with those with RS<12 receiving ET only, and those with RS≥12 receiving CT+ET;

Table 2: Quality assessment of clinical utility studies: Oncotype DX

	Random sequence generation	Allocation concealment	Blinding (participants and personnel)	Blinding (outcome assessment)	Incomplete outcome data	Selective reporting
TAILORx[1]	High	High	High	Low	High	Unclear
MD Anderson Le Du 2015[5]	High	High	High	Low	High	Unclear
WSG PlanB[2-4]	High	High	High	Low	High	Unclear
Clalit Health Services[7, 8, 10]	High	High	High	Low	High	Unclear
Memorial Sloan Kettering[6]	High	High	High	Low	High	Unclear
SEER registry[11, 12]	High	High	High	Low	High	Unclear
High/low/unclear relates to risk of bias on each criterion						

REFERENCES

1. Sparano JA, Gray RJ, Makower DF, Pritchard KI, Albain KS, Hayes DF*, et al.* Prospective validation of a 21-gene expression assay in breast cancer. *New Engl J Med* 2015;373:2005-14.
2. Gluz O, Nitz U, Christgen M, Kates R, Clemens M, Kraemer S*, et al.* Prospective WSG phase III PlanB trial: Clinical outcome at 5 year follow up and impact of 21 Gene Recurrence Score result, central/local-pathological review of grade, ER, PR and Ki67 in HR+/HER2-high risk node-negative and-positive breast cancer. *Eur J Cancer* 2016d;57:S6.
3. Gluz O, Nitz U, Chrlstgen M, Kates RE, Clemens M, Kraemer S*, et al.* Prognostic impact of 21 gene recurrence score, IHC4, and central grade in high-risk HR+/HER2-early breast cancer (EBC): 5-year results of the prospective Phase III WSG PlanB trial. *J Clin Oncol Conf* 2016a;34.
4. Gluz O, Nitz UA, Christgen M, Kates RE, Shak S, Clemens M*, et al.* West German Study Group Phase III PlanB trial: First prospective outcome data for the 21-gene recurrence score assay and concordance of prognostic markers by central and local pathology assessment. *J Clin Oncol* 2016b;34:2341-9.
5. Le Du F, Gonzalez-Angulo AM, Park M, Liu DD, Hortobagyi GN, Ueno NT. Effect of 21-gene RT-PCR assay on adjuvant therapy and outcomes in patients with stage I breast cancer. *Clin Breast Cancer* 2015;15:458-66.
6. Wen HY, Krystel-Whittemore M, Patil S, Pareja F, Bowser ZL, Dickler MN*, et al.* Breast carcinoma with an Oncotype Dx recurrence score <18: Rate of distant metastases in a large series with clinical follow-up. *Cancer* 2017;123:131-7.
7. Stemmer SM, Steiner M, Rizel S, Geffen D, Nisenbaum B, Peretz T*, et al.* First prospectively-designed outcome study in estrogen receptor (ER)+ breast cancer (BC) patients (pts) with N1mi or 1-3 positive nodes in whom treatment decisions in clinical practice incorporated the 21-gene recurrence score (RS) result. *Ann Oncol Conference: 41st European Society for Medical Oncology Congress,* ***ESMO*** 2016;27.
8. Stemmer SM, Steiner M, Rizel S, Soussan-Gutman L, Geffen DB, Nisenbaum B*, et al.* Real-life analysis evaluating 1594 N0/Nmic breast cancer patients for whom treatment decisions incorporated the 21-gene recurrence score result: 5-year KM estimate for breast cancer specific survival with recurrence score results <30 is >98%. *Cancer Research Conference: 38th Annual CTRC AACR San Antonio Breast Cancer Symposium San Antonio, TX United Stat* 2016;76.
9. Roberts MC, Miller DP, Shak S, Petkov VI. Breast cancer-specific survival in patients with lymph node-positive hormone receptor-positive invasive breast cancer and Oncotype DX Recurrence Score results in the SEER database. *Breast Cancer Res Treat* 2017;163:303-10.
10. Stemmer S, Steiner M, Rizel S, Soussan-Gutman L, Ben-Baruch N, Bareket-Samish A*, et al.* Clinical outcomes in early-stage breast cancer treated according to the recurrence score results: Evidence from a large prospectively-designed cohort study. In. Academic-In-Confidence; 2017.
11. Petkov VI, Miller DP, Howlader N, Gliner N, Howe W, Schussler NC*, et al.* Outcome disparities by age and 21-gene recurrence score (RS) in hormone receptor positive (HR+) breast cancer (BC). *J Clin Oncol Conf* 2016;34.
12. Roberts M, Petkov VI, Miller DP, Shak S, Howlader N, Cronin K*, et al.* Breast cancer specific survival in patients with node-positive hormone receptor positive invasive breast cancer and Oncotype DX recurrence score results in the SEER database. *J Clin Oncol Conf* 2016;34.
13. National Institute for Health and Care Excellence. Tumour profiling tests to guide adjuvant chemotherapy decisions in people with breast cancer (update of DG10) - final scope. London; 2017.
14. Nitz U, Gluz O, Christgen M, Kates RE, Clemens M, Malter W*, et al.* Reducing chemotherapy use in clinically high-risk, genomically low-risk pN0 and pN1 early breast cancer patients: five-year data from the prospective, randomised phase 3 West German Study Group (WSG) PlanB trial. *Breast Cancer Res Treat* 2017; 10.1007/s10549-017-4358-6.

15. Petkov VI, Miller DP, Howlader N, Gliner N, Howe W, Schussler N, *et al.* Breast-cancer-specific mortality in patients treated based on the 21-gene assay: a SEER population-based study. *NPJ Breast Cancer* 2016;2:16017.