Systematic review of the clinical effectiveness and cost-effectiveness of photodynamic diagnosis and urine biomarkers (FISH, ImmunoCyt, NMP22) and cytology for the detection and follow-up of bladder cancer

G Mowatt,[1]* S Zhu,[1] M Kilonzo,[2]
C Boachie,[1] C Fraser,[1] TRL Griffiths,[3]
J N'Dow,[4] G Nabi,[4] J Cook[1] and L Vale[1,2]

[1]Health Services Research Unit, Institute of Applied Health Sciences, University of Aberdeen, UK
[2]Health Economics Research Unit, Institute of Applied Health Sciences, University of Aberdeen, UK
[3]Department of Cancer Studies and Molecular Medicine, University of Leicester, UK
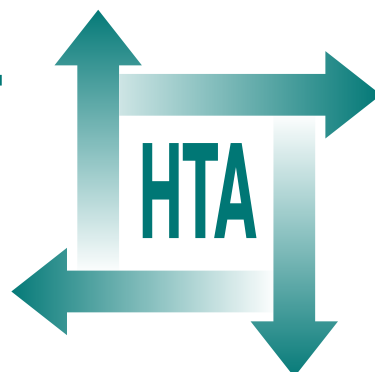[4]Academic Urology Unit, Department of Surgery, University of Aberdeen, UK

*Corresponding author

Executive summary

Health Technology Assessment
NIHR HTA programme
www.hta.ac.uk

Executive summary

Background

Bladder cancer is the sixth most common cancer in the UK, affecting more than 10,000 people each year. Around 75–85% of patients are diagnosed as having non-muscle-invasive disease, which, despite treatment, has a probability of recurrence at 5 years of 31% (95% CI 24% to 37%) to 78% (95% CI 73% to 84%). Inspection of the bladder [flexible cystoscopy using white light (CSC)] facilitated with local anaesthesia and voided urine cytology (involving the examination of cells in voided urine to detect the presence of cancerous cells) are currently the routine initial investigations of the bladder in patients with haematuria or other symptoms suggestive of bladder cancer. If CSC or urine cytology are suspicious, a rigid white light cystoscopy (WLC) under general or regional anaesthesia is performed with transurethral resection of bladder tumour (TURBT) where applicable. However, WLC may fail to detect some tumours. Photodynamic diagnosis (PDD) is a technique that could potentially be used to enhance tumour detection. Also, since the mid-1990s many urine biomarker tests for detecting bladder cancer have been developed, including fluorescence in situ hybridisation (FISH), ImmunoCyt and nuclear matrix protein (NMP22).

Objectives

This review aims to assess the clinical and cost-effectiveness of PDD compared with WLC, and urine biomarkers (FISH, ImmunoCyt, NMP22) and cytology for the detection and follow-up of bladder cancer.

Methods

Electronic searches were undertaken to identify published and unpublished reports. The databases searched included MEDLINE, MEDLINE In-Process, EMBASE, BIOSIS, Science Citation Index, Health Management Information Consortium (HMIC) and the Cochrane Controlled Trials Register as well as current research registers. The date of the last searches was April 2008. The types of studies considered for test performance were randomised controlled trials (RCTs), non-randomised comparative studies and diagnostic cross-sectional studies that reported the absolute numbers of true and false positives and negatives. Only RCTs were considered for studies reporting effectiveness. Participants had symptoms suspicious for bladder cancer or were previously diagnosed with non-muscle-invasive disease. The tests considered were (1) PDD compared with WLC or (2) FISH, ImmunoCyt, NMP22 or cytology, with a reference standard of histopathological examination of biopsied tissue.

One reviewer screened the titles and abstracts of all reports identified by the search strategy and data extracted included full-text studies, with checking by a second reviewer. Two reviewers independently assessed the quality of the diagnostic studies using a modified version of the QUADAS instrument and the quality of the effectiveness studies using a checklist adapted from Verhagen and colleagues.

The results of the individual studies were tabulated and sensitivity, specificity, positive and negative likelihood ratios, and diagnostic odds ratios (DORs) calculated. Separate summary receiver operating characteristic (SROC) curves were derived for different levels of analysis. Meta-analysis models were fitted using hierarchical summary receiver operating characteristic (HSROC) curves. Summary sensitivity, specificity, positive and negative likelihood ratios and DORs for each model were reported as median and 95% confidence interval (CI). For studies reporting effectiveness outcomes meta-analysis was employed to estimate a summary measure of effect, with dichotomous outcome data combined using relative risk (RR). Results were reported using a fixed-effect model in the absence of statistical heterogeneity.

An economic model was constructed to assess the cost-effectiveness of alternative diagnostic and follow-up strategies for the diagnosis and management of patients suspected of having bladder cancer. The model described care pathways from initial presentation, through diagnosis and treatment over a 20-year time horizon. A total of 26 different strategies were considered in the economic model, which represented plausible ways in which the tests might be used for the diagnosis

and follow-up of patients with bladder cancer. Of these 26, eight strategies that appeared to perform best in the deterministic analysis were further considered in a probabilistic analysis. The clinical effectiveness data from the systematic review (summarised below) were incorporated into the model. In the base-case analysis it was assumed that the underlying risk of disease within the target population was 5%. Costs for treatments and interventions with strategies were derived from the literature review in the UK setting, in particular NHS resources. The mean cost per test for PDD was £1371, WLC £937, CSC £441, cytology £92, NMP22 £39, ImmunoCyt £54 and FISH £55. TURBT cost from £2002 to £2436 depending upon whether it was assisted by WLC or PDD respectively. Additional subsequent treatments were also included, which were based upon those typically adopted within the UK NHS. A cost–utility analysis was not possible as part of the base-case analysis because of a lack of relevant utility data. Hence, cost-effectiveness (life-years, cases of true positives) and cost–consequence analyses were conducted. Sensitivity analyses were conducted to assess the uncertainties in estimates and assumptions.

Results

A total of 27 studies enrolling 2949 participants reported PDD test performance. In the pooled estimates for patient-level analysis, based on direct evidence, PDD had higher sensitivity than WLC (92%, 95% CI 80% to 100% versus 71%, 95% CI 49% to 93%) but lower specificity (57%, 95% CI 36% to 79% versus 72%, 95% CI 47% to 96%). In the pooled estimates for biopsy-level analysis, based on direct evidence, PDD also had higher sensitivity than WLC (93%, 95% CI 90% to 96% versus 65%, 95% CI 55% to 74%) but lower specificity (60%, 95% CI 49% to 71% versus 81%, 95% CI 73% to 90%).

Across studies, the median sensitivities (range) of PDD and WLC for detecting lower risk, less aggressive tumours were broadly similar for patient-level detection [92% (20% to 95%) versus 95% (8% to 100%)], but sensitivity was higher for PDD than for WLC for biopsy-level detection [96% (88% to 100%) versus 88% (74% to 100%)]. However, for the detection of more aggressive, higher risk tumours the median sensitivity of PDD for both patient-level [89% (6% to 100%)] and biopsy-level [99% (54% to 100%)] detection was higher than those of WLC [56% (0% to

100%) and 67% (0% to 100%) respectively]. The superior sensitivity of PDD was also reflected in the detection of carcinoma in situ (CIS) alone, both for patient-level [83% (41% to 100%) versus 32% (0% to 83%)] and biopsy-level [86% (54% to 100%) versus 50% (0% to 68%)] detection.

Four RCTs enrolling 709 participants comparing PDD with WLC reported effectiveness outcomes. The use of PDD at TURBT resulted in fewer residual tumours at check cystoscopy (pooled estimate RR 0.37, 95% CI 0.20 to 0.69) and longer recurrence-free survival (pooled estimate RR 1.37, 95% CI 1.18 to 1.59) compared with WLC. The advantages of PDD at TURBT in reducing tumour recurrence (pooled estimate RR 0.64, 95% CI 0.39 to 1.06) and progression (pooled estimate RR 0.57, 95% CI 0.22 to 1.46) in the longer term were less clear.

A total of 71 studies reported the performance of biomarkers (FISH, ImmunoCyt, NMP22) and cytology in detecting bladder cancer. In total, 14 studies enrolling 3321 participants reported on FISH, 10 studies enrolling 4199 participants reported on ImmunoCyt, 41 studies enrolling 13,885 participants reported on NMP22 and 56 studies enrolling 22,260 participants reported on cytology. In the pooled estimates, based on indirect evidence, sensitivity was highest for ImmunoCyt and lowest for cytology. FISH (76%, 95% CI 65% to 84%), ImmunoCyt (84%, 95% CI 77% to 91%) and NMP22 (68%, 95% CI 62% to 74%) all had higher sensitivity than cytology (44%, 95% CI 38% to 51%). However, cytology had higher specificity (96%, 95% CI 94% to 98%) than FISH (85%, 95% CI 78% to 92%), ImmunoCyt (75%, 95% CI 68% to 83%) or NMP22 (79%, 95% CI 74% to 84%).

Cost-effectiveness

Although the differences in outcomes and costs between these strategies appear to be small, the decision about which strategy to adopt depends upon society's willingness to pay for additional gain. The most effective strategy in terms of true positive cases (44) and life-years (11.66) was a strategy of CSC and ImmunoCyt followed by PDD in initial diagnosis and CSC followed by WLC in follow-up. This strategy had, however, an incremental cost per life-year of over £270,000. The least effective strategy was cytology followed by WLC in initial diagnosis and follow-up (total average cost over 20 years = £1403 and average life expectancy = 11.59). This strategy was most likely to be considered cost-effective when society's

willingness to pay was less than £20,000 per life-year. Over most of the ranges of willingness to pay values there appeared to be no strategy that would have a likelihood of being cost-effective more than 50% of the time, but four of the eight strategies included in the probabilistic sensitivity analysis were each associated with an approximately 20% chance of being considered cost-effective. Three of these four strategies involved the use of a biomarker or PDD.

Sensitivity analyses

The sensitivity analyses indicated that the order of the least to the most costly strategies remained the same when discount rates, RR rates and performance of CSC were changed. The results were most sensitive to the pretest probability of disease (5% in the base case). At a 1% probability it is most likely that the least costly (and least effective) strategy of cytology followed by WLC for both diagnosis and follow-up would be cost-effective. At a 20% prevalence the more effective strategies (in terms of diagnostic performance) are more likely to be worth their increased cost.

Discussion

PDD has higher sensitivity (fewer false negatives) than WLC and so will detect cases of bladder cancer missed by WLC, but its lower specificity will result in more false positives. The advantages of PDD's higher sensitivity in detecting bladder cancer overall, and also more aggressive, higher risk tumours, have to be weighed against the disadvantages of a higher false-positive rate, which leads to additional, unnecessary biopsies of normal tissue being taken and potentially additional unnecessary investigations being carried out and the resulting anxiety caused to patients and their families.

In the four studies reporting effectiveness outcomes, such as tumour recurrence, the administration of single-dose adjuvant chemotherapy following TURBT, which can reduce recurrence rates by up to 50% in the first 2 years, varied, making it difficult to assess what the true added value of PDD might be in reducing recurrence rates in routine practice.

Based on indirect comparisons, all three biomarkers had higher sensitivity, but lower specificity, than cytology in detecting bladder cancer. A urine biomarker test such as ImmunoCyt could potentially replace some cytology tests if higher sensitivity (fewer false negatives) is considered more important than higher specificity (fewer false positives). However, if higher specificity is considered more important then cytology would remain the test of choice.

Linking diagnostic performance to long-term outcomes required a number of assumptions to be made about the structure of the economic model and its parameters. Some assumptions were based on non-UK study data; it is unclear whether such data are applicable to the UK setting. One assumption concerned starting age and the length of time over which the benefits from a diagnostic strategy may accrue. In the base-case analysis a time period of 20 years and starting age of 67 years were used, although the impact of shorter time horizons and older starting age were explored in the sensitivity analyses. When either the time horizon was reduced or the starting age was increased, the incremental cost per life-year increased as the costs of initial diagnosis and treatments were not offset by survival and life-year gains.

Conclusions

Implications for service provision

PDD has higher sensitivity than WLC in detecting bladder cancer and is better at detecting more aggressive, higher risk tumours, including CIS, but has lower specificity. Based on limited evidence, the use of PDD at TURBT compared with WLC results in fewer residual tumours at check cystoscopy and longer recurrence-free survival, whereas the advantages of PDD at TURBT in reducing tumour recurrence and progression in the longer term are less clear. In the pooled estimates ImmunoCyt had the highest sensitivity and cytology had the highest specificity, with all three biomarkers having higher sensitivity, but lower specificity, than cytology.

Taking into account the assumptions made in the model, the strategy of CSC and ImmunoCyt followed by PDD in initial diagnosis and CSC followed by WLC in follow-up is likely to be the most costly and the most effective (£2370 per patient and 11.66 life-years). There appeared to be no strategy that would have a likelihood of being cost-effective more than 50% of the time over most of the ranges of willingness to pay values. Nevertheless, strategies involving biomarkers and/or PDD provide additional benefits at a cost that society might be willing to pay. Strategies involving

cytology are unlikely to be considered worthwhile. Strategies that replaced WLC with PDD provided more life-years but it is less clear whether they would be worth the extra cost.

Recommendations for research

Further research is required in the following areas:

- RCTs including economic evaluations comparing PDD with rigid WLC at TURBT plus adjuvant immediate single-dose intravesical chemotherapy in patients diagnosed with bladder tumours at CSC.
- Diagnostic cross-sectional studies comparing FISH with ImmunoCyt, NMP22 BladderChek point of care test and voided urine cytology within the setting of the British Association of Urological Surgeons and the Renal Association diagnostic algorithm for the diagnosis of patients with haematuria. Data produced should be incorporated into an economic evaluation.
- Studies to collect health state utilities are needed. These may come from further prospective studies or as part of future RCTs.

- The trade-off between process of care and short-term (diagnostic outcomes) and longer-term outcomes needs to be explored using recognised preference elicitation methodology in a way that can be incorporated into future economic evaluations.
- The impact that an incorrect diagnosis (false-negative result) has on patients either at diagnosis or at follow-up in terms of future survival, quality of life and costs.

Publication

NIHR Health Technology Assessment programme

The Health Technology Assessment (HTA) programme, part of the National Institute for Health Research (NIHR), was set up in 1993. It produces high-quality research information on the effectiveness, costs and broader impact of health technologies for those who use, manage and provide care in the NHS. 'Health technologies' are broadly defined as all interventions used to promote health, prevent and treat disease, and improve rehabilitation and long-term care.

The research findings from the HTA programme directly influence decision-making bodies such as the National Institute for Health and Clinical Excellence (NICE) and the National Screening Committee (NSC). HTA findings also help to improve the quality of clinical practice in the NHS indirectly in that they form a key component of the 'National Knowledge Service'.

The HTA programme is needs led in that it fills gaps in the evidence needed by the NHS. There are three routes to the start of projects.

First is the commissioned route. Suggestions for research are actively sought from people working in the NHS, from the public and consumer groups and from professional bodies such as royal colleges and NHS trusts. These suggestions are carefully prioritised by panels of independent experts (including NHS service users). The HTA programme then commissions the research by competitive tender.

Second, the HTA programme provides grants for clinical trials for researchers who identify research questions. These are assessed for importance to patients and the NHS, and scientific rigour.

Third, through its Technology Assessment Report (TAR) call-off contract, the HTA programme commissions bespoke reports, principally for NICE, but also for other policy-makers. TARs bring together evidence on the value of specific technologies.

Some HTA research projects, including TARs, may take only months, others need several years. They can cost from as little as £40,000 to over £1 million, and may involve synthesising existing evidence, undertaking a trial, or other research collecting new data to answer a research problem.

The final reports from HTA projects are peer reviewed by a number of independent expert referees before publication in the widely read journal series *Health Technology Assessment*.

> **Criteria for inclusion in the HTA journal series**
>
> Reports are published in the HTA journal series if (1) they have resulted from work for the HTA programme, and (2) they are of a sufficiently high scientific quality as assessed by the referees and editors.
>
> Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search, appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.