The economic evaluation of diagnostic technologies for Health Technology Assessment

An illustration using Positron Emission Tomography

Dr. Andrew Walker, University of Glasgow

Paper prepared for the Health Economics Study Group meeting, UEA Norwich, January 2002.

To illustrate points, this paper includes preliminary results that will change; please do not quote them.

Acknowledgement
This paper is based on work carried out on behalf of Health Technology Board for Scotland. HTBS encouraged and supported the preparation of this paper but does not necessarily agree with the views expressed.

Introduction

Every application of the principles of economic evaluation involves fresh challenges, whether these are in terms of defining the benefit measure, handling uncertainty, or whatever. Checklists of (generally agreed) 'good practice' points are carefully worded to be flexible enough to cope with this variation from one application to another and, in general, economists do not see the need to change or supplement them.

The Health Technology Board for Scotland (HTBS) is currently involved in a Health Technology Assessment (HTA) of Positron Emission Tomography (PET), a new type of diagnostic imaging technology. Several recent HTAs have been carried out in this field, in Australia and Denmark amongst others. However, the economic focus has been on reviewing published cost-effectiveness evaluations of PET. The HTBS economic evaluation is trying to go beyond this to provide 'value-added' to the literature rather than summarising it. To date this has focused on the use of PET in the staging of non-small cell lung cancer (NSCLC) as a prelude to possible surgical intervention. The model structure is based on the literature but modified in the light of discussions with Scottish clinicians. The aim is then to apply Scottish data on costs, case-mix and diagnostic protocols. Data on sensitivity and specificity will come from a new meta-analysis of the clinical literature.

The economic evaluation has identified a number of unanticipated issues. Some of these are specific to the evaluation of a diagnostic imaging technology, while others have more general application within HTA. The aim of this paper is to highlight, discuss and try to resolve these issues. This will be of interest to those involved in HTA in general and those considering a diagnostic imaging technology in particular.

A brief account of the economic evaluation of PET

What is PET?

From the patient's point-of-view, PET is another type of scanner, not dissimilar from CT or MRI. However, the fact that a PET scanner is typically based in a nuclear medicine department rather than in radiology is indicative of a fundamentally different technology. CT and MRI provide images of the anatomical detail of the body, but while these are valuable in planning treatment they are not always sufficient. In the case of non-small cell lung cancer (NSCLC), for example, a CT scan usually gives a good picture of the size and site of the primary cancer. However, it is less accurate at detecting spread to lymph nodes or metastatic disease.

PET requires injection of specially produced radio-chemicals that emit positrons into the body. Uptake of the radiochemical varies among different cell types. Malignant tumours have high uptake and imaging can then be used for the detection and assessment of changes in physiology and metabolism. For example, the PET scan has the potential to show differences between healthy tissue and lymph nodes with malignant spread. The chemical used is a radiolabelled version of glucose but because it is short-lived (half-life of two hours) it is thought to have no adverse side effects.

Like CT and MRI, PET is a general imaging technology with (potentially) very broad applications. Most attention to date has focused upon cancer, heart disease and neurological symptoms, but even within these headings there might be a number of

different applications. PET is widely available in America with more than 75 centres. In Europe, Germany has led the way with about 40 scanners. In Britain there are two scanners available for NHS use, both located in London. An assessment of PET is timely because the NHS appears to have a low level of provision compared to some other countries but the technology has not yet diffused.

What was the remit of the HTA?

The Health Technology Board for Scotland was established in 2000 to provide NHS Scotland with high-quality information on the clinical effectiveness and cost-effectiveness of health technologies. HTBS is also a member of international HTA networks and has this broader community as a secondary audience for its work. One of the first three HTAs it undertook was an assessment of the role of PET imaging in cancer management.

The early stages of the HTA involved trying to define a list of potential applications within this field to allow literature searches to be carried out, experts to be consulted and so on. One of the problems was that the technology is still at quite an early stage of its development; there are a few indications where a lot of accuracy data have been reported and others where there are only a few studies.

The recent Australian report (Medical Standards Advisory Committee (MSAC (2000)) covered six potential applications of PET, devoting 3-4 pages of its report to each. However, HTBS's preliminary literature searches indicated a lot of evidence related to the accuracy of PET compared to existing technologies but little on the impact that this might have on the patient's subsequent management. Further, while the published economic evaluations in this field were wholly supportive of PET, preliminary analysis suggested various flaws that might undermine the results.

The decision was taken to focus upon one application of PET and the example chosen was staging of NSCLC. This is an important health issue in Scotland, where audit suggests quality of care could be improved (Gregor et al (2001)). PET is widely agreed to have a potential role and there is a substantial body of clinical evidence as well as three published economic evaluations. NSCLC has a greatly improved survival rate when surgery is possible, so accurate detection can add several years to a new patient's life expectancy. In addition, the NCCHTA review of research priorities in this area ranked this potential role of PET as the main priority (Robert and Milne (1999)).

Staging of NSCLC

Patients referred by their GP with suspected NSCLC have generally had a chest X-ray. The treatment with the highest survival rate is surgery, so identifying suitable patients is a priority. This involves (i) establishing whether the primary tumour is resectable, (ii) ruling out involvement of lymph nodes, (iii) ruling out patients who secondary metastatic disease or who are unfit for surgery. In terms of TNM staging T is the primary tumour size, N is the involvement of nodes and M is distant disease. Ordinal numbers indicate the extent of disease, hence T1N0M0 is a resectable tumour, whereas T3N3M1 indicates advanced disease.

The initial investigation is by bronchoscopy (an endoscopic investigation of the bronchus and trachea.) In about 60% of cases a biopsy of cytology sample is possible

that allows a definite diagnosis to be made. Fine-needle biopsy can give a definite diagnosis in a further 20-30%. Of the remaining 10-20% a few might be referred for surgery on the basis of a suspicious CXR showing a solitary tumour strongly suggestive of cancer. Most patients who do not have obvious advanced disease will go on for further investigation, however.

Resectability depends on size, extension to ribs and pleura, and the extent of nodal disease. CT of the lungs and abdomen (notably the liver and adrenal glands) is a reliable guide to tumour size. It also shows the *size* of lymph nodes, a factor that has some prognostic value. (It is common practice to regard lymph nodes that are more than 1cm in diameter as 'enlarged' in which case the probability of spread to that node is around 60%. Lymph nodes of less than 1cm have a probability of spread of around 2%.) CT will also show obvious metastases beyond the lymph nodes in the area scanned.

Those with lymph nodes that are not enlarged are candidates for surgery (subject to fitness). For those with enlarged lymph nodes, mediastinoscopy is the next line of investigation; this is an endoscopic investigation carried out external to the bronchus, accessed by making a small incision in the chest close to the throat. This investigation is (close to) definitive for lymph nodes that can be reached but in about a quarter of cases with affected lymph nodes mediastinoscopy cannot detect disease. Where there is no suggestion of lymph node involvement, patients are candidates for surgery. Where lymph node disease is detected, surgery is generally ruled out and the patient is referred to an oncologist.

Screening out patients with metastatic disease or who are unfit for surgery takes place throughout this process. Only patients with symptoms that are indicative of secondary disease will be investigated further e.g. bone scan for skeletal metastases.

How might PET help? There are four possible roles:

(i) investigate lymph nodes out of the reach of mediastinoscopy after that procedure has been employed, or
(ii) to replace mediastinoscopy as the investigation of enlarged lymph nodes, or
(iii) to investigate the 10% of patients in whom bronchoscopy / fine needle biopsy do not give a definite diagnosis, or
(iv) as an additional procedure in investigation of distant metastases.

PET is unlikely to replace CT: the anatomical detail the latter offers is valued by clinicians.

What the evidence says

As an example of the type of clinical evidence available, a meta-analysis published in 1999 suggested the sensitivity of CT for affected lymph nodes was 60% whereas for PET it was 79% (Dwamena et al (1999)). The specificity for CT was 77% and for PET 91%. In each case the 95% confidence interval around the mean estimate was +/- 2%, suggesting the differences were unlikely to be due to chance.

Various limitations apply. The results of meta-analysis can be questioned in various ways. There is also a particular danger of publication bias because the review only considered studies included in Medline. Investigators may not write up their results if

they are not as good as those in the literature: they risk their reputation with peers as well as a loss of custom (if they operate in a market, as in America). Similarly, journal editors may be more interested in something that takes the technology to a new plane rather than a mildly negative study. Finally, the data cause concern to economic modellers because they take no account of the preceding diagnostic information available - for example, what is the sensitivity of PET following a negative CT?

The other issue is how this information is used to change patient management. As an example, Saunders et al (1999) reported on 97 NSCLC patients being considered for surgery. They report a change in management in 36 (37%), as follows (with my comment in italics):

- in 6 cases further investigation was required - *the significance of this is unclear: presumably there were no other changes in management?*
- in 15 a planned operation was cancelled because of previously unsuspected lymph node involvement
- in 11 an operation was enabled by excluding mediastinal spread - *in Scotland, this group would be operated upon anyway*
- in 4 an operation was enabled as a diagnosis was established - *ditto*

In the Scottish context, therefore, the evidence suggests 15 patients in whom a planned operation was cancelled per 97 scanned, a rate of 1 cancelled operation per 6.5 scans. This rate is similar to that found in one of the two RCTs of the impact of PET on the surgery rate (van Tinteren et al (2001)). While the results are only available in abstract form to date the reduction in surgery reported is from 41% to 21% i.e. 1 operation averted per 5 scans.

The previous economic evaluations are all supportive of PET. They conclude:

- Gambhir et al (1996) - CT-plus-PET is more cost-effective than CT alone
- Scott et al (1998) - PET is most cost-effective when used in patients with a CT scan that shows normal sized lymph nodes (US$25,286 per life-year gained)
- Dietlein et al (2000) - "the implementation of whole-body PET with a full ring of detectors in the preoperative staging of patients with NSCLC and normal-sized lymph nodes is clearly cost-effective. However, patients with nodal-positive PET results should not be excluded from biopsy." (page 1598)

In addition, Robert and Milne (1999) quote an American HTA carried out by the Emergency Care Research Institute (ECRI), which showed that PET is only cost-effective following a negative CT, and is not cost-effective earlier in the diagnostic process. However, this report has not been published in a peer-reviewed journal and is not routinely available without a subscription to the company's database.

The economic analysis

A fairly simple model was assembled to draw together data on the following elements:

- epidemiological data on the underlying ('true') staging distribution in the population who may be eligible for PET;
- sensitivity, specificity, side-effects and cost for each type of diagnostic test; and
- protocols for how tests will be used in different scenarios as well as the way the results will be used and the impact on health.

This work is still in progress, so the results quoted later in this paper are based on values in the existing literature and will change when the new data are available.

Of the published economic studies Dietlein et al appears to have used the most sophisticated strategy for identifying and analysing the evidence. For example, they estimate separate sensitivity and specificity figures for PET according to whether lymph nodes were normal or enlarged on CT.

In other areas, Scottish data have been used, notably in defining current practice and estimating costs. The primary source here was the publication "Scottish Health Service Costs" (Information and Statistics Division (2000)), which lists hospital costs by specialty. It includes cost per in-patient cases (including mean length-of-stay, allowing a cost per day to be estimated), day cases, out-patient attendances, as well as costs for support services such as theatre time, selected radiology investigations, laboratory tests, PAM consultations, and so on. For example, the median stay for a lung resection is 8 days at the Western General in Glasgow. The Scottish average costs for an in-patient cardio-thoracic case is £3,875 for a 6.9-day stay. Increasing this pro rata to an 8-day stay gives a figure of £4,493.

The cost of PET was taken from one of the two existing centres in the UK, at St. Thomas's Hospital, London. The figure the centre quotes for NHS work is £784 for a body scan and £873 for a complex scan. A draft report from the Intercollegiate Standing Committee on Nuclear Medicine includes resource requirements to set-up and run a PET scanner (ICSCNM (2001)). Valuing this resource use using salary scales and including a capital charge (assumed life 10 years) gives a cost of £1,035 per scan if throughput is 1,000 patients per year.

Results

The detailed results of the analysis are not yet available. However, it is apparent even from the simple scans-to-saved surgery ratios quoted above that the balance is a very fine one. The ratio of the cost of a PET scan to the cost of an operation, 1:5.7, falls in between the scan-to-averted-operation ratio of Saunders et al (1: 6.5) and the van Tinteren et al RCT (1:5).

Of course, the position is not quite so clear-cut: for example, it takes no account of any investigations saved (possibly some mediastinoscopies, which cost around £431). It also assumes that the operations averted would have been typical - in practice, they could be quite quick if there is evident spread of the disease (so-called 'open-and-shut' surgery). Finally, based on their literature review Dietlein et al found that PET has a 76% specificity rate when CT shows enlarged nodes: while this sounds relatively innocuous, it means that 24% of people with potentially resectable cancers are falsely classified as having nodal involvement and are thus sent for radiotherapy instead. From Dietlein et al's own assumptions, this reduces the potential life expectancy of each of these patients by almost 2 years.

The published economic evaluation results indicate that PET is most cost-effective when confined to patients with a normal CT. In the HTBS model (based on Dietlein et al's literature review), PET's specificity is much higher (96%) but its sensitivity is lower (74%). The comparable figures for mediastinoscopy are 100% and 72%, suggesting PET's advantage is very slight. In the HTBS model the PET strategy

actually involves a net *loss* of life-years. It should be emphasised, however, that even in the economic evaluations that are supportive of PET, the gains in life expectancy are very small. Scott et al, for example, found that PET following a normal CT in creases life expectancy by 2.6 days compared to the status quo. Clearly, it would only take a very small change in assumptions to make the denominator negative and the conclusion would be very different.

The tentative model compares the diagnostic protocol listed above to using PET after CT irrespective of node size, with doctors assumed to act on the results of PET. (CT does not guide patient management but it is still valued for the anatomical details provided, whether for surgery or for planning radiotherapy.) For a cohort of 100 patients the model predicts a reduction in surgery from 87 to 66 procedures. Out of this reduction, 5.7 of the 21 are false positives. Ironically, the users of PET would not know that this was the case - they would simply see a reduction in surgery in line with van Tinteren et al's findings.

As has been emphasised these results are based on the existing literature and will change when data from the HTBS work becomes available. All that can be concluded at present is that the supportive economics literature should not necessarily be taken at face value.

Possible issues for discussion

There are at least eight issues that arise from this HTA and economic evaluation:

- the breadth of the comparison
- the value of searching the literature
- costs - average or marginal?
- savings and the value of resource use
- what is the definition of benefit and how can this be put into practice?
- the limits of modelling in the face of limited clinical evidence
- what to do when the clinical evidence runs out
- what level of detail is required of the economic evaluation?

These are discussed in turn below. Each section concludes with an attempt to summarise the key issues; this should not restrict the discussion, however.

1. Breadth of comparison

Diagnostic technologies have many potential applications, so HTA exercises tend to be very broad: for example, the Australian HTA (MSAC) considered the role of PET in general while HTBS considered the application to cancer. As noted above, even the latter does not give the economist a 'ready-made' question to tackle - at one point HTBS had identified 12 potential applications to study, each of which might have required an economic evaluation.

There is a danger, however, of the economic study being driven by the clinical research. The main thing decision-makers ask of a new technology in terms of the clinical evidence is whether it is more effective than the existing service and whether there are safety issues. As a result, the review searches the literature for evidence comparing the new technology to the status quo in order to create a series of data extraction tables are created. In this case, the search might be confined to PET versus CT (or even mediastinoscopy).

The problem with basing the economic evaluation on these tables is that this does not necessarily reflect the economic efficiency question that decision-makers want answered. Assuming that in some uses PET does show advantages relative to CT albeit at increased cost, the question a decision-maker asks might be:

- "But would I get more benefit from investing in other ways to reduce unnecessary operations for NSCLC?" Another option would be more routine brain scans to look for metastatic disease.
- "But would I get more benefit from investing in other ways to improve diagnostic investigation of NSCLC?" For example, the £1 million that PET costs each year (at a minimum) could fund two CT scanners to reduce waiting times. In Glasgow even an urgent scan takes 2-3 weeks, with non-urgent scans taking 12 weeks.
- "But would I get more benefit from investing in other NSCLC services or other cancer services?" For example, the recent Scottish Cancer Investment Plan (Scottish Executive (2001a)) sets out proposals for spending an additional £10 million of earmarked funding. The schemes included could be seen as those at the (economic) margin - they thus accurately reflect the opportunity cost of investing in PET.
- "But would I get more benefit from investing in other health services?" NICE have found it acceptable to make judgements that have been informed by net cost per QALY gained data.

The economic evaluation outlined estimates the net costs and the net QALY impact of PET compared to the status quo, but certain assumptions are required: for example, the PET cost is based on an assumed throughput of 1,000 patients. In fact, there are about 3500 new cases of NSCLC in Scotland each year (Scottish Executive (2001b)), so no more than 500 or so would be eligible for PET (and some of these may be unable or unwilling to travel to the scanner). This means that the economist must either find another cost-effective use for the 600+ spaces on the scanner or increase the cost of scans (assuming average costs are appropriate, an issue to be discussed later in this paper).

A possible alternative approach to the economic appraisal would be to define and address the issues that health service decision-makers face, including:

- are inputs divisible, for example, through purchasing a PET scanner in conjunction with other purchasers?
- if not (or the 'purchaser' is a region in the first place), then what is the capacity of the scanner in terms of patient throughput?
- in terms of economic efficiency, which patients should have priority access to this capacity?
- given the economically optimal case-mix, how does the estimated total benefit compare with alternative uses of the resources?
- will any capacity be freed up elsewhere in the system as a result of the introduction of PET? what benefits will this yield in their alternative use?

This has similarities to the approach of Sassi et al (1997), who suggest identifying the optimal use of the pool of resources involved in the diagnostic process, not a single application (p.624). The problem, of course, is that this requires a far bigger evaluation - in the PET case, the economist would have to go back to the 12 questions mentioned above. However, the quality of the clinical evidence becomes progressively weaker and hence modelling becomes more speculative. To address the

question of whether purchasing the scanner and staffing it (a fairly indivisible input, unless staff are part-time), we would then have to ask a benefit maximisation question. In other words, PET is an all-or-nothing investment of £1 million; what are the alternative benefits that can be assembled for different 'packages' of services? This has some similarities to the method for economic evaluation proposed by Sutton (1997).

Given that decision-making in practice is fairly crude (and is certainly time limited) it might be more practical to try to identify a plausible package of services that represent the opportunity cost of £1 million for each of the decision-makers questions above. The investments in the Scottish Plan are specified in terms of inputs and short-term outcomes; to make true comparisons, however, quantified long-term outcomes would be required.

Issues from this section

Is HTA focussing on an abstract question from the point-of-view of informing the final decision? What can be done about this? Is this within the control of the economist? If not, what arguments should be advanced to those who draw up the remit for such exercises?

2. The value of the economics literature search

The three published economic evaluations identified were all supportive of PET imaging in staging NSCLC. Several reviews of the clinical evidence have simply quoted these results in passing and concluded that PET is cost-effective. A closer examination of the data suggests the issue is far less clear-cut, although it took a careful examination of the assumptions and reworking of the models to establish this. This finding has several implications:

- first, in a time-limited economic evaluation, carrying out a review of the economic evaluation literature might not be cost-effective. The time spent devising search strategies, reading abstracts, ordering and reading papers is non-trivial and has alternative valuable uses. The results can be questioned anyway on grounds of international transferability but the work above also suggests that reworking with other data may change the conclusions. The models provide helpful frameworks, but the analysis *per se* contributed little.

- second, it might be that economic evaluation alongside HTA needs to establish a different approach. If this is how unreliable the economic evaluation literature really is then economists should always be aiming to carry out an original economic evaluation, even if it is in the form of a relatively simple and transparent spreadsheet model. Having used this framework to identify the main benefits and resource use, the literature search should be used to find evidence for these variables. For example, the estimation of QALY gain from PET scanning requires utilities for NSCLC after resection and after radiotherapy - these might be available from sources other than evaluations of PET. Similar comments apply to the model of the natural history of NSCLC over time and to epidemiological data (such as the prevalence of lymph node involvement). Little thought has been given to how to search most efficiently for such studies.

- third, the finding that the 'true' cost-effectiveness might be at odds with the published literature has implications for groups that rely on the published literature such as Cochrane or the Scottish Intercollegiate Guidelines Network (SIGN). Sometimes, groups interested in systematic reviews want to subject economic evidence to the same types of meta-analysis that are used for clinical work. If this had been undertaken in this example then a synthesis of the three published studies might have given a misleading result.

Issues from this section

This is N=1, but if others have similar experiences then do we need to start rethinking how we do our searches? I am not suggesting we abandon the review of previous attempts by economists to answer the HTA question, but in the context of building the model these may actually play quite a small role.

3. Average or marginal costs

Every textbook on economic evaluation will say that marginal costs are the appropriate concept for valuing resource use to compare to marginal benefits; average costs should not be used as they do not reflect the value of resources used in the production of the last unit. However, this rests upon the assumption that we are dealing with small changes in programmes - for example, is it cost-effective to invest a little more in cardiac surgery and a little less in hip replacements?

In many cases, however, economists use the average cost, sometimes citing the defence that in the long run all costs are variable so marginal costs equal long-run average costs. The HTA of PET seems slightly different. From a financial perspective, the scanner, the centre and its staff are close to an indivisible input: almost all of the resource must be in place for any scans to take place and after that the variable financial cost per scan is relatively small. From an economic point-of-view, however, staff time used in the process would be measured and valued. Such a 'bottom up' approach to costing would only include direct resource use, however. There are many vital items of resource use that the patient does not directly consume but without which their scan could not take place. Maintenance is an obvious example, as is quality control in the laboratory that produces the chemicals to be injected; cleaning, portering, the cost of the capital used, and so on, all play a role. Whether hospital overheads should also be attributed to an individual scan is a moot point (Wagner (1983)).

For all the words about marginal costs, therefore, the average cost seems more appropriate. The problem with this is that the cost per scan then varies with the throughput. If throughput is 1,000 patients per annum the cost is £1,035. If throughput goes down then the average cost rises (ceteris paribus), even if the time taken per scan is identical before and after. This can be of crucial importance in determining the cost-effectiveness of PET in staging NSCLC. One of the published economic studies showed this to be a key factor in the sensitivity analysis (Scott et al) - for PET to dominate CT, the cost must be less than US$1,700, whereas in the baseline analysis it was US$2,000.

Issues from this section

It would be going too far to say that costs are meaningless but the true significance is certainly quite elusive. Are there circumstances under which average cost is acceptable? Or do we just become accountants if we do this?

4. Savings and values

The method of calculating costs described above has some advantages over the healthcare resource group (HRG) system used in England and drawn together in the NHS Reference Costs (Department of Health (2000)). It is much more flexible to estimating costs based on clinicians' estimates of resource use for specific groups of patients. It is also more transparent because the assumptions about resource use and the elements of the costs used to value that resource use are quite apparent.

Under an HRG system, for example, a lung resection must be matched to a heading and cost. For example, in the NHS Reference Costs it could be matched to:

HRG code	Cost per case
D002 Complex thoracic procedure	£4,143
D003 Major thoracic procedure	£2,471
D005 Intermediate thoracic procedure	£1,901

However, this disguises what measurement and valuation of resource use has gone into the figures. It is also likely that average costs are still being used, even if the measurement of resource use is procedure-specific.

Under the system used by HTBS it is clear the median stay is 8 days and that we are assuming these are spent on a cardio-thoracic ward. The downside of transparency, of course, is that it leads to further questions. For example, Scottish Health Service Costs lists hospitals as having a dedicated cardio-thoracic ward; presumably, surgery undertaken in other hospitals takes place on a general surgical ward. The cost per day on a cardio-thoracic ward is £562 whereas on a general surgical ward it is £297, so over an 8-day stay the patient who is treated on a cardio-thoracic ward 'costs' £2,120 less than if they had been managed in general surgery! Similarly, mediastinoscopy can often be carried out as a day case - if this is administered through cardio-thoracic surgery it 'costs' £584 whereas under general surgery it 'costs' £300. The former figure is roughly £3 for every minute that the patient is in the hospital!

The answer, of course, is that the cost returns in Scottish Health Service Costs represent averages for a specialty. Cardio-thoracic surgery also includes cardiac surgery and even transplants. At the Western Infirmary in Glasgow, for example, lung resections make up less than 10% of the workload and they appear to be some of the least complex procedures undertaken. It is also notable that ISD, the collators of these costs, advise that they should not be used in the way described.

So one message is that while such costs are flexible they should also be handled with care. There is also a broader issue about the value that local health services place on resources. The figures in Scottish Health Service Costs are being used as proxies for opportunity costs, hence they can be seen as 'exchange rates' between different types of activity and inputs. For example, an hour in theatre costs (is valued at) £480, a day on a general surgical ward is valued at £297, and the resource use involved in

processing a haematology specimen is valued at £4.74. This means the exchange rate is one hour in theatre is equivalent value to 1.6 general surgical bed-days, which, in turn, is equivalent to 101 haematology tests. This assumption is rarely tested in practice - what economists really need to know are the true 'spot market' prices for these resources. For example, salary scales might tell us that an experienced nurses time costs £20 per hour but if it is very difficult to recruit one then the effective value on an hour of their time is far higher.

This is more than just a curiosity. In the costing section above it was noted that one surgical resection costs the same as 5.7 PET scans. Based on an 8-day stay for an operation, the resources devoted to every PET scan are assumed to be equivalent in value to 1.4 cardio-thoracic bed-days. In fact, PET frees up bed-days at a rate of 1.2 per scan (based on 6.5 scans to save an operation, 8 days per operation). If this is the case then, if all else is equal, PET scanning is not justified. But does this really reflect the value judgement of decision-makers? Waiting times for cardiac surgery are a problem so it could be argued that cardio-thoracic bed-days are 'at a premium' - in other words, the PET scan : bed-day exchange rate is higher than 1:1.4.

Issues from this section

It is common practice for economists to check the values of clinical variables to be used in their models with doctors. The values placed on health states are also set out clearly for comment. Costs are calculated from a resource use protocol and the most challenging issue raised is usually how the cost differs if length-of-stay were a bit longer. In fact there are fundamental questions about the real value (shadow price) of resources. Economists need to develop the skills to ask these questions as well.

5. What are the benefits?

As Gatsonis (2000) says, "The evaluation of imaging modalities is primarily an assessment of the value of information, which is subsequently incorporated into health care decision-making and practice." (p.681). It is recognised that the economic evaluation of screening programmes should include the value of information as distinct from the value of the change in health outcome (positive or negative). By extension, it could be argued that information has value in the context of diagnostic testing. In the NSCLC example, patients might test positive or negative for lymph node disease. If they are positive, surgery is ruled out - the patient might be relieved to avoid surgery they are now told would have been unnecessary, but this must be set against their recent diagnosis of cancer and the new information which suggests they have little more than a year to live. If patients test negative, they can proceed to surgery. Again, they may be pleased to be undergoing surgery and to know their disease is not as extensive as it could be. On the other hand, without medical knowledge they might not have considered the possibility in the first place.

It is also possible that doctors are reassured by the information they receive from diagnostic tests. Descriptions of the diagnostic decision-making process suggest doctors often make judgements based in incomplete information and that this causes them some unease; fuller information would reassure them that they had made the decision in the patient's best interests.

Of course, empirical testing could potentially resolve this. Apart from any practical problems with surveying this group at this point in their lives, such an exercise is

obviously not possible in the context of a HTA. Part of the appeal of QALYs is the body of literature that has been built up, with tools such as Dolan et al's EQ5D survey of values being especially valuable (Dolan et al (1995)). There are no comparable 'ready reckoners' for the value of information. Until research into this is commissioned, these benefits will continue to be noted but not quantified.

Sassi et al note that it is probably acceptable to omit the doctor's utility from the evaluation, but note that in practice doctors may not behave exactly as the model predicts and may require incentives (or regulation?) to act in the way assumed.

The other aspect of benefit that economic models do not capture is the possibility that using PET in staging NSCLC represents an early stage of the learning curve - either directly or indirectly the existence and availability of PET might stimulate further progress in imaging. In principle this could be handled in a sensitivity analysis but in practice it is more likely to be included in the text of the report as a factor to consider alongside cost-effectiveness.

Issues in this section
Intuitively, the benefits of PET must exceed the QALY gain but capturing this is difficult, especially in a secondary evaluation. Can we develop any system, even of crude approximations that ensures these factors do not go un-valued (and therefore playing little role in the evaluation)?

6. Can modelling compensate for a lack of clinical evidence?
Typically, economic evaluations of diagnostic technologies have used a modelling approach with an outcome measure that falls short of final health. Severens and van der Wilt (1999) reviewed economic evaluations of diagnostic tests published between 1992 and 1997 and identified through MEDLINE. Their findings were:

- Where it was possible to state which type of economic study was carried out, 21% were CMAs, 64% CEAs, 9% CUAS, with the remainder (6%) being a combination.
- The most common measures of benefit were: ‘specific measure’, case detected, life saved, QALY, patient, life-year saved.
- The main sources of clinical evidence were: 8% RCT, 32% prospective study (controlled?), 15% retrospective study, 30% modelling study (but where did these get their data from?), 15% other / combination / not stated.

In the case of PET in NSCLC there are numerous reports on sensitivity and specificity (referred to below as 'accuracy'), but only a small amount of literature on how the imaging results impact on patient management and virtually nothing on how it affects the patient's health. Even the accuracy evidence has been subject to criticism:

- the American Veterans' Association report speaks of "significant deficiencies" (quoted in MSAC);
- the NCCHTA report notes that these studies are open to bias and were based on small numbers of patients (Robert and Milne);
- MSCAC itself found "some biased selection of cases and/or biased selection of areas for confirmation biopsy (diagnostic bias and work-up bias) according to test results. Further the actual PET scan was included in the reference standard in some studies, leading to an over-estimate of diagnostic accuracy." and

- the American Medicare reimbursement committee went so far as to produce a table to describe the likely biases in the evidence available (see appendix to this paper).

The Appendix to this paper summarises the potential sources of bias in this type of evidence.

The MSAC report offers a good example of an evidence-based body grappling with this issue. They admit there is "no direct evidence" linking accuracy to outcomes, but they also note "documented examples where the results of PET have led to changes in patient management". They say, "If it is assumed that changes in management result in improvements in health outcomes, then it is reasonable to infer the improvements in diagnostic accuracy will lead to improved health outcomes. It is, however, not always clear how changed management will impact on clinical outcomes." (page viii). So how do they justify their decision? "It should be noted that this assumed relationship between diagnostic accuracy and health outcomes is not restricted to PET; the same assumptions will apply to most modern diagnostic technologies …" (page ix).

Similarly, the American Medicare coverage decision was to reimburse for PET scans in staging NSCLC, based on safety and accuracy; cost-effectiveness was not a factor. The chair of the committee admitted, "It raises the whole philosophy of what standard of evidence to hold a new imaging technology to, what clinical value does an added test make in clinical treatment." (quoted in Newman (2001)). An additional factor in America was political interest, with 20 senators from both main political parties lobbying the Secretary of Health and Human Services to increase the availability of PET scanning (Coleman et al (2001)).

A critical reading of these reports would be that the evidence base is skewed towards accuracy, that even this is subject to bias, but since this has been true of diagnostic technologies in the past then it would be hard to apply different standards now. The report concludes that while unrestricted funding for PET is not warranted, it is "potentially effective and potentially cost-effective" and hence should receive interim funding in nine different indications (including staging NSCLC).

An HTA group taking an evidence-based perspective may decide that the absence of evidence of an impact on outcomes (and even equivocal evidence on the change in patient management) means that the technology cannot be recommended for funding without further research. Should an economist support such a policy? Modelling gives economists a tool (and an attitude to uncertainty) that makes them less insistent on having proof to RCT standard for everything.

Sassi et al favour more formal modelling of each of the three stages i.e. accuracy, patient management, health outcome. They suggest three possibilities for modelling the process from accuracy through to outcome:

- simple inference (i.e. assumption) - this is unsatisfactory as there are few means of validating these assumptions.
- modelling the decision-making process - some authors such as Fischer et al (2001) appear to be building up to this approach by presenting their comparison of CT and PET in terms of pre-test and post-test probabilities of having a disease

(Fischer). However, Sassi et al warn that the process is almost certainly characterised by inductive reasoning.

- allow decision-making to be a 'black box' and simply model the empirical relationship. However, this assumes that the studies available are a good guide to what would happen in local practice. Surgical attitudes to the key stages of NSCLC appear to vary between countries.

In the case of PET in staging NSCLC, for example, the approach in the HTBS model sets out a diagnostic protocol, with pre-specified 'rules' that determine which treatment the patient will receive as a result. Survival is then assumed to be the average for other patients of that type. It could be argued that this is the 'best case' for PET in that the clinical decision-maker places complete faith in PET - a more cautious approach would reduce the 'savings' on surgery.

Any approach along these lines assumes that clinical decision-makers would follow these protocols in practice. Some examples will help to illustrate why this might not be the case:

- some surgeons believe that radical surgery offers improved survival even where lymph nodes are affected - in Australia, for example, limited lymph node involvement (N2 disease) is managed by chemotherapy followed by surgery, whereas in the UK surgery is not generally considered as an option;
- some patients request (demand?) surgery where they have disease which is on the borderline of being resectable e.g. T1N2M0; and
- the surgeon may believe that the metastatic disease can be treated as well e.g. a single small brain metastasis.

In addition, one of the attractions of PET is that it can detect very small foci of metastatic disease in the body that cannot be seen by any other viewing modality. This very uniqueness also causes suspicion amongst some doctors. They would argue that if the metastatic disease cannot be seen by any other means then it might be at a very early stage. It is even conceivable that resecting the primary tumour in these cases will allow the body's immune system to keep further spread in check for some time.

The problems are best seen in the two RCTs reported to date. Patients were randomised when their diagnosis was confirmed and received either conventional work-up or a protocol that included PET. One RCT showed no difference in the operation rate (97.5% versus 96.4%) (Boyer et al (2001)), while the other showed the rate being halved (41% vs. 21%) (van Tinteren et al). Not only is the effect size different but the control group baseline suggests fundamental differences between patient groups, clinical decision-making, or both - presumably the latter trial recruited at an earlier stage of the diagnostic process.

So the economist trying to model the impact of PET on treatment has a problem. The uncontrolled study by Saunders mentioned earlier suggests a 15% reduction in surgery rates, which is closest to the second RCT. Of course, it is possible to include a range of values in the sensitivity analysis but since decision-makers have no way of knowing what the true value of this might be it is unclear how this will help.

Issues from this section

Sassi et al have the right idea in trying to break these down into different stages but the actual modelling is enormously difficult. Has this ever been attempted and if so, how & with what degree of success?

7. What to do when the clinical evidence runs out

PET appears to be at stage 2 of Sculpher et al's classification (Sculpher et al (1997)). In other words it is a maturing innovation, characterised by case series and small RCTs with provision from a few specialist centres. If the classification is taken as a 'straight line' progression the next stage (close to widespread diffusion) would require large RCTs. For a diagnostic technology, however, this may not be quite so straightforward. Three barriers have been suggested with respect to RCTs of PET in the UK:

(i) the PET scanner is an indivisible input and few research councils can afford the capital cost plus revenue for new capacity. With existing scanners full to capacity there will be no research until somebody invests in additional capacity;
(ii) there has been difficulty in agreeing protocols for the two arms of trials; and
(iii) there is anecdotal evidence that doctors who refer patients to PET scanners already would now regard randomisation for a clinical trial as unethical.

A further issue would be the cost-effectiveness of using research resources in this way. The cost of the scanner has been noted above, but there is also the question of what would become of it if, after a few years, RCTs showed it was not cost-effective. More broadly, there is the question of the added value of the information that it would provide. As noted already there are two RCTs in this field with quite different results - if we take change in management as the appropriate endpoint then one is negative and one is broadly positive. From this, we can estimate how big a third trial would have to be and what effect size it would have to have to 'prove' (subject to standard significance tests) that PET was or was not effective and cost-effective in preventing operations. We know, however, that statistical evidence alone is rarely sufficient to clinch an argument of this type. This is compounded by concerns about using change in management as an endpoint because it limits the international transferability of the results. There are already concerns about the transferability of data with clinical endpoints but introducing a 'human' element as well (clinical judgement, clinical response to incentives, traditions of working within tightly defined protocols, patient demands, and so on) limits their applicability to another hospital, let alone to another country.

The cost of the research might be considerable, especially when one considers that the trials so far have only looked at one of the three or four possible uses of PET in staging NSCLC. There are a similar number of questions for other aspects of NSCLC, plus other cancers, plus cardiological and neurological applications. Of course, trials can be carried out simultaneously (they may have to be to keep the scanner running to capacity) but the complexity is considerable.

If further research is the chosen option, then economics can help by constructing preliminary economic evaluation models that can be used to either (i) identify the key variables and target subsequent research on these and (ii) to estimate sample sizes that will allow differences in cost-effectiveness to be proven. They can also guide data

collection in the trial. Note, however, that some of the literature in this area is less than encouraging. Drummond and Coyle (1998) simulated a pilot study at the end of their main study of radiotherapy in NSCLC, for example, and found that the final cost of radiotherapy did not fall within the 95% confidence interval that would have been predicted by the pilot study. Further research is required to assess whether this was misfortune or a more general problem.

Issues from this section

Models might be the best way forward, supplemented by specific pieces of research designed to assess the values for key variables identified through cruder modelling. In some cases RCTs will be needed but in other cases this will not be cost-effective. Are there examples that will help to guide this work and what issues arise?

8. Level of detail required

Suppose that an HTA body declared a cost-effectiveness threshold below which technologies would be approved and above which they would be rejected. Then an economic evaluation that was 'fit-for-purpose' (in the sense of enabling a decision to be made) would focus on threshold analysis around the cut-off point. This would make the economics effort far more cost-effective since it would only have to 'prove' (subject to statistical limitations on chance) that the technology fell one side of the line or the other. It is rare in practice for HTA bodies to adopt this policy. How, then, should the economist try to allocate their own time efficiently?

There is a *potential* dilemma between carrying out the evaluation to an academic standard (loosely defined as meeting the standard required to be published in a peer-reviewed journal) and an evaluation that is sufficient upon which to make a decision. In part this debate has been held already (see e.g. Stevens et al (1995), Freemantle and Mason (1999)) but the timescale for HTAs is bringing this issue back to the fore. To rushed an approach and there may be a serious methodological flaw which undermines the results. Too studious an approach and there will be little or no new information for HTA decision-makers to use.

In the context of PET, possible decisions include:

- fund for NHS according to a protocol,
- fund for research to fill gaps in evidence base,
- do not fund now but reconsider at some point in the future,
- do not fund with no plans to reconsider,
- some combination of NHS and research use.

An economic evaluation could be described as 'fit for purpose' if it guided the committee to a decision (although of course this is not to say that the economist should have a preference). The potential conflict with academic standards might be:

- a premium is placed on transparency in order to make the evaluation accessible to non-economists (and non-statisticians) among the decision-makers. This might involve settling for point estimates where a more sophisticated statistical approach might be more widely accepted.
- compromising on data inputs such as the use of Scottish Health Service Costs or taking utility values from the literature even though the source study is not of especially high quality. In an evaluation that is 'fit-for-purpose' these might be justified on the grounds that they make little difference to the economic

conclusion of the HTA. However, viewed from another perspective these might be seen as cavalier compromises in danger of bias.

The picture is further complicated when, as in the case of PET, the quality of the clinical evidence and a 'philosophical' argument about the appropriateness of different endpoints are also likely to be factors.

This is not an easy problem to resolve but more research into 'evidence-based corner-cutting' in economic evaluation would be helpful. One example in the literature is the work carried out in costing colorectal cancer care which suggested that the gains in accuracy from counting items other than length-of-stay and theatre time (plus intensive care?) were small. If similar work established how a crude (but quick) approximation related to the 'gold standard' thorough approach advocated in textbooks then the dilemma would be partly resolved.

Issues from this section

Is the distinction between ‘academic standards’ and ‘fit-for-purpose’ genuine or a false dichotomy? Is it a matter for concern if economic evaluations used as the basis for HTAS are rejected when submitted for publication in peer-review journals?

Summary

The HTA has thrown up a number of interesting questions with regard to evaluating a diagnostic technology, evaluating a technology at a particular stage of its development and the role of modelling relative to clinical trials. The design and execution of the economic evaluation has involved a number of judgements on the part of the HTA team, yet other issues remain unresolved, as discussed above. A fuller discussion of these points will guide HTAs of diagnostic technologies in the future.

References

Boyer M., Viney R., Fulham M. et al. 'A randomised trial of conventional staging with or without positron emission tomography in patients with stage 1 or 2 non-small cell lung cancer' *Abstract 1233*, ASCO 2001.

Coleman R., Tesar R. and Phelps M. 'HCFA and expanded coverage of PET' ***Journal of Nuclear Medicine*** 2001; 42: 11N-12N.

Department of Health "The New NHS – 2000 Reference Costs". Web address: http://www.doh.gov.uk/nhsexec/refcosts.htm.

Dietlein M., Weber K., Gandjour A., Moka D., Theissen P., Lauterbach K. and Schicha H. 'Cost-effectiveness of FDG-PET for the management of potentially operable non-small cell lung cancer: priority for a PET-based strategy after nodal-negative CT results' ***European Journal of Nuclear Medicine*** 2000; 27: 1598-1609.

Dolan P., Gudex C., Kind P. and Williams A. "A Social Tariff for EuroQol: Results from a UK General Population Survey" Discussion Paper 138, Centre for Health Economics, University of York, September 1995.

Drummond M. and Coyle D. 'The role of pilot studies in the economic evaluation of health technologies' ***International Journal of Technology Assessment in Health Care*** 1998; 14: 405-418.

Dwamena B., Sonnad S., Angobaldo J. and Wahl R. 'Metastases from non-small cell lung cancer: mediastinal staging in the 1990s - meta-analytic comparison of PET and CT' ***Radiology*** 1999; 213: 530-536.

Fischer B., Morternsen J. and Hojgaard L. 'Positron emission tomography in the diagnosis and staging of lung cancer: a systematic quantitative review' ***Lancet Oncology*** 2001; 2: 659-666.

Freemantle N. and Mason J. 'Not playing with a full DEC: why development and evaluation committee methods for appraising new drugs may be inadequate' ***BMJ*** 1999; 318: 1480-1482.

Gambhir S., Hoh C., Phelps M., Madar I. and Maddahi J. 'Decision tree sensitivity analysis for cost-effectiveness of FDG-PET in the staging and management of non-small cell lung cancer' ***Journal of Nuclear Medicine*** 1996; 37: 1428-1436.

Gatsonis C. 'Design of evaluations of imaging technologies: development of a paradigm' ***Academic Radiology*** 2000; 7: 681-683.

Gregor A., Thomson C., Brewster D., Stroner P., Davidson J., Fergusson R., Milroy R. on behalf of the Scottish Cancer Trials Lung Group and the Scottish Cancer Therapy Network 'Management and survival of patients with lung cancer in Scotland diagnosed in 1995: results of a national population based study' ***Thorax*** 2001; 56: 212-217.

Information and Statistics Division (ISD), NHS Scotland "Scottish Health Service Costs for Year Ended 31st March 2000". Web address: www.show.scot.nhs.uk/isd/Scottish_Health_Statistics/subject/Costs/2000/costs2000.pdf

Intercollegiate Standing Committee on Nuclear Medicine (ICSCNM) 'Position paper on a strategy for the provision of PET in the UK' (March 2001).

Lottes G., Gorschluter P., Kuwert T., Adam D. and Schober O. 'Costs of F-18-FDG PET: experience with a satellite concept' ***Nuklearmedizin*** 1998; 37: 159-164.

Medicare Services Advisory Committee (MSAC) "Positron Emission Tomography" (Australia, March 2000). Web address: www.health.gov.au/haf/msac

Newman L. 'Medicare Coverage Group assess effectiveness of new technologies' ***Journal of the National Cancer Institute*** 2001; 83: 805-806

Robert G. and Milne R. 'Positron emission tomography: establishing priorities for health technology assessment' ***Health Technology Assessment*** 1999; 3(16) (NCCHTA, Southampton).

Sassi F., McKee M. and Roberts J. 'Economic evaluation of diagnostic technology - methodological challenges and viable solutions' ***International Journal of Technology Assessment in Health Care*** 1997; 13: 613-630.

Saunders C., Dussek J., O'Doherty M. and Maisey M. 'Evaluation of fluorine -18-fluorodeoxyglucose whole body positron emission tomography imaging in the staging of lung cancer' ***Annals of Thoracic Surgery*** 1999; 67: 790-797.

Scott W., Shepherd J. and Gambhir S. 'Cost-effectiveness of FDG-PET for staging non-small cell lung cancer: a decision analysis' ***Annals of Thoracic Surgery*** 1998; 66: 1876-1883.

Scottish Executive (2001a) "Cancer in Scotland: Action for Change. Implementation / Investment Plan 2001-02". Web address: www.scotland.gov.uk/library3/health/csii-00.asp

Scottish Executive (2001b) "Cancer Scenarios: An aid to planning cancer services in Scotland in the next decade". Web address: www.scotland.gov.uk/library3/health/csatp-00.asp

Sculpher M., Drummond M., Buxton M. 'The iterative use of economic evaluation as part of the process of health technology assessment' ***Journal of Health Services Research and Policy*** 1997; 2: 26-30.

Severens J. and van der Wilt G-J. 'Economic evaluation of diagnostic tests: a review of published studies' ***International Journal of Technology Assessment in Health Care*** 1999; 15: 480-496.

Stevens A, Colin-Jones D, Gabbay J. '"Quick and clean": authoritative health technology assessment for local health care contracting' ***Health Trends*** 1995; 27: 37-42

Sutton M. "Personal paper: how to get the best health outcome for a given amount of money" ***BMJ*** 1997; 315: 47-49.

van Tinteren H., Hoekstra O., Smit E. et al 'Randomized controlled trial (RCT) to evaluate the cost-effectiveness of positron emission tomography (PET) added to conventional diagnosis in non-small cell lung cancer (NSCLC)' *Abstract 1885*, ASCO 2001

Wagner J. 'The feasibility of economic evaluation of diagnostic procedures' ***Social Science and Medicine*** 1983; 17: 861-869.

Appendix - Evaluating the accuracy of a diagnostic test

The American Medicare includes a table of the potential biases in accuracy studies in diagnostic technologies:

IDEAL	USUAL PRACTICE	RESULT
Study subjects are consecutive patients seen in a typical clinical setting with a chief complaint	Subjects selected because they have had the diagnostic gold standard	Overestimates sensitivity, underestimates specificity
All patients who get the index test also get the reference test	Patients with negative results on the index test often don't get the diagnostic gold standard	Overestimates sensitivity, underestimates specificity
The person who interprets the index test is blinded to all other information	The person who interprets the index test knows the clinical history and the results of the diagnostic gold standard	Overestimates sensitivity and specificity
The person who interprets the reference test is blinded to all other information	The person who interprets the diagnostic gold standard knows the clinical history and the results of the index test	Overestimates sensitivity and specificity
The reference test is a valid measure of the disease state	The diagnostic gold standard imperfectly measures the disease state	The measured test performance could be either better or worse than the true performance

From page 10 of "National Medicare Coverage Decision on FDG-PET" (15th December, 2000, Medicare Coverage Advisory Committee)