WORK IN PROGRESS:
PLEASE DO NOT QUOTE WITHOUT CONSULTING THE AUTHORS

CAN ECONOMIC EVALUATIONS

CROSS THE CHANNEL?*

by

Michael Drummond[1]
Stephanie Boulenger[2]
John Nixon[1]
Philippe Ulmann[2,3]
Stephen Rice[1]
Gerard de Pouvourville[2,4]

1. University of York, York, United Kingdom.
2. Collège des Économistes de la Santé, Paris, France.
3. Conservatoire National des Arts et Métiers, Paris, France
4. CREGAS, Unité INSERM U537, Le Kremlin-Bicêtre, France

Correspondence to: Professor M F Drummond, Centre for Health Economics, University of York, Heslington, York, YO10 5DD, United Kingdom.

* Paper presented at the 1st Franco-British Meeting on Health Economics, Paris, 14-16 January 2004.

ABSTRACT

Background: Several commentators have identified problems concerning the potential lack of generalisability in economic evaluation results. For example, Späth *et al* (1999) reviewed 26 international studies in the field of adjuvant therapy for breast cancer and found that none were applicable to the French setting. With this issue in mind the aim of the present study is to answer the following questions: (i) can the results of studies be considered generalisable from France to the UK, and *vice versa*? (ii) what are the main reasons for any lack of generalisability? (iii) what could be done, in future studies, to increase the generalisability of results? (iv) how can international databases of economic evaluations, such as the European Network of Health Economic Databases (EURO NHEED), help the users of studies assess the level of generalisability in findings?

Methods: Economic evaluations covering all health technologies and involving the UK and France were located using a previous study by Barbieri *et al* (2003*)* and searches of the UK's NHS Economic Evaluation Database (NHS EED) and the French *Connaissances et Décision en Économie de la Santé (CODECS*) database. Studies were then analysed using a generalisability checklist developed by the authors in the light of previous work on this topic and the database templates used by NHS EED and CODECS. A sub-checklist of the most important items was then derived from the full checklist and a summary score obtained for each study. This approach aimed at determining the degree to which results could be interpreted and replicated in another setting. A summary score for each study was calculated as the percentage (based on the ratio) of correctly addressed points in the study divided by the number of relevant items for each study. For studies providing cost estimates for France and the UK, the results were converted into purchasing power parities (PPP) values in order to identify underlying causes and differences in cost-effectiveness results between the two countries.

Results: 28 economic evaluations met the inclusion criteria: In the overall checklist the range of scores was 35-93% and the mean score (standard deviation) was 69.6% (15.8%). The results for the sub-checklist were very similar. Analysis of the ranking for each sub-category of the checklist revealed that costing, modelling details, assessment of data variability, effectiveness reporting, assessments of generalisability, details of the study population and discounting are areas that need more attention by authors of economic evaluation studies in terms of improving generalisability. Underlying differences in cost-effectiveness results (which ranged from 2.6 to 700%) are often accounted for by price differences between France and the UK, or organisational differences in the delivery of health care. Details are provided for each study.

Conclusions: Although the scores achieved in the sample were surprisingly high, in order to improve the generalisability of economic evaluations authors need to be more explicit and detailed in describing and reporting their studies, particularly in relation to population/sample characteristics and in the derivation of cost estimates. If they are to provide added value to their users, international databases such as EURO NHEED should adopt a comprehensive approach towards assessing and critiquing generalisability, and present their results in a common currency. Further work is planned to improve and develop the results arrived at to date.

1. INTRODUCTION

Users of economic evaluations need to assess whether the results of published studies are relevant to their setting. However, there are a number of factors, varying from place to place, that may limit the generalisability of study results. These include variations in the epidemiology of disease, relative prices, the availability of health care resources and clinical practice patterns [11].

The extent to which economic evaluation results are generalisable has been studied previously. Barbieri *et al* [3] reviewed studies of pharmaceuticals in Europe. They found that 6 out of 44 estimates of cost-effectiveness could be considered generalisable according to their definition, but that the level of generalisability could be conditional on the economic study design.

For example, in trial-based economic evaluations, where the analyst used the pooled trial results for resource utilisation, the cost-effectiveness results were always generalisable. On the other hand, where the data on resource utilisation were allowed to vary from country to country, the cost-effectiveness results were never generalisable.

The other main finding from the study by Barbieri *et al* (*ibid*) is that the differences in cost-effectiveness results from country to country were rarely systematic. That is, a decision-maker in Country A could not normally infer that, because a drug has been studied in Country B and found to be cost-effective, it would be cost-effective in Country A.

In another study, Späth *et al* [33] reviewed the literature to identify economic evaluations of adjuvant therapy in women with breast cancer. They identified 26 studies in all, six of which met their criteria for selection. However, none of these were considered generalisable to the French health care system, the main reason being that cost data were not reported in a transparent way.

Databases of economic evaluation, such as the French CODECS (Connaissances et Décision en ÉConomie de la Santé) database and UK's NHS EED (the NHS Economic Evaluation Database), seek to help decision-makers assess the reliability and relevance of published economic evaluations. They do this by critically appraising studies according to a pre-determined checklist. In a study of the usefulness of economic evaluations, undertaken in two English health authorities using NHS EED, decision-makers reported that the concern about the lack of generalisability of economic evaluation results was one of the main reasons why they considered the published literature to be of limited use [16].

Therefore, if databases of economic evaluation are to be of use to decision-makers, they need to provide assessments of the generalisability of published studies. This need has become more apparent following the initiation of the Euro NHEED (European Network of Health Economic Databases) project[1], funded by the European Union, which includes the CODECS and NHS EED databases as well as five other European countries in which databases will be implemented. Here the objective is to provide decision-makers across Europe with information on the reliability and relevance of economic evaluations undertaken in all the member states.

Although the definition of generalisability in the field of health economics is not firmly established, for the purposes of this study we consider the generalisability of a study to be 'conditional on the elements in the papers (mainly methodology and levels of reporting) that are present (or absent) and which could allow (or prevent) someone using the study in order to apply it to their own context.'

[1] Details of this project, including participating countries and centres, are available at http://www.euronheed.org.

Therefore, the objectives of this paper are: (i) to assess the level of generalisability of economic evaluations of health care programmes and treatments undertaken in both the UK and France; (ii) to identify the main reasons for any lack of generalisability in study results; (iii) to assess what could be done, in future studies, to improve the generalisability of results; and (iv) to outline the ways in which databases of economic evaluations can better assist the users of studies to assess the generalisability in study findings.

2. METHODS

Identification and selection of studies

The first step was to select a sample of full health economic evaluations (i.e. studies in which a comparison of two or more treatments or care alternatives was undertaken and in which both the costs and outcomes of the alternatives were examined) published in peer-reviewed journals. The papers selected mainly concerned drugs, but also medical devices. Besides having to be a full economic evaluation, the geographical context of the studies had to be, at least, France or the UK.

The starting point for the identification and the selection of the studies was a paper [3] on the generalisability of economic evaluations of pharmaceuticals in Western Europe. The goal of the paper was to analyse the sources of variation between study results, to see if these were systematic between countries and to assess whether the extent of variation in study results between countries was important for decision-making. From the list of papers selected for the article, the multi-country studies that included at least France and the United Kingdom (UK) were identified, along with single-country studies from France or the UK, which studied the same health technology, using the same methodology. This enabled comparisons to be made regarding the results between the two countries.

A second step was to search the CODECS and NHS EED[2] databases for additional multi-country studies that included at least France and the UK. In the next phase of this research it is intended to match, according to the health technology studied, French and UK single-country studies contained in NHS EED and CODECS. This will allow us to have a larger number of studies in our sample and increase the level of reliability of the present research.

In all, 36 papers were considered in the present analysis. Barbieri *et al's* paper allowed the identification of 23 studies, while the searches of the CODECS and the NHS EED databases identified the 13 remaining articles. Of these, 28 were multi-country studies and 8 were single-country studies. Twenty-eight papers have been analysed to date. The remaining eight papers, were either not full economic evaluations (one was a methodological paper and one was a literature review) or had not been abstracted by either CODECS or NHS EED.

Description of studies and abstracts

For the purposes of analysis and clarification, studies were placed into one of four sub-categories. The first sub-category was 'multi-country studies' (MS) which refers to studies in which the UK and France were involved (usually as part of a larger group of countries), and for which cost estimates were made and reported for both France and the UK.

The second sub-category was 'multi-country studies (1)' (MS1-UK or MS1-F), which refers to a multi-country economic evaluation in which the costing was performed for only one country, in this case either the UK or France.

[2] **NHS EED**: http://nhscrd.york.ac.uk/; **CODECS**: www.inserm.fr/codecs/codecs.nsf

The third sub-category was 'multi-country study (pooled)' (MS-p) for which costing was based on pooled data from a number of countries that included both the UK and France.

The fourth and final sub-category was that of single-country studies (SC-UK and SC-F), which refers to studies carried out independently in both the UK and France for the same health technology.

Once the studies had been identified, the associated papers and NHS EED and/or CODECS abstracts were retrieved. As far as possible the NHS EED/CODECS abstracts were used as the principal source for the completion of the generalisability checklist, but the paper was referred to in situations where the abstracts indicated there was (required) further information in the paper, or where the required information was known not to be routinely recorded in the structured abstracts (e.g. descriptive comments regarding the sensitivity analyses conducted in studies).

Development of the generalisability checklist

To evaluate the degree of generalisability of the papers, a checklist was developed, initially based on the Späth *et al* [33] approach. The goal, however, of the new checklist was to provide a means of assessing generalisability according to the agreed definition (provided in the background section).

The generalisability checklist (Table 1) was developed through discussion between all the co-authors, in conjunction with the methodological guides that are used by the CODECS [7] and NHS EED [29]databases for the production of structured abstracts.

The checklist is divided into six main sections: the subject and key elements of the study (questions Q1 to M2), characteristics of the methods used to measure clinical outcomes (E1 to E8), the measure of health benefit(s) used in the economic analysis (B1 to B5), the costs (C1 to C11), discounting (D1 to D4) and discussion by the authors (S1 and O1). For each question, answers were classified as follows: Yes, Partially, No/Information not provided, Not Applicable (N/A).

The checklist was independently piloted by two of the co-authors using the first six studies in the sample. Following this exercise, the checklist was then modified in order to make it as objective, efficient (in not duplicating questions) and comprehensive as possible. This involved clarifying and rephrasing some questions, and eliminating or adding some others. The main addition to the first draft was the section on discounting, which was felt to be an essential element in the assessment of generalisability, since the recommended discount rates vary from jurisdiction to jurisdiction.

When the final checklist was agreed, it was completed for each study identified in the literature search. As many studies as possible have been assessed by two or more of the authors and consensus reached through discussion, for questions where the assessors' answers differed. However, due to time constraints, a number of studies have been checked (to date) by only one of the co-authors, but this was always done after blindly assessing two or more papers with one of the other co-authors and discussing any differences before reaching agreement on the final assessment.

As the checklist had grown beyond the anticipated number of items, many of which could be considered to relate to 'internal validity' issues, it was decided to develop a more concise sub-set of items considered to be the most important for assessing generalisability. The thought was that, whilst a study could have a high score on the overall checklist, it may be deficient in several important areas. In order to achieve this, each co-author independently selected from the full generalisability checklist the sub-set of questions they felt were most essential to

judge the generalisability of a paper. As differences of opinion were found between the co-authors, only checklist questions that were selected by more than three (out of five) co-authors were included in the final sub-set.

The sub-checklist contains 16 questions (HT1, HT2, SE2, P1, SP1, SP3, E7, E8, B5, C1, C5, C6, C7, C9, S1 and O1) shown as starred items in Table 1.

Table 1 Generalisability Checklist

Study question	
Q1	Is the study question clearly stated?
Q2	Are the alternative technologies justified by the author(s)?
Health technology	
HT1*	Is the intervention described in sufficient detail?
HT2*	Is(are) the comparator(s) described in sufficient detail?
Setting	
SE1	Is the setting (inpatient, outpatient, home, community) of the study clearly specified?
SE2*	Is(are) the country(ies) in which the economic study took place clearly specified?
Perspective	
P1*	Do the authors state which perspective they adopted?
Studied population	
SP1*	Does the article clearly state the target population of the health technology?
SP2	Are the population characteristics described? (e.g. age, sex, health status, socio-economic status, etc.)
SP3*	Does the article provide sufficient detail about the study sample(s)?
SP4	Is the study sample representative of the stated population?
Modelling	
M1	If a model is used is it described in sufficient detail?
M2	Are all the sources used to construct the model given?
Effectiveness	
E1	If a single study is used are the methods of data collection adequately described (sample selection, study design, allocation, follow-up)?
E2	If a single study is used are the methods of data analysis adequately described (ITT/per protocol or observational data)?
E3	Are side-effects or adverse effects addressed in the analysis?
E4	Do the authors undertake and report statistical analyses?
E5	If based on a review/synthesis of previous published studies, are review methods adequately described (search strategy, inclusion criteria, sources, judgement criteria, combination, investigation of differences)?
E6	If based on opinion, are the methods used to derive estimates adequately described?
E7*	Is the level of reporting of the effectiveness results adequate?
E8*	Does the article provide the results of a statistical analysis of the effectiveness results?
Benefit measure	
B1	Do the authors specify any summary benefit measure(s) used in the economic analysis?
B2	Do the authors report the basic method of valuation of health states or interventions?
B3	Do the authors specify the source(s) of health states (e.g. specific patient population or the general public)?
B4	Do the authors specify the valuation tool used?
B5*	Is the level of reporting of benefit data adequate (incremental analysis, statistical analyses)?
Cost	
C1*	Are all the cost components/items used in the economic analysis presented?
C2	Are the methods used to measure costs components/items provided?
C3	Are the sources of resource consumption data provided?
C4	Are the sources of unit price data provided?
C5*	Are unit prices for resources given?
C6*	Are costs and quantities reported separately?
C7*	Is the price year given?
C8	Is the time horizon given for each element of the cost analysis?
C9*	Is the currency unit reported?
C10	Is a currency rate conversion given?
C11	Does the article provide the results of a statistical analysis of cost results?
Discounting	
D1	Was the summary benefit measure(s) discounted?
D2	Were the cost data discounted?
D	Do the authors specify the rate(s) used?
D4	Were discounted and not discounted results reported?
Variability	
S1*	Are quantitative and/or descriptive analyses conducted to explore variability from place to place?
Generalisability	
O1	Did the authors discuss caveats regarding the generalisability of their results?

* = items selected to form the generalisability sub-checklist (as outlined below)

Summary of analysis methods

To assess if study results were comparable across the two countries, the cost and/or cost-effectiveness ratios were compared for those studies that had comparable outcomes for both the UK and France. Furthermore, it was thought beneficial to compare the study results between France and the UK for appropriate sub-categories (studies providing results for both countries) in order to identify the source of differences. This was achieved by converting the results of these studies to purchasing power parity rates (PPP), which allows adjusting for the level of prices between countries. The most convenient common currency to adopt was US$, the results being given for the price year reported in the original study.

To evaluate the potential generalisability of the results of each study, a percentage score of generalisability was derived for both the full checklist and the sub-checklist. Each question in the checklists had four possible responses, as outlined above, for the assessor to select. To facilitate the calculation of a percentage score, the responses were given the following scores: 1 for "Yes", 0.5 for "Partial" and 0 for "No/No information". "No information" was penalised equally as "No" since the provision of the relevant information was considered to be essential for the generalisation of the study results. When the response to a question was "N/A", the question was excluded from the scoring by reducing the denominator accordingly. A summary (percentage) score was therefore derived using the following formula:

$$\frac{1}{n - x} \sum_{i} S_i \times 100$$

where $i = 1,..,n$, 'n' is the number of questions, 'x' is the number of questions for which the response was N/A: not applicable and 'S' is the score of each question[3].

A number of statistical analyses were performed on the results to investigate various relationships. These included descriptive reporting of mean and standard deviation for total and sub-sets of the results from the checklist, and a correlation analysis between the overall score and the reduced generalisability sub-set of items.

3. RESULTS

The principal results are shown in Appendix 1 (study details, overall (and sub-set) checklist score for each study, price year, plus original and converted cost and/or cost-effectiveness results), and Appendix 2 (differences in cost-effectiveness results between the UK and France for multi-country studies that include results for the UK and France). The results of the statistical tests are given in Table 2, Table 3, Table 4 and Appendix 3, as outlined in the following sections.

For the 28 studies assessed with the checklist (Table 2), the mean percentage score was 69.6% with a standard deviation of 15.8%. The marks are distributed between 35% and 93%. With the sub-checklist of questions the results were comparable to the original scoring system. Indeed, for the 28 studies very similar results were achieved in terms of the mean score but some differences were observed in the distribution (see Table 2): The mean percentage score for the sub-checklist was 70.9% and the standard deviation is 17.1%. The marks were distributed between 37% and 100% (maximum mark for two studies).

[3] This method is similar to previous approaches adopted by researchers in developing a quality score from checklists (for example Nixon and Pang (2000) 'Economic evaluations in Japan: A review of published studies, methodological issues and practice.' In PSAM5 (eds. S. Kondo and K. Futura, Universal Academy Press, Inc., Tokyo, Japan.)

Table 2: Scores with the two checklist systems

Range (%)	0-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100	**Mean percentage score**
Overall score	0	0	0	2	1	3	8	6	4	4	69.6
Sub-set score	0	0	0	1	3	2	7	8	1	6	70.9

The first observation is that the scores are rather high, perhaps more than expected, with a limited standard error (around 16%). Therefore most of the analysed studies obtained what might be considered a good mark. One possible explanation for this is that the studies selected have already been published in a peer-reviewed journal and selected by the NHS EED or CODECS databases and therefore the results might reasonably be expected to be skewed to the left. In other words, many of the items in the checklist, whether or not they concern generalisability issues, are satisfied by these studies since they went through several reviewing processes. This issue has to be explored further by concentrating on more focused generalisability criteria. This was attempted by using the sub-checklist. However, the results obtained with this sub-checklist were very similar to the full checklist.

We then tried to test for any potential relationship between the mark (overall and sub-set) and the characteristics of the study or the checking process. We used the Chi-square test to analyse the link "score/type of study" (multi-country studies *versus* single country studies) and the link "score/team of assessors" (York's team *versus* CES's team).

For both variables and with both scoring systems, we did not obtain any statistically significant relationships among the results. The best outcome was a significant test at the 20% threshold with the Chi-square test for score/team of assessors.

The number of studies in the sample is possibly too low to reveal any robust links, despite the fact that it was thought there would be a difference in scoring levels between the York and CES teams (York's mean score was 79% while CES's mean score is 65%). Indeed, if we analyse the discrepancies between centres and assessors, we can note that the answers to the questions of the checklist were sometimes different from one person to the other. Although every attempt was made, within the limitations of the study, to design questions in such a way as to eliminate different interpretations and to ensure clarity in terms checklist questions, we still observed differences between researchers for studies that were not evaluated by two researchers.

It will be necessary for the next steps of the study, to explore the sources of differences in the evaluation system by the two centres and the evaluators. This can be due to the heterogeneity in the quality of the studies and/or to the difference of interpretation and consequently in the attribution of the scores. In order to explore this issue, it will be necessary to double-check each economic study to identify sources of differences and in a second step, if required, refine the checklist.

Regarding the scores obtained by question and group of question, the analysis gave some valuable information. The two general sections (study question and setting) had very high scores (more than 0.9), confirming our previous comments about the consistency and quality of papers undergoing peer review. On the other hand, discounting, study population, and effectiveness sections obtained low scores (from 0.50 to 0.6, see Table 3). The detail for each checklist item is given in Appendix 3 and Table 4.

Two particularly important aspects for assessing generalisability are the questions on cost and effectiveness, which are ranked 6th and 9th out of the 12 sections of the checklist. The mean marks obtained by these two groups were respectively 0.691 for costs (which was close to the overall average of 0.696) and 0.610 for effectiveness. Both had a high standard deviation, being 0.436 for effectiveness and 0.415 for cost.

It should be noted that one question in the cost section ("Is the currency unit reported?") obtained for all the 28 studies the maximum score of 1, which is perhaps the least researchers and decision-makers can expect from an economic evaluation study.

Table 3 Individual marks for checklist items

Rank	Item (number)	Mean (s.d) mark per question	Mean No. of 0	Mean No. of 0.5	Mean No. of 1	Mean No. of N/A
1	Q (2) Study question	0.938 (0.213)	1	1.5	25.5	0
2	SE (2) Setting	0.911 (0.252)	1.5	2	24.5	0
3	P (1) Perspective	0.750 (0.443)	7	0	21	0
4	HT (2) Health technology	0.750 (0.382)	4.5	4.5	18	1
5	B (5) Benefit measure	0.741 (0.386)	2.8	2.8	10.1	11.6
6	C (11) Cost	0.691 (0.415)	6	4.4	16.1	1.5
7	M (2) Modelling	0.667 (0.373)	2	4	6	16
8	S (1) Variability	0.614 (0.299)	2	13	7	6
9	E (8) Effectiveness	0.610 (0.436)	5.4	3.5	9.4	9.8
10	O (1) Generalisability	0.607 (0.386)	6	10	12	0
11	SP (4) Study Population	0.599 (0.443)	7.5	4.3	12.3	4
12	D (4) Discounting	0.500 (0.493)	4.3	0.4	4.3	19.3

Note: the eleven items in the cost section, ranked 6th, gave a mean of 0.691 and a standard deviation of 0.415. For these 11 questions, the mean number (on the 28 studies) of 0 is 6, of 0.5 is 4.4, of 1 is 16.1 and only 1.5 for N/A.

However, if this criterion were excluded, the cost section would have obtained a mean mark of 0.658. Another important topic for assessing generalisability is the description of the study population characteristics. Since this section obtained a low mark of 0.599, and almost three out of four items in this section had a score between 0.4 and 0.5, these results are not very encouraging. Therefore researchers should seek to address this in future studies.

This led us to consider whether the overall checklist score or the sub-checklist score reflected the real level of generalisability of the paper. This issue would need to be explored further, using different procedures of evaluation (e.g. qualitative evaluation and expert opinion).

Table 4 Score (mean and standard deviation) for each checklist item

Checklist item	Mean	s.d.
Q1. Is the study question clearly stated?	1.00	0.00
C9. Is the currency unit reported?	1.00	0.00
D3. Do the authors specify the rate(s) used?	1.00**	0.00
SE2. Is(are) the country(ies) in which the economic study took place clearly specified?	0.96	0.19
SP1. Does the article clearly state the target population of the health technology?	0.93	0.22
B3. Do the authors specify the source(s) of health states (e.g. specific patient population or the general public)?	0.91*	0.29
Q2. Are the alternative technologies justified by the author(s)?	0.88	0.29
C1. Are all the cost components/items used in the economic analysis presented?	0.88	0.29
SE1. Is the setting (inpatient, outpatient, home, community) of the study clearly specified?	0.86	0.29
B2. Do the authors report the basic method of valuation of health states or interventions?	0.85*	0.32
E4. Do the authors undertake and report statistical analyses?	0.83	0.37
B1. Do the authors specify any summary benefit measure(s) used in the economic analysis?	0.82	0.38
C3. Are the sources of resource consumption data provided?	0.80	0.34
C7. Is the price year given?	0.80	0.39
C2. Are the methods used to measure costs components/items provided?	0.79	0.34
HT2. Is(are) the comparator(s) described in sufficient details?	0.77	0.35
C8. Is the time horizon given for each element of the cost analysis?	0.77	0.39
P1. Do the authors state which perspective they adopted?	0.75	0.43
HT1. Is the intervention described in sufficient detail?	0.73	0.41
M1. If a model is used is it described in sufficient detail?	0.71*	0.32
E7. Is the level of reporting of the effectiveness results adequate?	0.70	0.41
B4. Do the authors specify the valuation tool used?	0.64**	0.35
C4. Are the sources of unit price data provided?	0.64	0.40
M2. Are all the sources used to construct the model given?	0.63*	0.41
S1. Are quantitative and/or descriptive analysis conducted to explore variability from place to place?	0.61	0.30
E1. If a single study is used are the methods of data collection adequately described (sample selection, study design, allocation, follow-up)?	0.61	0.36
E2. If a single study is used are the methods of data analysis adequately described (ITT/per protocol or observational data)?	0.61	0.39
O1. Did the authors discuss caveats regarding the generalisability of their results?	0.61	0.39
B5. Is the level of reporting of benefit data adequate (incremental analysis, statistical)?	0.56	0.38
C5. Are unit prices for resources given?	0.55	0.36
E3. Are the side-effects or adverse effects addressed in the analysis?	0.55	0.49
E8. Does the article provide the results of a statistical analysis of the effectiveness results?	0.54	0.48
C10. Is a currency rate conversion given?	0.50	0.48
D2. Were the cost data discounted?	0.50*	0.50
SP2. Are the population characteristics described? (age, sex, health status, socio-economic status, etc.)	0.48	0.42
SP3. Does the article provide sufficient detail about the study sample(s)?	0.48	0.42
E5. Are review methods adequately described (search strategy, inclusion criteria, sources, judgement…) ?	0.45*	0.40
SP4. Is the study sample representative of the stated population?	0.43	0.48
C6. Are costs and quantities reported separately?	0.38	0.41
D1. Was the summary benefit measure(s) discounted?	0.36*	0.48
E6. If based on opinion, are the methods used to derive estimates adequately described?	0.36**	0.35
C11. Does the article provide the results of a statistical analysis of cost results?	0.35	0.47
D4. Were discounted and not discounted results reported?	0.25**	0.38

* : % of N/A in the total of answer is > 50%.
** : % of N/A in the total of answer is > 75%.

The comparison of study results for the 15 papers that included results for both the UK and France showed that twelve were in favour of France. Namely, that the cost-effectiveness (or utility) ratios were lower for France, or that the health technology studied was a dominant strategy compared to the reference strategy (or comparator). The percentage difference in cost-effectiveness ratios varied between 2.6 and 700%.

When considering the effectiveness results it was observed that these are quite similar between France and the UK, the interpretation of this being that the differences in the ratios result mainly from differences in the cost results. In three of the fifteen cases, the source of difference was merely a price effect, since the same resource consumption data were used for

each country varying only in unit costs. In the other cases, it was difficult to identify clearly the source of difference due to a lack of data presented in the article. Although it is clear that the differences are, by implication, due to differences in local resource use or prices, it was often difficult to identify the main reason. However, it was possible, in some cases, to pin-point the main cost item that made a difference as reported in Appendix 2.

4. DISCUSSION

This study has raised a number of issues that merit further discussion. These are discussed in turn.

How should generalisability be assessed?

One simple test of generalisability is whether the cost-effectiveness results obtained in a given study would also be obtained in another setting. (Data on the results obtained from a limited number of studies performed in both France and the UK are presented in this paper.) However, the philosophy behind this paper is that this simple test is not sufficient. Namely, users of economic evaluations are not only interested in the results generated, but also in whether the studies are methodologically sound. For this reason the generalisability checklist presented here includes a number of questions relating to the methodological approach adopted in published studies, rather than generalisability *per se*.

The implication is that the full checklist contains a large number of items, many of which can be found in standard methodological checklists for economic evaluations. This begs the question as to whether some items are more important than others with respect to generalisability. One approach, followed here, would be to identify a sub-set of checklist items. (Our finding was that there was considerable agreement between scores on the two checklists, suggesting that a study that reports well on generalisability issues also reports well overall.)

However, one of the problems of checklist is that, unless items are explicitly weighted, the implicit assumption is that all items are equally important. Therefore another approach would be to conduct a discrete choice experiment (DCE) in order to generate relative weights for checklist items. This approach has been used in one recent study generating overall quality scores for economic evaluations [9]. An alternative approach would be to specify reporting criteria that studies absolutely must meet in order to be considered generalisable.

Can the results of studies be considered generalisable from France to the UK and vice versa?

Only limited data are currently available on this issue. The comparisons for 15 economic evaluations performed in both France and the UK are shown in Appendix 1. One interesting finding, mirroring that of Barbieri *et al*[3], is that, in the vast majority of cases, the cost-effectiveness result for France is more favourable than that for the UK. If this finding were confirmed by further study, the implication would be that a decision-maker in France, on seeing a study performed in the UK, could be fairly confident that even more favourable results would be obtained for France. However, it is difficult from this study to determine the *extent* by which the results would be more favourable.

One limitation is that the comparisons of study results given here relate to the researchers' estimates from their primary, or 'best guess' analysis. Some authors also explore the uncertainty around their estimates, either through simple sensitivity analysis, probabilistic modelling or statistical analysis of cost-effectiveness data. Therefore it is possible that, although the point estimates of cost-effectiveness differ between France and the UK, the confidence intervals (or ranges) around those estimates may overlap.

What are the main reasons for any lack of generalisability?

Even less data are available on this issue (see Appendix 2). However, one important point is that the generalisability of study results can be conditioned by the methodology employed in the study. The vast majority of economic evaluations use the same effectiveness data for all countries, often because these come from clinical studies where the data are pooled. Therefore it is difficult to assess what would be the impact on generalisability if country-specific effectiveness data were used.

The same issue arises with respect to the resource use data in economic evaluations conducted alongside clinical trials. Sometimes these data are pooled and used for all countries. Sometimes country-specific resource use data are used. In the former case there is a risk that real variation between countries is being 'designed-out' of the evaluation. Also, the implication is that only price effects are being explored in such studies.

What could be done to increase the generalisability of study results?

The main implication of this study is that high quality reporting is important in helping the user to assess potential generalisability. That is, the user needs to know whether the methods and data used apply in his or her own setting.

In addition, authors of studies could explore, through sensitivity analysis, whether their results would apply in a different patient population, or a different healthcare setting. However, it would be unreasonable to expect much from study authors in this respect.

How can international databases of economic evaluations (like EURO NHEED) help users assess generalisability?

It might be more realistic to expect international databases to tackle issues of generalisability. First, a generalisability checklist could be applied to studies, in order to let users know which studies (in principle) report in sufficient detail to enable assessments of generalisability to be made.

Secondly, it might be possible to provide a commentary on the features of the patient population, treatment interventions or healthcare setting that might limit generalisability.

Thirdly, it might be possible to highlight important items of resource consumption, or unit costs of resources, that are unlikely to be generalisable to other settings.

These initiatives may help. However, generalisability of studies could only be explored in detail if the model used in the economic evaluation were made freely available so that users could populate it with their own data. However, this would approximate to repeating the study in each and every setting.

What are the main limitations and needs for further research?

The first priority is to complete the analysis of all the papers identified, using two or more assessors. Secondly, the reasons for differences of opinion between assessors need to be explored, so that a standardized, and agreed, assessment of all the papers can be obtained.

Thirdly, the generalisability checklist should be validated using a wider range of health economic experts. This validation process would also include a debate about whether more explicit weights are required for the checklist items and whether some items must always be fulfilled in order for a paper to be deemed generalisable.

Finally, those papers in which results for France and the UK can be compared should be analysed in more detail in order to understand more fully those factors that influence cost-effectiveness, or are subject to the greatest uncertainty.

5. CONCLUSIONS

Although restricted to two countries, this study suggests two main conclusions. On the one hand, the high rating obtained on items related to basic guidelines in the achievement of economic evaluations indicate a strong trend towards the standardization of methods and of the influence of peer-reviewed publications. On the other hand, the lower ratings obtained for items more directly relevant to generalisability, suggest that there is still much to do. Moreover, the variability of scores indicate that assessment is not context-free, and that implicit qualitative criteria have been taken into account by the assessors. This does not necessarily mean that generalisability is low, but that transferring results from one country to another is still a matter of judgment as much as of controlled objective evaluation. Finally, it is somewhat surprising that there is no difference shown between multi-country studies and single country studies: one could expect the former to offer more potential with respect to generalisability than the latter.

If the aim is actually to improve this situation, what can be done? Lessons can be learned from Table 3, but a first point can be made before dealing with the design of studies. Peer reviewed journals generally impose strict constraints on authors regarding the number of pages, tables and figures. Moreover, some journals, mainly medical ones, even have a policy of charging authors for extra pages, which can be a strong disincentive to publish sufficient details, and thus limit generalisability. Some published articles in very prestigious journals are even too concise to assess correctly their full scientific value. This may explain why there is some variability in the scoring of items. Indeed, improving generalisability means that authors have to provide results for several target populations, provide data on resource use and unit costs, and a detailed one-way and multi-parameter sensitivity analysis, so that a reader from another country can easily infer from the publication a plausible result for his or her setting. One way to get round this issue of restricted publication space is to recommend that authors stress in their sensitivity analysis what the key parameters are that determine the results of the evaluation. Statistical methods can probably be designed around probabilistic sensitivity analysis in order to identify which parameters have more influence on the conclusions: a multivariate analysis of the outcomes of sensitivity analysis, using parameters as explanatory variables, could be considered.

Referring back to Table 3, a relatively low score on an item does not indicate that it is more difficult to obtain generalisability for it. For example, ‘discounting’ scores low, but should be an easy issue to settle through sensitivity analysis. An agreement could probably be found amongst economists on a range of discounting rates to be used as standard. More simply, it could be convened that all publications should include at least one scenario with a conventional discount rate agreed upon *ex ante*.

For the other items that score lower than 0.7, with a high standard deviation, the task may be more difficult. For example, the definition of a study population can be contingent on effectiveness data and on specific national concerns. When studies refer to the same clinical trials, they all have at least the same base case. This is the case, in general, in multi-country studies, but single country studies may rely on domestic trials. Moreover, providing cost-benefit results for different sub-groups requires specific analyses, and it is difficult to coordinate efforts *ex ante* among different research teams to agree on common relevant target groups.

The task with cost may be easier, if publications progressively provide detailed information on resource use and on unit costs. Nevertheless, there will always be other relevant factors, linked to practice patterns, organizational differences, or differences in terminology or classification schemes. For example, medical services are not described the same way in

different countries: there are different procedure classification schemes and DRG-like systems are not the same around the world.

Thus, specific recommendations could be made item by item, to improve generalisability. A more generic process could rely on peer-reviewing: editorial boards could adopt a policy whereby a submitted paper should be reviewed by readers from different countries. Each reviewer would be asked to put a special emphasis on generalisability from their national perspective, so as to encourage authors to provide supplementary analyses. Obviously, this would increase time to publication which may not be welcomed by authors. Thus, another process could be for journals to ask academics from other countries to write comments and raise generalisability issues once a paper is published. Internet forums could probably provide a medium to organize such inter-country discussions, and health economists associations or databases such as CODECS, NHS EED and EURO NHEED help to build up such forums.

REFERENCES

1. Ament A, Baltussen R, Duru G, Rigaud-Bully C, De Graeve D, Ortqvist A (2000) Cost-effectiveness of pneumococcal vaccination of older people: a study in 5 western European countries. Clin Infect.Dis. 31:444-50
2. Armstrong SH, Ruckley CV (1997) Use of a fibrous dressing in exuding leg ulcers. Journal of Wound Care 6:322-4
3. Barbieri M, Drummond MF, Willke R, Chancellor J, Jolain B, Towse A (2003) Are the results of economic evaluations generalisable? Evidence from studies of pharmaceuticals in Western Europe. *Value in Health* 6(6):710 (Abs).
4. Belcaro G, Laurora G, Nicolaides AN, Agus G, Cesarone MR, Desanctis MT (1998) Treatment of severe intermittent claudication with PGE1 a short-term vs a long-term infusion plan a 20-week, European randomized trial analysis of efficacy and costs. Angiology 49:885-894
5. Berger K, Fischer T, Szucs T (1998) Cost-effectiveness analysis of paclitaxel and cisplatin versus cyclophosphamide and cisplatin as first-line therapy in advanced ovarian cancer. A European perspective. European Journal of Cancer 34:1894-901
6. Borghi J, Guest J (2000) Economic impact of using mirtazapine compared to amitriptyline and fluoxetine in the treatment of moderate and severe depression in the UK. European Psychiatry 15:378-87
7. Boulenger S (2002) CODECS: Connaissances et Décision en ÉConomie de la Santé: Première base de données francophone sur l'évaluation économique en santé (Knowledge and Decision in Health Economics: The first database in economic evaluation for French-speaking countries). La Lettre du Collège Des Économistes de la Santé 4:2
8. Brown M, van Loon J, Guest J (2000) Cost-Effectiveness of Mirtazapine Relative to Fluoxetine in the Treatment of Moderate and Severe Depression in France. European Journal of Psychiatry 14:15-25
9. Chiou C-F, Hay JW, Wallace JF, Bloom B, Neumann PJ, Sullivan SD, Yu H-T, et al (2003) Development and validation of a grading system for the quality of cost-effectiveness studies. Medical Care 41:32-44
10. Combier E, Levacher S, Letoumelin P, Joseph A, Pourriat J, De Pouvourville G (1999) Cost-effectiveness analysis of the terlipressin-Glycerin trinitrate combination in the pre-hospital management of acute gastro-intestinal haemorrhage in cirrhotic patients. Intensive Care Medicine 25:364-370

11. Drummond M, Bloom B, Carrin G, Hillman A (1992) Issues in the cross-national assessment of health technology. International Journal of Technology Assessment in Health Care 8:671-682
12. Einarson T, Oh P, Gupta A, NH S (1997) Multinational pharmacoeconomic analysis of topical and oral therapies for onychomycosis. Journal of Dermatological Treatment:229-35
13. Group C-IIaHE (2001) Reduced costs with bisoprolol treatment for heart failure. An economic analysis of the second Cardiac Insufficency Bisoprolol Study (CIBIS-II). European Heart Journal 22(12):1021-1031
14. Grover S, Coupal L, Zowall H, Alexander C (2001) How cost-effective is the treatment of dyslipidemia in patients with diabetes but without cardiovascular disease? Diabetes Care 24:45-50
15. Hansson L, Lloyd A, Anderson P, Kopp Z (2002) Excess morbidity and cost of failure to achieve targets for blood pressure control in Europe. Blood Pressue 11:35-45
16. Hoffman C, Nixon J, Stoykova BA, Drummond MF, Misso K (2002) Do Decision-Makers Find Economic Evaluations Useful? Results of Recent Focus Group Research in the UK. Value in Health 5:71-78
17. Hutton J, Brown R, Borowitz M, Abrams K, Rothman M, Shakespeare A (1996) A new decision model for cost-utility comparisons of chemotherapy in recurrent metastatic breast cancer. Pharmacoeconomics 9 Suppl 2:8-22
18. Iveson T, Hickish T, Schmitt C, Van Cutsem E (1999) Irinotecan in second-line treatment of metastatic colorectal cancer: improved survival and cost-effect compared with infusional 5-FU. European J ournal of Cancer 35:1796-804
19. Jönsson B, Johannesson M, Kjekshus J, Olsson A, Pedersen T, H W (1996) Cost-effectiveness of cholesterol lowering. Results from the Scandinavian Simvastatin Survival Study (4S). European Heart Journal 17:1001-7
20. Keown PA, Balshaw R, Krueger H, Baladi JF (2001) Economic analysis of basiliximab in renal transplantation. Transplantation 71:1573-1579
21. Kobelt G, Jonsson L (1999) Modeling cost of treatment with new topical treatments for glaucoma. International Journal of Technology Assessment in Health Care 15:207-219
22. Launois R, Reboul-Marty J, Henry B, Bonneterre J (1996) Pharmacoeconomics. A cost-utility analysis of second-line chemotherapy in metastatic breast cancer. Docetaxel versus paclitaxel versus vinorelbine 10:504-21
23. Leese B, Hutton J, Maynard A (1992) A comparison of the costs and benefits of recombinant human erythropoietin (Epoetin) in the treatment of chronic renal failure in 5 European countries. Pharmacoeconomics 1:346-356
24. Legendre CM, Norman DJ, Keating MR, Maclaine GD, Grant DM (2000) Valaciclovir prophylaxis of cytomegalovirus infection and disease in renal transplantation: an economic evaluation. Transplantation 70:1463-1468
25. Levy P, Lechat P, Leizorovicz A, Levy E (1998) A cost-minimization of heart failure therapy with bisoprolol in the French setting: an analysis from CIBIS trial data. Cardiac Insufficiency Bisoprolol Study. Cardiovascular Drugs and Therapy 12:301-5
26. Levy-Piedbois C, Durand-Zaleski I, Juhel H, Schmitt C, Bellanger A, Piedbois P (2000) Cost-effectiveness of second-line treatment with irinotecan or infusional 5-fluorouracil in metastatic colorectal cancer. Annuls of Oncology 11:157-61
27. Lindgren P, Jonsson B, Redaelli A, Radice D (2002) Cost-effectiveness analysis of exemestane compared with megestrol in advanced breast cancer: a model for Europe and Australia. Pharmacoeconomics 20:101-8
28. Malek M, Cunningham-Davis J, Malek L, Paschen B, Tavakoli M, Zabihollah M (1999) A cost minimisation analysis of cardiac failure treatment in the UK using CIBIS trial

data. Cardiac Insufficiency Bisoprolol Study. International Journal of Clinical Practice 53:19-23
29. NHS Centre for Reviews and Dissemination (2001) Improving access to cost-effectiveness information for health care decision making: the NHS Economic Evaluation Database. York: University of York.
30. Rutten-Van Molken M, Van Dorrslaer E, Till M (1998) Cost-effectiveness analysis of formoterol versus salmeterol in patients with asthma. Pharmacoeconomics 14:671-84
31. Simpson K, Hatziandreu E, Andersson F, Shakespeare A, Oleksy I, Tosteson A (1994) Cost effectiveness of antiviral treatment with zalcitabine plus zidovudine for AIDS patients with CD4+ counts less than 300/microliters in 5 European countries. Pharmacoeconomics 6:553-62
32. Smith I, Terhoeve PA, Hennart D, Feiss P, Harmer M, Pourriat JL, Johnson IAT (1999) A multicentre comparison of the costs of anaesthesia with sevoflurane or propofol. British Journal of Anaesthesia 83:564-570
33. Späth HM, Carrere M-O, Fervers B, Philip T (1999) Analysis of the eligibility of published economic evaluations to a given health care system. Health Policy 49:161-177
34. Stalhammar N, Carlsson J, Peacock R, Muller-Lissner S, Bigard M, Porro G (1999) Cost effectiveness of omeprazole and ranitidine in intermittent treatment of symptomatic gastro-oesophageal reflux disease. Pharmacoeconomics 16:483-97
35. Underwood S, Godman B, Salyani S, Ogle J, Ell P (1999) Economics of Myocardial Perfusion imaging in Europe- The EMPIRE study. European Heart Journal 20:157-166

Appendix 1 Studies included in the analysis with scores, original and converted results (original price years used)

Ref	Type	Original results taken from the paper	Price year	Results converted using the PPP rates	% Score (overall)	% Score (sub-list)
[1]	MS	Invasive pneumococcal disease: Cost/QALY: Scotland = £12,211.44, France = 123,915.72FF. Pneumococcal pneumonia: Cost/QALY: Scotland = £198, France = 13,734 FF.	1995	Invasive pneumococcal disease: Cost/QALY: Scotland = US$ 18,787£, France = US$ 19,276. Pneumococcal pneumonia: Cost/QALY: Scotland = US$ 305, France = US$ 2,136.	76	77
[5]	MS	Cost per life-year saved of Paclitaxel and cisplastin compared with cyclophosphamide and cisplatin: US$ 6,642 (8648/1,302) for France and US$ 6,403 (8112/1,267) for the UK.	NG	Cost per life-year saved of Paclitaxel and cisplastin compared with cyclophosphamide and cisplatin: US$ 5,163 for France and US$ 6,388 for UK.	60	55
[12]	MS	Cost per symptom-free day: The figures are all approximate because the results are presented in a chart. Ciclopirox: 0.2 (FR) versus 0.5 (UK). Amorolfine: 0.20 (FR) versus 0.5 (UK). Itraconazole: 0.45 (FR) versus 0.4 (UK). Terbinafin: 0.3 (FR) versus 0.4 (UK). Griseofulvin: 0.6 (FR) versus 1.1 (UK).	1996	Cost per symptom-free day: The figures are all approximate because the results are presented in a chart. Ciclopirox: 0.16 (FR) versus 0.49 (UK). Amorolfine: 0.16 (FR) versus 0.49 (UK). Itraconazole: 0.35 (FR) versus 0.39 (UK). Terbinafin: 0,23 (FR) versus 0.39 (UK). Griseofulvin: 0.47 (FR) versus 1,08 (UK).	50	50
[14]	MS	Cost per life year saved 40 year old men, diabetes: France = 21588 FF, UK = £2564; 40 year old women, diabetes: France = 62,929 FF, UK = £6,627; 60 year old men, CVD: France = 25,894 FF, UK = £2,983; 60 year old women, CVD: France = 33,076 FF, UK = £3,873.	1998	Cost per life year saved 40 year old men, diabetes: France = US$ 3,324, UK = US$ 3,945; 40 year old women, diabetes: France = US$ 9,690, UK = US$ 10,195; 60 year old men, CVD: France = US$ 3,987, UK = US$ 4,589; 60 year old women, CVD: France = US$ 5,093, UK = US$ 5,959.	53	36
[19]	MS	Cost per life year saved: France = 31,646 FF, UK = £6,983	1995	Cost per life year saved: France = US$ 4,923, UK = US$ 10,743.	70	54
[31]	MS	Cost per life year saved: France = 91,600 FF, UK = £15,264	NG	1994 PPPs, Cost per life year saved: France = 13,826 US$, UK = 23,483 US$	65	61
[34]	MS	The mean total direct costs for 1 year: UK = £292 for omeprazole 20 mg, £286 for omeprazole 10 mg, and £307 for ranitidine 150 mg. France FF 3,122 for omeprazole 20 mg, FF 2,993 for omeprazole 10 mg, and FF 3,235 for ranitidine 150 mg.	1998 and 1995	1998 PPPs. The mean total direct costs for 1 year: UK = US$ 449 for omeprazole 20 mg, US$ 440 for omeprazole 10 mg, and US$ 472 for ranitidine 150 mg. France US$ 48 for omeprazole 20 mg, US$ 461for omeprazole 10 mg, and US$ 498for ranitidine 150 mg.	97	92

Ref	Type	Original results taken from the paper	Price year	Results converted using the PPP rates	% Score (overall)	% Score (sub-list)
[6] ([8])	SC-UK	UK. Mirtazapine was dominant compared to amitriptyline. Six months' treatment with mirtazapine compared to fluoxetine increased the proportion of successfully treated patients by 22% at a net additional cost to the NHS of £27 per patient. The direct NHS cost of managing a patient with moderate and severe depression with mirtazapine was £413 per patient over seven months and £420 over six months, compared with £448 for amitriptyline and £394 for fluoxetine (7 months). The indirect societal cost of managing a patient with moderate and severe depression with mirtazapine was £1,513 per patient over seven months, compared with £1,519 for amitriptyline and £1,360 with mirtazapine or fluoxetine over six months.	1997	Mirtazapine was dominant compared to amitriptyline. Six months' treatment with mirtazapine compared to fluoxetine increased the proportion of successfully treated patients by 22% at a net additional cost to the NHS of US$ 43 per patient. The direct NHS cost of managing a patient with moderate and severe depression with mirtazapine was US$ 656 per patient over seven months and US$ 667 over six months, compared with US$ 711 for amitriptyline and US$ 625 for fluoxetine (7 months). The indirect societal cost of managing a patient with moderate and severe depression with mirtazapine was US$ 2,402 per patient over seven months, compared with US$ 2,411 for amitriptyline and US$ 2,159 with mirtazapine or fluoxetine over six months.	92	90
[8] ([6])	SC-F	France. Additional cost per successfully treated patient = 3,343 FF. Cost for society = 98,883 FF for mirtazapine versus 99,310 FF for fluoxetine. Cost for social security = 117 FF more for mirtazapine versus fluoxetine.	1995	Additional cost per successfully treated patient = US$ 520. Cost for society = US$ 15,382 for mirtazapine versus US$ 15,448 for fluoxetine. Cost for social security = US$ 18,20 more for mirtazapine versus fluoxetine.	35	25
[17] ([22])	SC-UK	UK. £2,431/QALY (docetaxel versus paclitaxel). Total cost per patient = £8,233 for docetaxel versus £8,013 for paclitaxel.	1994	UK. US$ 3,740/QALY (docetaxel versus paclitaxel). Total cost per patient = US$ 12,666 for docetaxel versus US$ 12,328 for paclitaxel.	81	100
[18] ([26])	SC-UK	UK.£7,695/LYG for irinotecan versus B1-5-FU and £11,947 IRI versus B2-5-FU	1998	UK. Cost/LYG: US$ 11,839 for irinotecan versus B1-5-FU and US$ 18,380 for irinotecan versus B2-5-FU.	91	95
[22] ([17])	SC-F	France. Docetaxel is the dominant strategy. Savings = 6,800 FF compared to vinorelbine and 700 FF compared to paclitaxel. Total cost = 250,400 FF for docetaxel, 251,100 FF for paclitaxel and 257,200 FF for vinorelbine.	1993	Docetaxel is the dominant strategy. Savings = 1,037 US$ compared to vinorelbine and 107 US$ compared to paclitaxel. Total cost = 38,173 US$ for docetaxel, 38,280 US$ for paclitaxel and 39,210 US$ for vinorelbine	77	70
[26] ([18])	SC-F	France. Cost per year of added survival: between 9,344 US$ and 10,137 US $ (already adjusted with PPP).	1999		70	55
[25] ([28])	SC-F	France. Cost per patient: 18,975 FF for bisoprolol and 23,306 FF for placebo. Cost per patient who already had heart failure: 21,806 FF for bisoprolol and 27,177 FF for placebo.	1995	Cost per patient: US$ 2,952 for bisoprolol and US$ 3,626 for placebo. Cost per patient who already had heart failure: US$ 3,392 for bisoprolol and US$ 4,228 for placebo.	69	65
[28] ([25])	SC-UK	UK. Cost per patient: £716.7 for bisoprolol and £725.8 for standard regimen.	1998	Cost per patient: US$ 1,103 for bisoprolol and US$ 1,117 for standard regimen.	89	80

Ref	Type	Original results taken from the paper	Price year	Results converted using the PPP rates	% Score (overall)	% Score (sub-list)
[13]	MS	France: Bisoprolol is a dominant strategy with a cost per patient of 31,762 FF (vs 35,009 for placebo and £4,722 for UK). UK: CER = £3,000/per additionnal patients alive (or £5,500/per life year saved for bisoprolol.	1998	France: Bisoprolol is a dominant strategy with a cost per patient of US$ 4,891 (vs US$ 5,391 for placebo and US$ 7,265 for the UK). UK: CER = US$ 4,615/per additionnal patients alive (or US$ 8,462 per life year saved for bisoprolol.	67	50
[30]	MS	Data pooled (Italy, Spain, France, Switzerland UK and Sweden)Outcomes were not significantly different. The average cost per episode free day for formoterol and salmeterol respectively was $10.60 and $12. Mean cost per patient reaching a clinically relevant improvement in quality of life was $1,602.18 and $1,825.13.	1995		93	92
[27]	MS	In the 1,080-day model, the cost per LYG with exemestane over megestrol was euro 11,169 in Australia, euro 6,911 in Belgium, euro 6,966 in France, euro 1,353 in Germany, euro 10,638 in Italy, euro 13,016 in The Netherlands, euro 7,806 in Spain, and euro 11,733 in the UK.	1999	Cost/LYG: US$ 7,181 for France and US $ 10,476 for the UK	4	18
[23]	MS	Cost/QALY's gained: France: $89,279 (563,710 FF). UK: $176,075 (£103,145).	1988	PPP rates not available prior to 1990.	46	38
[21]	MS	Cost of standard therapy = 2 389 FF ($US 398) for France and £380 ($US 627) for the UK. Cost of latanoprost therapy = 2 087 FF ($US 348) and £307 ($US 507). Cost of dorzolamide therapy = 2 305 FF and £324. Cost of therapy combining timolol and pilocarpine = 2 305 FF and the one with brimonidine = £325.	1997	Cost of standard therapy = US$ 368 for France and US$ 603 for the UK. Cost of latanoprost therapy = US$ 321 for France and US$ 487 for the UK. Cost of dorzolamide therapy = US$ 355 for France and US$ 514 for the UK. Cost of therapy combining timolol and pilocarpine = US$ 351 for France. Cost of therapy combining timolol and brimonidine = US$ 516 for the UK.	64	75
[10]	MS1-F	France: Cost per death avoided = 25,849 FF	1997	Cost per death avoided = US$ 3981.	75	85
[15]	MS	Costs avoided = 233 million Euros for France and 178 million Euros for the UK	2000	248 millions for France and 167 millions for the UK	57	56
[35]	SC-UK	Total 2-year costs (coronary artery disease present/absent): strategy 1: £4,453/£710; strategy 2: £3,842/£478; strategy 3: £3,768/£574; strategy 4: £5,599/£1,475; user centres: £5,563/£623; non-user centres: £5,428/£916. Strategy 1 = exercise ECG/coronary angiography; Strategy 2 = 0exercise ECG/MPI/coronary angiography; Strategy 3 = MPI/coronary angiography; Strategy 4 = coronary angiography.		Total 2-year costs (coronary artery disease present/absent): strategy 1: 6958/1109 US$; strategy 2: 6003/747 US$; strategy 3: 5888/897 US$; strategy 4: 8748/2305 US$; users: 8692/973; non-users: 8481/1431.	80	90

Ref	Type	Original results taken from the paper	Price year	Results converted using the PPP rates	% Score (overall)	% Score (sub-list)
[2]	MS1-UK	Cost to heal one ulcer: £1,425.97 for the Hydrofibre group and £1374.61 for the Alginate group. In incremental cost-effectiveness terms: (£1,425.97 - £1,374.61)/(6-2) = 12.84 (derived from the results by the abstractor).	1995	Cost to heal one ulcer: 2194 US$ for the Hydrofibre group and 2115 US$ for the Alginate group. In incremental cost-effectiveness terms: 2194 - 2115)/(6-2) = 19.75 (derived from the results by the abstractor).	72	83
[32]	MS-p	Mean cost between Belgium, France, Netherlands and UK (per country data not available) Total mean cost: Propofol only group at $31.9 compared to $19.70 in the propofol-sevoflurane group and $18.8 in the sevoflurane only group.	NG		66	91
[4]	MS-p	The cost to achieve an improvement in walking distance of 1 metre was ECU 35.6 for the LTP group and ECU 9.45 for the STP group	NG	No price year	63	71
[24]	MS1-F	France only: Dominant strategy for the group D+R- (seropositive donor/ seronegative recipient) and CER = 62,429 FF for group R+ (seropositive recipient)	1998	Dominant strategy for the group D+R- (seropositive donor/ seronegative recipient) and CER = US$ 9,613 for group R+ (seropositive recipient).	87	75
[20]	MS-p	When priced at zero, basiliximab therapy reduced the costs by Can$4,554 per patient. As long as basiliximab costs less than this amount it is a dominant treatment. When the costs were calculated for an additional 4 years, the costs of basiliximab therapy per patient were Can$11,908 less than for a control group patient. Thus, provided that basiliximab therapy costs less than this, it is the dominant treatment.	1999	When priced at zero, basiliximab therapy reduced the costs by US$ 3,827 per patient. As long as basiliximab costs less than this amount it is a dominant treatment. When the costs were calculated for an additional 4 years, the costs of basiliximab therapy per patient were US$ 10,007 less than for a control group patient. Thus, provided that basiliximab therapy costs less than this, it is the dominant treatment.	67	67

Notes: MS = multi-country studies with cost valuations for both the UK and France; MS1-UK, MS1-F = multi-country economic evaluation in which the costing was performed for only one country, in this case either the UK or France; MS-p = multi-country study for which costing was based on pooled data from a number of countries that included both the UK and France; SC-UK, SC-F = independent study carried out in either the UK or France but on the same health technology; NG = not given.

Appendix 2 Differences in results between the UK and France for relevant studies (mulit-country)

Ref	Type	Difference between UK and France	Price effect?	Sources of differences
[1]	MS	Invasive pneumococcal disease: Cost/Qaly: FR>UK - 2,6%. Pneumococcal pneumonia: Cost/QALY: FR>UK - 700%	Cost measured with a model: country-specific data. Not only a price effect.	ALOS: FR>UK (13.1 vs 7.8). Mortality rate: FR>UK (25.8 vs 37.9).
[5]	MS	UK>FR - 24%	Country-specific cost data. Not only a price effect.	Ratio almost equal but total cost multiplied by 2 for France compared to UK. Not enough data to identify the source of differences.
[12]	MS	Ciclo: UK>FR - 200%. Amo: UK>FR - 200%. Itra: FR>UK - 11%. Terb: UK>FR - 70%. Griseo: UK>FR - 130%	Price effect	Cost of the management of onychomycosis higher in UK (approximately x2) for all drugs except Itraconazole. Not enough data to identify the source of differences.
[14]	MS	1) UK>FR - 19%. 2) UK>FR - 5.2%. 3) UK>FR - 16%. UK>FR - 17%.	Same benefit measure used for each country. Based on independent local costing studies conducted in each country and were provided by Merck.	Not enough data.
[19]	MS	UK>FR - 118%.	Price effects.	
[31]	MS	UK>FR - 70%.	Standard treatment algorithms for each country. Adjustments according to local practice.	
[34]	MS	Omep 20 mg: FR>UK - 7%. Omep 10 mg: FR>UK - 5%. RANIT: FR>UK - 5.5%.	Price effect.	
[13]	MS		Country specific cost data. Not only a price effect.	Cost of cardiax transplantation (UK>FR) + Hospital costs.
[30]	MS			
[27]	MS	UK>FR - 46%	Country specific cost data. Not only a price effect.	Not enough data to identify the source of differences.
[23]	MS	UK>FR - 97%	Country specific cost data. Not only a price effect	Annual epoetin cost per patient (US$ 9587 vs US$ 5769) and number of units transfused.
[21]	MS	1) UK>FR - 64% 2) UK>FR - 52% 3) UK>FR - 45% 4) UK>FR - 47%	Country specific cost data. Not only a price effect.	One of the main differences identified is the cost per patient of the tests (UK>FR). Not enough data to isolate the price effect and identify the source of differences.
[15]	MS	FR>UK - 49%	39,000 cases avoided in France versus 31,000 in the UK. Costs per event higher in France (almost twofold for AMI, CHF and Stroke).	

Appendix 3 Distribution of scores according to the sub-checklist items on generalisability

Range	0-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	80-90%	90-100%	Mean (s.d)
No. of questions with the mark	**0**	**0**	**1**	**4**	**4**	**6**	**7**	**7**	**8**	**6**	**0.696 (0.158)**
Ques. Heading (No.)											
Q (2) Question									**Q2** (28/0)[1]	**Q1*** (28/0)	**0.938 (0.213)**
HT (2) Health technology								**HT2**(28/0)**HT1**(26/2)			**0.750 (0.382)**
SE (2) Setting									**SE1**(28/0)	**SE2**(28/0)	**0.911 (0.252)**
P (1) Perspective								**P1**(28/0)			**0.750 (0.443)**
SP (4) Study Population					**SP2** (27/1) **SP3** (21/7) **SP4** (20/8)					**SP1**(28/0)	**0.599 (0.443)**
M (2) Model							**M2**(12/16)	**M1**(12/16)			**0.667 (0.373)**
E (8) Effectivness				**E6** (7/21)	**E5** (11/17)	**E3** (21/7) **E8** (27/2)	**E1** (18/10) **E2** (18/10)	**E7**(27/1)	**E4** (18/10)		**0.610 (0.436)**
B (5) Benefit						**B5** (25/3)	**B4** (7/21)		**B2**(10/18) **B1** (28/0)	**B3**(11/17)	**0.741 (0.386)**
C (11) Cost				**C6** (28/0) **C11**(24/4)		**C5** (28/0) **C10**(15/13)	**C4** (28/0)	**C2** (28/0) **C8** (28/0)	**C1** (28/0) **C3** (28/0) **C7** (28/0)	**C9*** (28/0)	**0.691 (0.415)**
D (4) Discount			**D4** (6/22)	**D1** (11/17)		**D2** (12/16)				**D3*** (6/22)	**0.500 (0.493)**
S (1) Variability							**S1** (22/6)				**0.614 (0.299)**
O (1) Generalis.							**O2** (28/0)				**0.607 (0.386)**