Appendix

Methodological Framework

At first, as part of the problem structuring phase, the decision problem and the aims of the analysis are defined, and the relevant decision-makers and other key stakeholders are identified. Next, as part of the model building phase, objectives and/or relevant criteria are identified in order to reflect decision-makers' goals and areas of concern, and attributes are selected to operationalise the criteria. In addition, under the same phase, selection of the alternative options takes place and evidence on their performance across the selected criteria is identified. Following that, under the model assessment phase, the performance of options against the criteria is assessed (i.e. scoring) and criteria are weighted according to their relative importance (i.e. weighting). Subsequently, as part of the appraisal phase, scores and weights are combined in order to produce overall WPV scores (i.e. aggregation), taking the form of a value index. In combination with sensitivity analysis, the results are examined and their robustness is analysed. Finally, as part of action planning, the outcome of the analysis can be used to inform policy making, relating to resource allocation and coverage decisions.

Evidence Considered and Alternative Treatments Compared (Model Building)

As part of NICE TA255 [40], for the case of cabazitaxel in combination with prednisone, NICE primarily considered clinical evidence coming from one phase III, randomised, open-label, multicentre trial (*TROPIC*) investigating the use of cabazitaxel plus prednisone (or prednisolone) compared to mitoxantrone plus prednisone (or prednisolone) in men with hormone-refractory metastatic prostate cancer. Patients had to be aged over 18 years with an Eastern Cooperative Oncology Group (ECOG) performance score of 0–2, and with evidence of disease progression during or after completion of docetaxel-containing treatment [33]. The same clinical trial was used by TLV as part of a health economic exercise (no formal appraisal).

As part of TA259 [38], the decision problem considered whether treatment with abiraterone plus prednisolone was clinically effective compared with mitoxantrone (with or without prednisolone) or best supportive care for castration-resistant metastatic prostate cancer previously treated with a docetaxel-containing regimen. NICE primarily considered clinical evidence coming from a phase III, placebo-controlled, randomised, double-blind, multicentre trial (*COU-AA-301*), investigating the use of abiraterone in combination with prednisone (or prednisolone) versus placebo in combination with prednisone (or prednisolone), in men whose disease had progressed on or after docetaxel therapy [34].

Patients were aged over 18 years, with an Eastern Cooperative Oncology Group (ECOG) performance score of 0–2. A similar decision problem was adopted in TLV TA4774/2014 for the case of abiraterone in combination with prednisolone versus prednisolone on its own for patients who had received docetaxel or comparable chemotherapy, with clinical evidence coming from the *COU-AA-301* trial [42].

As part of TA316 [39], for the case of enzalutamide NICE primarily considered clinical evidence coming from a phase III randomised double-blind placebo-controlled study (*AFFIRM*) which investigated the use of enzalutamide plus best supportive care[1] (i.e. with or without the use of prednisone or other glucocorticoids) compared with placebo plus best supportive care [35]. Eligible patients were aged over 18 years, with metastatic hormone-relapsed prostate cancer who had previously received 1 or 2 cytotoxic chemotherapy regimens, at least 1 of which contained docetaxel. Patients who had received abiraterone or treatment with any other investigational agents that block androgen synthesis were excluded. A similar decision problem was adopted in TLV TA2775/2013 for the case of enzalutamide versus best supportive care for patients who had progressed during or after docetaxel treatment, with clinical evidence base from the *AFFIRM* study [41].

In addition, as part of NICE TA316 evidence for abiraterone plus prednisone from the COU-AA-301 trial was also considered in order to indirectly compare enzalutamide versus abiraterone (plus prednisone) using placebo as a common comparator whereas TLV TA4852/2014 used the same pivotal trials to compare enzalutamide versus abiraterone, either (1) when treatment with hormonal therapy has not worked or when treatment has not worked in men without symptoms or with only mild symptoms that do not need chemotherapy yet (i.e. pre-chemotherapy), or (2) adult men with progressive disease during or after docataxel-based chemotherapy (i.e. post-chemotherapy); none of these two scopes were considered.

Setting Attribute Ranges and Reference Levels (Model Building)

For the case of clinical therapeutic attributes, "lower" reference levels were based on BSC figures, coming from the median of the respective placebo arm of the *AFFIRM* trial, with the exception of the HRQoL attribute (EQ-5D index score) that was based on the utility of stable disease with no treatment coming from past NICE TAs [38, 40]. The "higher" reference levels were derived by adding a 20% absolute improvement to the performance level of the best performing option, besides for the case of the HRQoL attribute (EQ-5D index score) that was based on the general Swedish population [51]. The rationale was to design a hybrid type of value scale possessing characteristics from both "local" and "global" reference levels [60],

[1] Best supportive care in AFFIRM could include radiopharmaceuticals, analgesics, bisphosphonates, hormonal therapies, corticosteroids, and radiotherapy

reflected respectively by "satisfactory" performance (proxied by BSC) and "ideal" performance (proxied by a 20% improvement among the options considered), corresponding to the 0 and 100 anchor levels of the value function scale respectively. This could offer a flexibility margin to be able to incorporate the performance of future improved options with options performing better than the satisfactory level scoring more than 0. Consequently two reference levels within the attribute range were defined in most cases: i) the "lower" reference level (x_l) (i.e. BSC-based satisfactory performance), acting on the same time also as the minimum limit of the attribute range (x_*); and ii) the "higher" reference level (x_h) (i.e. 20% better than the best performing option), acting on the same time as the maximum limit of the attribute range (x^*) to give x_*= x_l < x_h = x^*.

Similar but reverse logic was used for setting the reference levels in the "treatment discontinuation" attribute of the safety cluster; the "lower" reference level was defined to be equal to the BSC (i.e. placebo) arm of the *AFFIRM* trial. However, contrary to the logic adopted so far for the therapeutic impact criteria, the "higher" reference level was not set equal to 20% worse than the best performing option (because the lower the performance the higher the value), but rather equal to the minimum natural limit of the attribute scale (i.e. 0%) which was regarded as an "ideal" level. In turn, the minimum limit of the scale was derived by worsening the performance of the worst performing treatment option by 20%. A similar approach was used for setting the reference levels of the qualitative "contraindications" attribute, defining the "higher" reference level to be equal to the maximum (i.e. most attractive) limit of the attribute scale (i.e. no known contraindications) and the "lower" reference level equal to the minimum (i.e. least attractive) limit of the attribute scale, based on the performance of the alternative options therefore acting as reference levels of a "local" scale.

For the innovation attributes, the "higher" reference level was set either equal to 20% better than the best performing option for the case of natural quantitative attributes (e.g. number of new indications for which the technology is investigated in a given clinical development stage), or equal to the maximum limit of the scale for the case of constructed qualitative attributes (e.g. the existence of any special instructions, the technology's relative market entrance in regards to its ATC Level), reflecting a "global" versus "local" scaling approach respectively. Given that the BSC performance was irrelevant to be used as satisfactory level in the innovation attributes, and the fact that any efforts to derive a "satisfactory" level would by definition be subjective in nature, the minimum limit of the scale for each attribute was used as a "lower" reference level. Therefore the "lower" reference level was based on the worst performance plausible as inferred from the lowest limit of the scales, both for the case of natural quantitative attributes (e.g. 0 number of new indications for which the technology is investigated in a given clinical development stage), and the case of

constructed qualitative attributes (e.g. worst possible combination of special instructions, 5th entrance at an ATC level).

For the socioeconomics attribute (impact on direct costs), the "higher" reference level was based on the BSC's impact on cost (i.e. £0 impact on costs), given that by definition impact on costs for all treatment options are incremental to BSC, and the "lower" reference level was derived by adding a 20% absolute increment to the worst performing option (i.e. to the one with the biggest impact on costs).

Table A1: Pre-workshop attribute reference levels and basis of selection

Cluster	Attribute name	Attribute metric	Lower level	Basis	Higher level	Basis
THERAPEUTIC IMPACT	Overall survival	months	13.6	BSC	22.1	20% higher than the best performing option
	Health related quality of life	utility (EQ-5D)	0.72	Utility used for progressive disease in TA259	0.82	Utility scores of general population
	Radiographic tumour progression	months	2.9	BSC	10.6	20% higher than the best performing option
	PSA response	% patients	1.5	BSC	64.8	20% higher than the best performing option
SAFETY PROFILE	Treatment discontinuation (% of patients)	% patients	10	BSC	0	Maximum limit of the scale
	Contra-indications	types of contra-indications	Hypersensitivity + hepatic impairment + low neutrophil counts	Minimum limit of the scale	No contra-indications	Maximum limit of the scale
INNOVATION LEVEL	ATC Level 1	relative market entrance	5	Minimum limit of the scale	1	Maximum limit of the scale

	ATC Level 2	relative market entrance	5	Minimum limit of the scale	1	Maximum limit of the scale
	ATC Level 3	relative market entrance	5	Minimum limit of the scale	1	Maximum limit of the scale
	ATC Level 4	relative market entrance	5	Minimum limit of the scale	1	Maximum limit of the scale
	ATC Level 5	relative market entrance	5	Minimum limit of the scale	1	Maximum limit of the scale
	Phase 1	number of new indications	0	Minimum limit of the scale	10	20% higher than the best performing option
	Phase 2	number of new indications	0	Minimum limit of the scale	16	20% higher than the best performing option
	Phase 3	number of new indications	0	Minimum limit of the scale	2	20% higher than the best performing option
	Marketing authorisation	number of new indications	0	Minimum limit of the scale	1	20% higher than the best performing option
	Delivery Posology	types of delivery system & posology combinations	Oral, every day - one off + IV, every 3 weeks - 1 hour*	Minimum limit of the scale	Oral, every day - one off*	Maximum limit of the scale

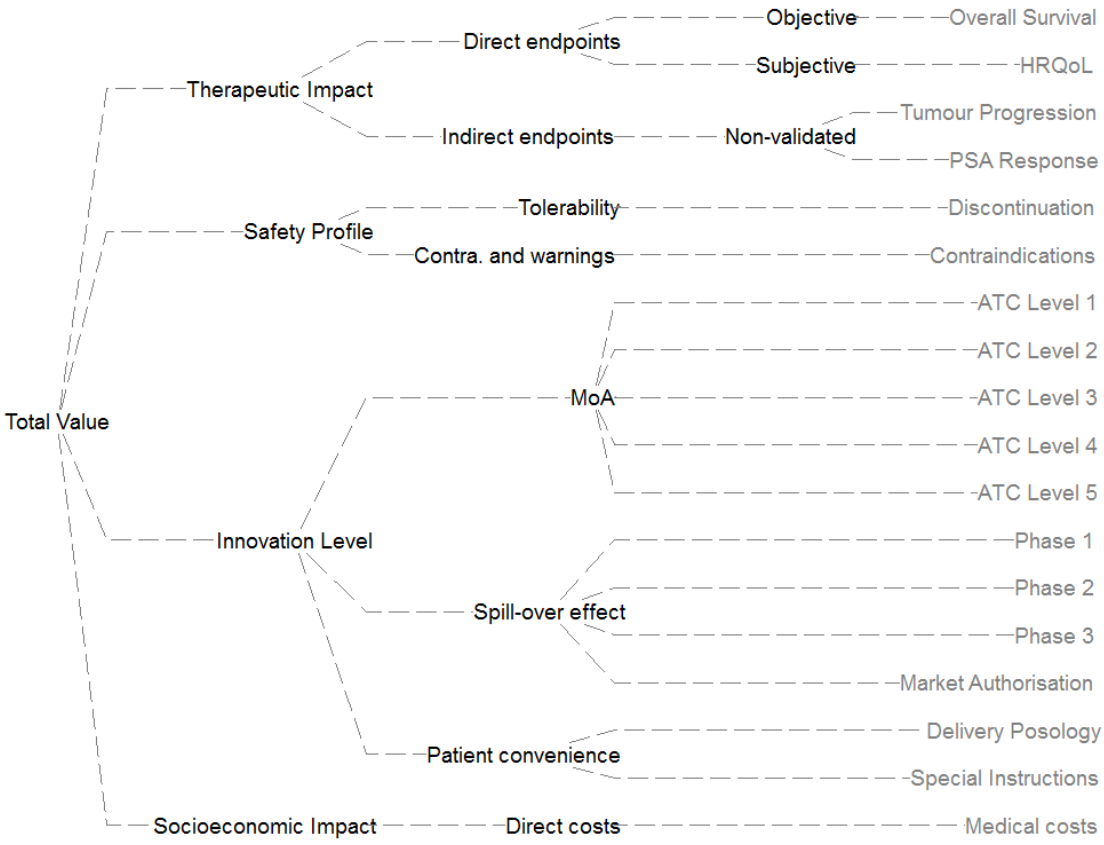
	Special instructions	types of special instructions	No food + concomitant and/or pre-medication*	Minimum limit of the scale	None*	Maximum limit of the scale
SOCIO-ECONOMIC IMPACT	Medical costs impact	GBP (£)	10,000	20% higher than the worst performing option (rounded up)	0	BSC

Decision Conference (Model Assessment and Appraisal)

In terms of the decision-aiding methodology used, the author acted as an impartial facilitator with the aim of enhancing content and process interaction, while refraining from contributing to the content of the group's discussions, essentially guiding the group in how to think about the issues but not what to think [62]. In terms of facilities, the room of the workshop had a Π-shaped meeting table for all the participants to have direct eye-to-eye contact, with an overhead projector screen surrounded by whiteboards. The M-MACBETH software was operated using a laptop, the screen of which was connected to the projector.

The workshop lasted two half-days, three to four hours each, with a short coffee break around the middle of each session. In the first day, the workshop started with an overview of the MCDA methodology adopted and the description of the value tree. The preliminary version of the value tree (Figure A1) was then presented and analysed cluster by cluster. At the beginning of each cluster the value tree was validated; the various criteria were explained, followed by a group discussion relating to their relevance and completeness. As a result of this iterative process, some of the criteria were excluded because they were perceived as irrelevant or non-fundamental. Then, value functions were elicited for the different criteria and relative weights were assigned within the clusters. Finally, relative weights were assigned across clusters, enabling the calculation of the options' overall WPV scores.

Figure A1: Preliminary value tree for metastatic prostate cancer (pre-workshop)*



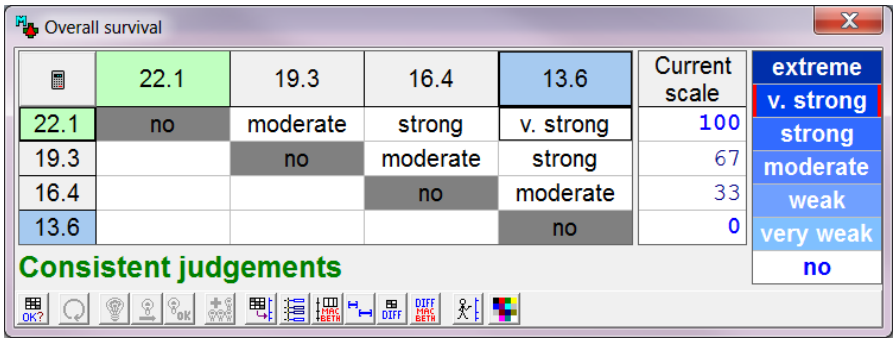
Abbreviations: Contra. = Contraindications; MoA = Mechanism of action; HRQoL = Health related quality of life; PSA = Prostate-specific Antigen; ATC = Anatomical therapeutic chemical;
*Image produced using the Hiview3 software version 3.2.0.4

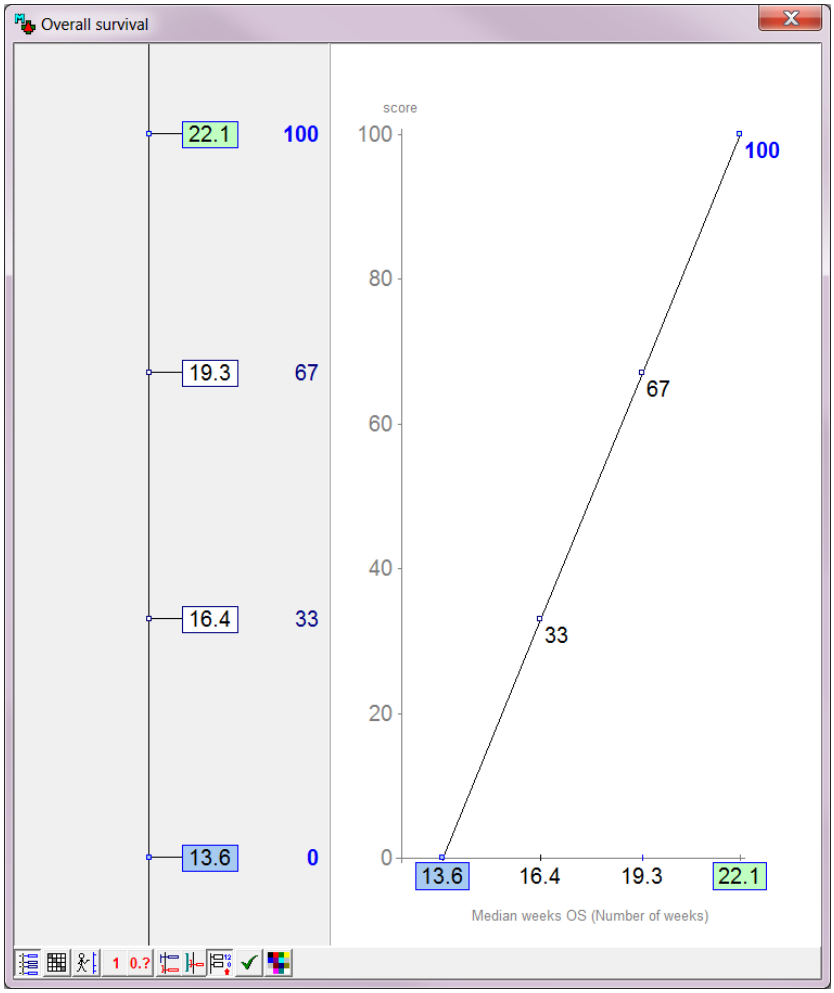
MCDA Technique (Model Assessment and Appraisal)

MACBETH uses seven semantic categories ranging between "no difference" to "extreme difference", in order to distinguish between the value of different attribute levels. Based on these qualitative judgements of difference and, by analysing judgemental inconsistencies, it facilitates the move from ordinal preference modeling, a cognitively less demanding elicitation of preferences, to a quantitative value function. The approach has evolved through the course of theoretical research and real world practical applications, making it an interactive decision support system that facilitates decision-makers' communication. An example of the type of questioning being asked would be "What do you judge to be the difference of value between x' and x'' ?" where x' and x'' are two different attribute levels of attribute x, across the plausible range (i.e. $x_* \leq x', x'' \leq x^*$). The underlying conversion of the MACBETH value judgements into a value scale for the Overall Survival attribute (Figure

A2), followed by the respective scoring of the alternative treatment options as performed by the M-MACBETH software is described below as an illustrative example.

Figure A2: Example of value judgements matrix for the Overall Survival attribute and its conversion into value functions.*





Caption: In the Overall Survival example, the question asked was the following: "*What do you judge to be the difference of value between 13.6 months OS and 22.1 months OS? No difference, very weak, weak, moderate, strong, very strong, or extreme?*" Once a consensus was reached, the next question came along: "*What do you judge to be the difference of value between 16.4 months OS and 22.1 months OS? No difference, very weak, weak, moderate, strong, very strong, or extreme?"* The same process was followed until value judgments for all the different combinations of attribute levels were elicited, filling in the different rows from the right-hand side (i.e. lower range) to the left-hand side (i.e. higher range). *Image produced using the M-MACBETH (beta) software version 3.0.0

Step 1: Input of MACBETH judgements (starting point): these are the value judgements provided by the group (via consensus), acting as the input for the M-MACBETH software.

	22.1	19.3	16.4	13.6
22.1		moderate	strong	v.strong
19.3			moderate	strong
16.4				moderate
13.6				

Extreme	6
Very strong	5
Strong	4
Moderate	3
Weak	2
Very weak	1

Step 2: Initial conversion of the diagonal MACBETH judgements into value scores: the initial value scale for the six categories of "Very weak" – "Extreme" corresponds to value scores 1 – 6 respectively.

	22.1	19.3	16.4	13.6
22.1		moderate 3	strong	v.strong
19.3			moderate 3	strong
16.4				moderate 3
13.6				

Extreme	6
Very strong	5
Strong	4
Moderate	3
Weak	2
Very weak	1

Step 3: Conversion of the v(22.1)–v(16.4) "Strong" value judgement entails that:

v(22.1)–v(16.4) = v(22.1)–v(19.3) + v(19.3)-v(16.4),

v(22.1)–v(16.4) = 3 + 3,

v(22.1)–v(16.4) = 6

	22.1	19.3	16.4	13.6
22.1		moderate 3	strong 6	v.strong
19.3			moderate 3	strong
16.4				moderate 3
13.6				

Extreme	6
Very strong	5
Strong	4
Moderate	3
Weak	2
Very weak	1

Step 4: Therefore, the initial value score for the "Strong" category has to be changed accordingly. Similarly, the "Very strong" and "Extreme" categories have to also accommodate this change, in order to preserve the ordering of the categories.

	22.1	19.3	16.4	13.6
22.1		moderate 3	strong 6	v.strong
19.3			moderate 3	strong
16.4				moderate 3
13.6				

Extreme	8
Very strong	7
Strong	4 - 6
Moderate	3
Weak	2
Very weak	1

Step 5: Next conversion of the second diagonal of MACBETH judgements into value scores: conversion of the v(19.3)–v(13.6) "Strong" value judgement difference is also equal to 6. In other words (similarly to Step 3), conversion of the v(19.3)–v(13.6) "Strong" value judgement entails that:

v(19.3)–v(13.6) = v(19.3)–v(16.4) + v(16.4)-v(13.6),

v(19.3)–v(13.6) = 3 + 3,

v(19.3)–v(13.6) = 6

	22.1	19.3	16.4	13.6
22.1		moderate 3	strong 6	v.strong
19.3			moderate 3	strong 6
16.4				moderate 3
13.6				

Extreme	8
Very strong	7
Strong	4 - 6
Moderate	3
Weak	2
Very weak	1

Step 6: Conversion of the v(22.1)–v(13.6) "Very strong" value judgement difference entails that: v(22.1)–v(13.6) = v(22.1)–v(19.3) + v(19.3)-v(16.4) + v(16.4)-(13.6)

v(22.1)–v(16.4) = 3 + 3 + 3

v(22.1)–v(16.4) = 9

	22.1	19.3	16.4	13.6
22.1		moderate 3	strong 6	v.strong 9
19.3			moderate 3	strong 6
16.4				moderate 3
13.6				

Extreme	8
Very strong	7
Strong	4 - 6
Moderate	3
Weak	2
Very weak	1

Step 7: Therefore, the value scores for the "Very strong" category has to be changed accordingly. Similarly, the "Extreme" category has to also accommodate these changes.

	22.1	19.3	16.4	13.6
22.1		moderate 3	strong 6	v.strong 9
19.3			moderate 3	strong 6
16.4				moderate 3
13.6				

Extreme	10
Very strong	7 - 9
Strong	4 - 6
Moderate	3
Weak	2
Very weak	1

Step 8: Final value scores will act as benchmarks for the value scale (i.e. value function) based on which the alternative treatment options will be scored.

	22.1	19.3	16.4	13.6
22.1		moderate 3	strong 6	v.strong **9**
19.3			moderate 3	strong **6**
16.4				moderate **3**
13.6				0

Extreme	10
Very strong	7 - 9
Strong	4 - 6
Moderate	3
Weak	2
Very weak	1

Step 9: The MACBETH value scale is normalised into a normalised 0-100 value scale.

	MACBETH value scale	Normalised value scale
22.1	9	100
19.3	6	66.67
16.4	3	33.33
13.6	0	0

Step 10: Using the normalised value scale, the respective performance of the alternative treatments options is converted into value scores.

Treatment Options	Performance (months)	Value scores
Abi	15.8	26.2
Caba	15.1	17.9
Enza	18.4	56.3

Following the elicitation of value functions, criteria baseline weights can be elicited. Questions of direct importance for a criterion such as "*How important is a given criterion*?" are known to be as one of the most common mistakes when making value trade-offs because they are assessing them independent of the respective attribute ranges [70]. In contrast, indirect weighting techniques that assess value trade-offs in tandem with the respective ranges of attributes should be employed. For example, the quantitative swing weighting technique asks for judgments of relative value between 'swings' (i.e. changes from standard lower level x_* to higher reference level x^* on each x-th attribute) taking the form "*How would you rank the relative importance of the criteria, considering their attributes ranges relative to 100 for the highest-ranked criterion considering its range*?". Each swing, i.e. a relative change from a lower attribute level to a higher attribute level, is valued between 0 and 100, with the most valuable swing anchored as 100 [28]. Normalised weights are then calculated, as a proportion of each swing weight, so the normalised weights are summed up to 100%. Instead, relative attribute weights were calculated using an alternative qualitative swing weighting protocol, by using the MACBETH procedure to elicit the differences in attractiveness between the lower and higher reference levels of the different attributes, initially at individual level and then at criteria cluster level (i.e. by considering multiple attribute swings on the same time) [69].

Finally criteria preference value scores and the respective weights can be combined together through an additive aggregation approach as described in equation 2 (if the adequate conditions of complete and transitive preferences are met as well as multi-attribute preferential independence conditions – see [28]).