The additive utility assumption of the QALY model revisited

Afshin Gandjour a,b,c,⁎, Amiram Gafni d

a The James A. Baker III Institute for Public Policy, Rice University, Houston, USA
b Institute of Health Economics and Clinical Epidemiology, University of Cologne, Cologne, Germany
c Pennington Biomedical Research Center/Louisiana State University, Baton Rouge, USA
d Centre for Health Economics and Policy Analysis and the Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada

A R T I C L E   I N F O

Article history:
Received 6 October 2009
Accepted 3 November 2009
Available online 11 November 2009

Keywords:
Cost-effectiveness analysis
Quality-adjusted life years

A B S T R A C T

Quality-adjusted life years are valid representations of the preferences of individuals for health outcomes only under a set of restrictive assumptions. One of the key assumptions is additive utility independence (AUI). Recently, Bleichrodt and Filko [Bleichrodt and Filko, 2008. Journal of Health Economics 27 (5), 1237–1249] presented a new test for AUI, the test for generalized marginality (GM). Based on a student survey showing that violations observed at an individual level cancel out at the group level they concluded that use of the QALY model for economic evaluations in health care is supported. In this comment we argue that this conclusion is not warranted for 2 independent reasons: (i) the GM test is not sufficient to claim AUI both at an individual and group (i.e., aggregated) level and (ii) the student survey is not appropriate to make generalized statements about preferences at the population level.

© 2009 Elsevier B.V. All rights reserved.

Quality-adjusted life years (QALYs) are valid representations of the preferences of individuals for health outcomes (i.e., life time health profiles) only under a set of restrictive assumptions. Piskin et al. (1980) were the first to define the conditions for the case of a chronic health state (HS) (i.e., the case where an individual will be in the same HS for the remaining years of life) and for an individual who is an expected utility maximizer (i.e., a von Neumann-Morgenstern-type individual). It was shown (Mehrez and Gafni, 1991; Broome, 1993) that this additional condition must be satisfied – additive utility independence (AUI) – for the more general case of a life time health profile (i.e., experiencing more than one HS during the remaining years of life). When AUI holds, the utility of a life time health profile equals the sum of single-period utilities. The AUI assumption in addition to other assumptions allows these single-period utilities to be measured using conventional preference-based questionnaires (e.g., time trade-off and standard gamble methods).

While the term AUI is used by many authors in the health economics field (e.g., Drummond et al., 2005, p. 156; Johannesson, 1996, p. 202), the meaning is not consistent. We use a definition of AUI that is in line with Johannesson (1996, p. 202): “the utility of a health status in one period is independent of the health status in all other periods”. Johannesson refers to Broome (1993) when explaining AUI. Broome himself does not use this term, but talks about “strong separability”. Broome (1993, p. 152) states that the utility function underlying the QALY model has an additively separable form based on strongly separable preferences. Strong separability means “that a person’s preferences about the qualities of her life in any particular group of years are independent of the qualities of her life in other years”. It is interesting to note that for Broome it is “dubious” that “in the person’s preferences, qualities of life at different times are strongly separable”.

In a recent paper in this journal Bleichrodt and Filko (2008) (B & F) present a new test for AUI, the test for generalized marginality (GM). They argue that this test avoids some of the pitfalls of previous tests. They applied the test in a survey of 60 students and show that while AUI is violated at the individual level by most students; at an aggregated level AUI cannot be rejected at the 5 percent level of significance. In other words, violations observed at an individual level cancel out at the group level. The authors conclude that their result supports the use of the QALY model for economic evaluations in health care.

In the following we argue that this conclusion is not warranted for 2 independent reasons: (i) the GM test is not sufficient to claim AUI both at an individual and group (i.e., aggregated) level; and (ii) the survey is not appropriate to make generalized statements about preferences at the population level. The comment is structured as follows. First, we define AUI formally. Second, we discuss the GM test used by B & F and its limitations with respect to testing for AUI. And third, we discuss limitations of the empirical survey.

⁎ Corresponding author at: Pennington Biomedical Research Center/Louisiana State University, 6400 Perkins Road, Suite G3047, Baton Rouge, LA 70808, USA.
Tel.: +1 225 763 3051; fax: +1 225 763 3009.
E-mail address: afschin.gandjour@pbrc.edu (A. Gandjour).
1. Definition of additive utility independence

Consider, for simplicity, a potential 3-period health profile which consists of health states $a$, $b$, and $c$ and has no preceding HS. The utility of this profile can be described as an additive utility function described in Gandjour (2008):

$$u(a, b, c) = u(a) + u(b) + u(c) + \lambda(a)u(c) + \lambda(b)u(c) + \lambda(c)u(c)$$

$$+ \lambda(a)u(L(b)) + \lambda(b)u(L(c)) + \lambda(c)u(L(c))$$

(1)

where $u$ denotes utility, $\lambda$ is a weighting factor that values future health states from the perspective of prior periods and is $>0$ ($<0$) if future health has a positive (negative) impact on a prior period's utility, and $L$ describes the distribution of health states within a period.

The utility of the 3-period health profile is thus not only determined by expected mean future health, but also by anticipatory feelings about the uncertainty around the expected future health $L$. Furthermore, health states in the second and third period may also be influenced by preceding health states. For example, a preference valuation of future health may depend on present and past health states. The influence of prior health states on subsequent health states can also be captured by $\lambda$. For example, assume that HS $a$ negatively influences HS $b$. This will be captured as a negative $\lambda$ as HS $b$ from the perspective of HS $a$ will receive a lower preference value. In fact, it is known that some individuals with poor health do not want to live beyond a certain time and thus disvalue a continuation of their poor health in the future (Sutherland et al., 1982).

An alternative (but more complex) description would have been to consider the impact of HS $a$ on HS $b$ (as well as the impact of HS $b$ plus $c$) as a separate term in Eq. (1), but to exclude this impact from $\lambda$, i.e., the assessment of HS $b$ from the perspective of HS $a$. Note again that the above formulation holds when HS $a$ is not preceded by prior health states. If HS $a$ was preceded by prior health states, the impact of prior health states on health states $a$, $b$, and $c$ would need to be captured by additional terms. As the latter case does not affect the generalizability of our statements, we do not discuss it further in order to keep the presentation of our arguments simple.

Note that the utility function in Gandjour (2008) is different from a utility function that satisfies AUI as defined in the introduction (i.e., strong separability). While the utility function in Gandjour (2008) is additive, it includes terms ($\lambda$') representing the potential impact of past or future health states on the utility of subsequent or prior health states. That is, it is an additive utility function without necessarily being strongly separable.

AUI or “strong separability” for the utility function described in Eq. (1) implies that

$$\lambda(a)u(b) = \lambda(a)u(c) = \lambda(b)u(c) = \lambda(c)u(b)$$

$$= \lambda(a)u(L(c)) = \lambda(b)u(L(c)) = \lambda(c)u(L(c)) = 0$$

(2)

That is, there is no impact of future health states on the utility of prior health states and no impact of prior health states on the utility of future health states (all $\lambda$'s are zero). It is easy to see from Eq. (2) that when all $\lambda$'s are constant, the AUI condition is not met except for the special case where the constant is zero. That is, when all $\lambda$'s are constant but not zero, the QALY model does not hold.

2. Definition of generalized marginality

GM defined by B & F assumes that the following relationship holds:

$$(p: a_1, b_1, c_1; a_1', b_1', c_1')$$

$$\leftrightarrow (p: a_2, b_2, c_2; a_2', b_2', c_2')$$

(3)

where $p$ denotes probability and $a$, $b$, and $c$ are again 3 health states, resulting in a 3-period health profile. In Eq. (3) there are 2 indifferences, the left one is denoted by $I$ and the right one is denoted by $II$. Each indifference consists of 2 prospects. These 2 indifferences must also be equivalent as denoted by $\leftrightarrow$. The only difference between indifference $I$ and $II$ are mid-period health states ($b$'s): $b_i$ (in indifference $I$) is replaced by $b_{II}'$ (in indifference $II$) and $b_{II}'$ (in indifference $I$) by $b_{II}'$ (in indifference $II$). Two other versions of the GM test were applied, where indifference $I$ and $II$ differ in terms of first-period health states ($a$'s) and third-period health states ($c$'s), respectively.

From the above equivalence it follows that the difference between the first period of indifference $I$ and $II$ must equal the difference between the second period of indifference $I$ and $II$ for the utility function described in Eq. (1) GM tests the following:

$$p \times (u(b_1) - u(b_{II})) + \lambda(a_1)u(b_1)) = u(a_1)u(b_1)) + (1-p) \times (u(b_1') - u(b_{II})) + \lambda(a_1')u(b_1')) - \lambda(a_1)u(b_{II})) +$$

$$+ \lambda(a_1)u(b_{II} - b_{II})) = 0$$

(4)

$$\Rightarrow\lambda(a_1)u(b_1) = \lambda(a_1')u(b_1') - \lambda(a_II)u(b_{II})$$

$$\Rightarrow(1-p) \times (u(b_1') - u(b_{II})) + \lambda(a_1')u(b_1')) - \lambda(a_1)u(b_{II})) +$$

$$+ \lambda(a_1)u(b_{II} - b_{II})) = 0$$

GM thus holds if the value of HS $b$ is the same regardless of HS $a$. Note that in the GM test uncertainty around future health (i.e., $\lambda(a_II)(u(b_II)))$ cancels out.

3. Limitations of testing for generalized marginality

In the following we explain why the GM test is insufficient to claim AUI both at an individual and group (i.e., aggregated) level using as an example the utility function described in Eq. (1). In order for the GM test to confirm the validity of the conventional QALY model, it needs to show that the AUI condition as described in Eq. (2) is met. However, the GM test is not able to do so. While the GM test may provide results that are consistent with an additive utility function as described in Eq. (1), it does not tell us whether the condition specified in Eq. (2) is met. That is, it does not distinguish between an additive function which is not strongly separable and an additive and strongly separable one.

The reason is that GM only considers differences between profiles (between indifference $I$ and $II$), but does not take into account an impact of future or past health states on the utility of subsequent or prior health states. That is, if we look for differences between profiles, $\lambda(a_II)(u(b_II))$ might cancel out, but within a profile $\lambda(a_II)(u(b_II))$ might violate Eq. (2). Therefore, in a situation where the expected mean health in period 2 has a constant but non-zero impact on HS in period 1 (i.e., all $\lambda$'s with regard to the expected mean health are constant), GM might hold but AUI is violated. The constant but non-zero impact can be represented by the additive utility function described in Eq. (1), but it violates Eq. (2).

A potential error in the assumptions of the GM test might have occurred when the GM test by Miyamoto (1988) was applied to health profiles (Bleichrodt and Quigg, 1997). The GM model by
Miyamoto uses distinct outcomes (single states), which are different from health profiles, which have time inextricably bound to them and use multiple states. It seems that one can use Miyamoto’s GM test to demonstrate independence of health profiles as such, but not to show independence of the different states contained in a health profile.

Another situation where the GM test might hold but AUI is violated exists when the individual disvalues only the uncertainty in future health, but not to show independence of the different states contained in them and use multiple states. It seems that one can use Miyamoto’s GM test to demonstrate independence of health profiles as such, and use multiple states. It seems that one can use Miyamoto’s GM test only holds at an aggregated level because individual violations cancel out.

4. Limitations of the empirical survey

A generalized statement about preferences at the population level based on the type and size of participants in the survey and a test that uses a very small number of health profiles is not adequate. Yet, for example, the abstract of the paper makes such statement (“results support the use of QALYs (…) in economic evaluations of health care”). Note that this conclusion is stated in a way that it applies everywhere, not just in the jurisdiction where the study was done.

Reasons for a lack of generalizability are as follows. First, a survey of students seems not representative of the population of the Netherlands. It seems even questionable to state that the student sample is representative of the student population in the Netherlands. At least, we are not aware that the sample was drawn randomly. B & F claim “empirical studies exploring the difference in response patterns between students and samples from the general population observed no significant differences” and cite De Wit et al. (2000) and Bleichrodt et al. (2005) as evidence. However, de Wit compares patients (not students) to the general population and thus does not seem to answer the question. Thus, evidence is limited to only one study which we feel does not support this statement. We think that more than one study that compares a sample of students (N=69) to a sample of the general public (N=208) is needed to support such statement. And even this study does not assess profiles with varying health states over time and thus does not really relate to AUI testing. In addition, the sample size (N=60) seems to be too small to make generalized statement regarding the Dutch population (or the world population – as the authors do not qualify their finding as applying only to the Dutch population).

We are also concerned about the conclusion – AUI holds at an aggregated level, but not at the individual level – as it likely stems from individual violations cancel out. This, however, could well have been just due to chance and thus cannot be generalized (i.e., it does not mean that individual violations will always cancel out within and between other subgroups of the general population). There seem to be important differences between students and other groups in the population. We are concerned, for example, that elderly persons or parents may value declining future health differently than students. In conclusion, the findings that a phenomenon that was rejected at an individual level will hold at a group level at best requires some further exploration and justification before making such a strong statement.

In addition to questioning the generalizability of the sample (i.e., can we generalize from the answers of a sample of students regarding the preference pattern of the general population), we also question whether any generalized statement on AUI can be made based on 3 tests only. That is, we think that even if the survey population was representative of the whole population and AUI appropriately tested, a generalized claim would seem invalid. The reason is that the variety of health profiles is essentially endless and showing AUI for specific profiles cannot rule out AUI violation even for small changes in profiles. This relates again to the finding that the GM test only holds at an aggregated level because individual violations cancel out.

To illustrate this point consider the following example. Imagine a survey population consisting of 2 individuals A and B who have a utility function similar to the one described in Eq. (1). Individual A is optimistic with respect to future health but not anxious with respect to future uncertainty in health and individual B is anxious, but neither optimistic nor pessimistic with respect to future health. Furthermore, assume that both individuals assess the same 3-period health profile consisting of health states a, b, and c, where \( u(a) < u(b) < u(c) \) and \( L(b) = L(c) > 0 \), i.e., uncertainty exists both for health states b and c. For individual B’s negative as he is anxious with respect to future uncertainty in health.

If

\[
\lambda(a)(u(b))_A + \lambda(a)(u(c))_A + \lambda(b)(u(c))_A + \lambda(a)(u(L(b)))_B \\
+ \lambda(a)(u(L(c)))_B + \lambda(b)(u(L(c)))_B = 0
\]

holds, then both individuals show a violation of AUI at the individual level, but accord with AUI at the aggregated level (AUI violations of individuals A and B cancel out). Now consider that both individuals assess a new health profile where health states a and b are the same as above but health c does not have uncertainty. As \( \lambda(a)(u(L(c)))_B \) and \( \lambda(b)(u(L(c)))_B \) become zero, but all other summands stay the same, the above equation does not hold any more, i.e., AUI is also violated at an aggregated level.

Another reason why it seems important to test a broader range of profiles is the considerable number of factors that cast doubts about the validity of the AUI assumption as a generalized phenomenon (e.g., age, marital status/children, educational background of respondents, and duration of health states). Current evidence on the impact of demographic factors on quality of life measurement is limited and non-systematic (see, e.g., Froberg and Kane (1989), who are also cited by Dolan (2000)). Furthermore, we are not aware of any study comparing students and the general population or other demographic groups in terms of violations of QALY assumptions. However, the absence of such studies does not mean evidence for absence of a difference. The burden of proof is to the authors.

While B & F acknowledge that the number health profiles tested in their survey is limited, they still make a strong general statement. They cite tests conducted by Spencer and Robinson (2007) as additional support for their claims. However, Spencer and Robinson did not examine AUI. They tested for utility independence, a weaker condition that does not necessarily imply the conventional QALY model. Therefore, their tests do not support the statement in the abstract that “results support the use of QALY (…) in economic evaluations”. Furthermore, they also used a small sample of students for their empirical test and thus their study suffers from the same problem of generalizability of findings.

5. Discussion

The GM test used by B & F is not sufficient to claim AUI both at an individual and group (i.e., aggregated) level. The reason is that it does not distinguish between an additive utility function which is not strongly separable and an additive and strongly separable utility function. Furthermore, the survey used for the empirical testing is not appropriate to make generalized statements about preferences at the population level. First, the survey population is small (N=60) and is not representative of the population of the Netherlands (or even the students’ population in the Netherlands). Second, showing AUI for specific profiles cannot rule out AUI violation even for small changes in these profiles as well as other profiles.
We agree that it is not easy to show that an assumption (e.g., AUI) holds at a population level. It is much easier to show that an assumption does not always hold as you only need to show one example to prove it. Thus, we think that a student survey with a limited number of profiles may well be suited to disprove an assumption. To prove that an assumption always holds (or holds in the vast majority of cases) requires much more extensive testing, e.g., use of a representative sample of the public and a broad range of profiles.

Leaving aside the various models and tests, we would like to stress the counter-intuitiveness and questionable face validity of the AUI assumption as noted also by Broome (1993) and others. This led us to question the statement by B & F about the group level. AUI means that people are, for example, not optimistic, pessimistic, or anxious about their future health. AUI is not consistent with a behavior where people who are anxious about a decline in their future health may start preventive activities or people with a sad outlook may commit suicide. The way people rate their past and future health may depend on their age and on their disease. Each disease may have a different effect on a person’s outlook. This is why we stated that the number of profiles and the age range of respondents in the B-F paper are not sufficient. Even arguing that some students are, for example, optimistic with regard to their future health and others are pessimistic so that on average students are neither optimistic nor pessimistic needs more testing. To further infer that individuals in the general population (or in any subgroup of the general population) are neither optimistic nor pessimistic (including patient groups with very old people or a sad outlook) is even a bigger leap of faith.

References