NIH Public Access
Author Manuscript
Ann Intern Med. Author manuscript; available in PMC 2012 September 28.

Published in final edited form as:
Ann Intern Med. 2007 January 16; 146(2): 133–141.

Incorporating Quality of Evidence into Decision Analytic Modeling

R. Scott Braithwaite, MD, MSc, **Mark S. Roberts, MD, MPP**, and **Amy C. Justice, MD, PhD**
Yale University School of Medicine and Connecticut Veterans Affairs Healthcare System, New Haven, Connecticut, and University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania

Abstract

Our objective was to illustrate the effects of using stricter standards for the quality of evidence used in decision analytic modeling. We created a simple 10-parameter probabilistic Markov model to estimate the cost-effectiveness of directly observed therapy (DOT) for individuals with newly diagnosed HIV infection. We evaluated quality of evidence on the basis of U.S. Preventive Services Task Force methods, which specified 3 separate domains: study design, internal validity, and external validity. We varied the evidence criteria for each of these domains individually and collectively. We used published research as a source of data only if the quality of the research met specified criteria; otherwise, we specified the parameter by randomly choosing a number from a range within which every number has the same probability of being selected (a uniform distribution). When we did not eliminate poor-quality evidence, DOT improved health 99% of the time and cost less than $100 000 per additional quality-adjusted life-year (QALY) 85% of the time. The confidence ellipse was extremely narrow, suggesting high precision. When we used the most rigorous standards of evidence, we could use fewer than one fifth of the data sources, and DOT improved health only 49% of the time and cost less than $100 000 per additional QALY only 4% of the time. The confidence ellipse became much larger, showing that the results were less precise. We conclude that the results of decision modeling may vary dramatically depending on the stringency of the criteria for selecting evidence to use in the model.

Many clinicians and policymakers view decision analytic modeling as interesting but ultimately esoteric and unhelpful (1). Anecdotally, model users find it difficult to know when they can “trust the model” and often express bewilderment when confronted with data sources from a wide range of study designs, personal opinions, and convergence with the target population. While modelers commonly use methods to consider the random error of individual data sources (2, 3), they seldom model the uncertainty that arises when the evidence for the data is weak or is of questionable applicability to the model’s target population. Because this uncertainty is hard to detect, it may be an insidious source of modeling error.

Similar issues arise in evidence-based reviews of the literature, in which authors combine studies that have different methods and populations to form a summary measure (4). However, these reviews often encounter less skepticism because evidence-based reviews

© 2007 American College of Physicians

Requests for Single Reprints: R. Scott Braithwaite, MD, MSc, Yale University School of Medicine, Veterans Affairs Connecticut Healthcare System, Mailstop 11 ACSLG, 950 Campbell Avenue, West Haven, CT 06516; ronald.braithwaite@va.gov.
Current Author Addresses: Drs. Braithwaite and Justice: Yale University School of Medicine, Veterans Affairs Connecticut Healthcare System, Mailtop 11 ACSLG, 950 Campbell Avenue, West Haven, CT 06516
Dr. Roberts: University of Pittsburgh School of Medicine, 200 Meyran Avenue, 2nd Floor, Pittsburgh, PA 15213.

Potential Financial Conflicts of Interest: None disclosed.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

characterize the quality of each study's evidence (5–7). Readers can then use their own personal thresholds for quality of evidence to decide whether to use the results in their decision making. We sought to emulate the transparency of systematic reviews in the domain of decision analytic modeling.

Several authors have published methods to quantify the uncertainty conferred by weak study designs. These include bias modeling (8) and Bayesian approaches, such as the confidence profile method (9, 10). While these methods differ in their specifics, they require the analyst to identify each bias that may have affected each individual data source, to quantify the impact of that bias, and then to modify each parameter distribution in the model so that it reflects the bias. These methods promise great accuracy, potentially yielding probability distributions that precisely reflect what is known about each parameter. However, these methods are sometimes intractable—it may be impossible to identify and quantify every bias that affects each model parameter. Responding to skepticism about complex decision analytic models that readers cannot understand (1), authors may be reluctant to add another opaque methodologic layer that deals with bias in the evidence. For all their merits, those methods for accounting for biased evidence have not become routine practice.

We sought to illustrate the principle behind adjusting for biased evidence by using a transparent method for incorporating quality of evidence into decision analytic modeling. Our approach does not involve a conceptual advance beyond these previously published methods. Rather, we identify powerful and simplifying heuristics that sacrifice precision to gain transparency and feasibility.

We demonstrate how to incorporate the evidence valuation hierarchy used by the U.S. Preventive Services Task Force (USPSTF) within decision analytic models. The question we chose to evaluate was the cost-effectiveness of directly observed therapy (DOT) for an individual with newly diagnosed HIV infection and a moderately low CD4 cell count (0.1 to 0.2×10^9 cells/L).

Methods

We first discuss how we adapted the USPSTF method to our task of grading the quality of evidence that we might use in our decision model. Next, we describe how we specified a minimally acceptable, or "qualifying," grade of evidence. If the quality of evidence was beneath this grade, we considered it to be too weak for us to use in the model. Finally, we varied our threshold for defining adequate quality of evidence to measure the effect of evidence quality on the results produced by the model.

The basic idea underlying this approach is that when a data source has poor-quality evidence, we should not use it to estimate a model parameter. Instead, we should assume that we know little about the parameter's true value and specify it by using a probability distribution with few embedded assumptions (probability distributions that are "uninformative" or "wide"). A probability distribution is a way of expressing what we know about a number whose true value is uncertain. It is narrow if we are certain and wide if we are uncertain. We chose to use *uniform* distributions (that is, it is equally likely that a parameter has any value within a specified range) to construct our uninformative distributions; however, other types of wide distributions (such as shallow triangular distributions) are consistent with this approach and may also be suitable. It is important to note that we used uninformative distributions only if no data source met our qualifying grade of evidence; otherwise, we based the parameter distribution on the statistical uncertainty of the qualifying data source.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Adapting the Analytic Framework of Evidence-Based Medicine

As recommended by the USPSTF (4), we evaluated quality of evidence according to 3 separate domains of study design: hierarchy of research design, internal validity, and external validity (Table 1). Hierarchy of research design (which we subsequently refer to as *study design*) measures the extent to which a study's design differs from a controlled experiment. The hierarchy ranges from randomized, controlled clinical trial (level 1, most favorable) to expert opinion (level 3, least favorable). *Internal validity* measures how well a study's conclusions apply to the members of the study sample itself (11) (Table 1) and includes determinants of study quality or bias minimization, such as concealment of randomization and loss to follow-up. It is graded as good, fair, or poor. *External validity* measures how well a study's inferences should apply to members of the target population (11) (Table 1) and includes such determinants of generalizability as whether clinical, social, or environmental circumstances in the studies could modify the results from those expected in another clinical setting. The USPSTF did not specify a grading system for external validity, so we simply graded this attribute as high or low. The USPSTF classification scheme treats research design, internal validity, and external validity as largely independent of each other. Therefore, although some would argue that any study with poor internal validity will necessarily have poor external validity, the USPSTF approach preserves the flexibility to rate these 2 attributes independently.

The USPSTF developed its methods to use with experimental variables (that is, interventions). Therefore, it gives lower-quality scores to studies that rely strictly on observation. Decision analytic models commonly include variables that are necessarily observed and are impossible to measure by experiment (for example, cost or features of the natural history of disease, such as the mortality rate due to age-, sex-, and race-related causes). For this reason, we did not want to automatically award low study design grades to studies measuring nonexperimental variables. Therefore, we classified high-quality studies that measured nonexperimental variables as level 1 rather than level 2.

Two of the authors independently classified the design, internal validity, and external validity of 17 data sources (12–25; one unpublished study; expert opinion) for the DOT model by using this adapted classification scheme (Table 2). They disagreed only once and quickly resolved the difference. Some investigators may prefer more robust review procedures, such as those used for meta-analyses (for example, review in duplicate with decision rules for adjudicating discrepancies, blinded review) (26).

Specifying Qualifying Grades of Evidence

Every time we used the model, we specified qualifying grade or grades of evidence for 1 or more of our evidence domains (study design, internal validity, and external validity). A "qualifying" grade of evidence means we could use a data source in our model only if that source met or exceeded the qualifying grade of evidence. For example, if we specified that the model must contain only data from studies using level 1 study design, we used only data sources with level 1 design and excluded studies with weaker design (level 2-1, level 2-2, level 3 [see Table 1]).

Varying the Qualifying Grade of Evidence in Sensitivity Analyses

After performing a base-case analysis in which we used all evidence (regardless of its quality), we performed 4 separate sensitivity analyses on the basis of quality of evidence. Three of them consisted of imposing the strictest evidence criteria for 1 of the 3 evidence domains while using all evidence for the other 2 domains. In the fourth analysis, we simultaneously set all 3 evidence criteria to their strictest levels.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Estimation of Parameter Input Distributions

When more than 1 study met the qualifying grade of evidence, we used the data source with the most statistically precise estimate. Specifically, we used the mean and 95% CI from the study to define the corresponding parameter's distribution. We chose this approach for its simplicity, as well as to reflect common modeling practice. However, an alternative approach would be to perform a formal meta-analysis using all data sources that met the qualifying grade of evidence (assuming that the studies were sufficiently homogeneous to justify combining them).

When data sources for a particular model parameter did not meet the qualifying grade of evidence, we did not base the parameter point estimate on those sources. Instead, we assumed a uniform distribution over a range that was sufficiently wide and inclusive to encompass the likely range of model users' beliefs (that is, prior probability distributions) about the true value of that parameter. To minimize our dependence on assumptions, we specified uniform distributions that were neutral, thereby not favoring any particular direction of effect. For example, when we imposed a strict qualifying grade of evidence for external validity, none of the studies of the effectiveness of DOT met our criterion, so we specified this parameter by a uniform distribution centered on the null effect (DOT had no effect on adherence).

Decision Analytic Model

We constructed a simple decision analytic model (Figure 1) by using standard methods to specify parameters as probability distributions (that is, a range of probabilities in which some probabilities may be more likely than others). In the model, individuals may or may not receive DOT, and DOT may or may not affect whether individuals take HIV treatment. HIV treatment, in turn, influences whether people live or die. We used this simple model solely to illustrate this approach. The model differs completely from the HIV decision model published elsewhere by Braithwaite and colleagues (27–29). We deliberately made it as simple as possible to ensure that the complexity of the model would not be a barrier to understanding the concept of incorporating quality of evidence into a decision model, which applies to any expected value decision model, no matter how complex.

We used a type of model that generates a result from each of many runs in which the model draws a value at random from the probability distributions for each parameter input. Because the parameter values vary from run to run, the results vary from run to run. We used a cost-effectiveness model, so each run generated a value for incremental cost and incremental effectiveness. In accord with accepted practice for presenting model results, we display the results in 2 different ways. First, we show a *confidence ellipse*, which indicates the portion of the cost-effectiveness plane (that is, a 2-dimensional graph of incremental cost and incremental effectiveness) in which 95% of results of a run are likely to occur (30). The confidence ellipse is analogous to confidence intervals. Smaller confidence ellipses correspond to more precise results, and wider confidence ellipses correspond to less precise results. Second, we show an *acceptability curve*, which shows the proportion of observations that fall beneath a range of hypothetical thresholds that society is willing to pay for health benefits (31). A range of $50 000 to $100 000 per quality-adjusted life-year (QALY) is commonly used as a de facto standard for cost-effectiveness. If a high proportion of observations falls below this range, it is evidence in favor of cost-effectiveness. We ran each simulation 1000 times.

Role of the Funding Source

The funding source had no role in the study's design, conduct, or analysis and in the decision to submit the manuscript for publication.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Results

To perform a sensitivity analysis by quality of evidence, we varied each of the 3 evidence criteria (study design, internal validity, external validity) across their ranges, first individually and then collectively. When we imposed no evidence criteria, we used all 17 data sources for the parameter estimates in the model. When we included only studies that met the most selective grade of evidence for study design (level 1), we excluded 4 of the data sources and based our parameter estimates on the remaining 13 studies. When we included only studies that met the most selective grade of evidence for internal validity (good), we excluded 8 of the 17 studies and based our parameter estimates on the remaining 9 studies. Specifying the most selective grade for external validity (high) excluded 12 of the 17 studies, and specifying all 3 evidence criteria simultaneously excluded 14 of the 17 studies (Table 2).

Precision of Model Results

When we imposed no evidence criteria and used all data sources in the model, the 95% confidence ellipse was narrow (Figure 2, *A*), suggesting precise model results. In addition, the area contained within the confidence ellipse was mostly to the right of the $100 000 cost-effectiveness threshold, suggesting that DOT was cost-effective.

When we required that the study design of data sources had to have level 1 evidence, the cost-effectiveness ellipse became much larger (Figure 2, *B*), indicating a substantial increase in uncertainty about model results. The long axis of the ellipse rotated considerably counterclockwise, indicating worse cost-effectiveness.

Requiring a “good” rating for the internal validity of data sources resulted in an intermediate-size cost-effectiveness ellipse compared with requiring a level 1 rating for study design. Nonetheless, the size of the ellipse, and therefore the uncertainty of the model results, increased substantially compared with when we used all data sources (Figure 2, *C*).

When we required a “high” rating for the external validity of data sources, the confidence ellipse became larger than with the other individual evidence criteria, indicating imprecise modeling results. Very few results were on the favorable side of the cost-effectiveness threshold (Figure 2, *D*).

Imposing all 3 evidence criteria simultaneously yielded a confidence ellipse nearly identical to that resulting from setting the external validity criterion to “high.” Few results were on the favorable side of the cost-effectiveness threshold.

Model Results for Cost-Effectiveness

When we imposed no evidence criteria (Figure 3, *A*), 85% of simulation runs were cost-effective, assuming willingness to pay $100 000 per QALY, and 99.0% of all runs showed DOT to have a beneficial effect. The overall incremental cost-effectiveness of DOT was $78 000 per QALY.

When we mandated level 1 study design (Figure 3, *B*), only 17% of simulation runs were cost-effective, although 99.6% of runs showed effectiveness. The overall incremental cost-effectiveness increased to $227 000/QALY.

Mandating good internal validity (Figure 3, *C*) affected the cost-effectiveness of DOT less adversely than did the study design criterion but still resulted in only 20% of runs being cost-effective. Most (89%) runs showed effectiveness. The cost-effectiveness remained

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

unfavorable compared with when we did not impose limitations on evidence criteria ($158 000/QALY).

Requiring high external validity (Figure 3, *D*) resulted in a tiny fraction of simulation runs being cost-effective (4.5%). Directly observed therapy was no longer consistently effective; 53% of simulation runs now showed DOT to be ineffective. The overall incremental cost-effectiveness now increased to greater than $6 million/QALY.

Imposing all 3 evidence criteria simultaneously yielded similar results: 4.3% of runs were cost-effective, 49% of runs showed effectiveness, and the cost-effectiveness was extremely unfavorable.

Discussion

Our results suggest that quality of evidence may sometimes have a profound impact on the results of decision analytic models. When we did not specify a qualifying level of evidence, DOT was always effective and was usually cost-effective, with 85% of simulations showing an incremental cost-effectiveness more favorable than $100 000 per QALY and an overall incremental cost-effectiveness of $78 000 per QALY. These results were not markedly different from the far more complex HIV simulation published by Goldie and colleagues (25), in which DOT was estimated to have an incremental cost-effectiveness of up to $75 000 per QALY across a wide range of effectiveness assumptions. However, when we specified strict qualifying levels of evidence, our results became much less precise, and the confidence ellipse became commensurately more expansive. Directly observed therapy was often ineffective, and it was seldom cost-effective. Therefore, with constraints concerning quality of evidence, our results became substantially more pessimistic than the results of Goldie and colleagues.

This approach can make the quality of evidence assumptions behind decision analytic modeling more transparent and perhaps more believable. Model users who subscribe to the philosophy that “any data are better than no data” will probably base inferences on model results that incorporate all data sources, regardless of evidence quality. In contrast, users who subscribe to the philosophy that “my judgment supersedes all but the best data” would probably base inferences only on model results that incorporate the highest quality sources. In many respects, this approach constitutes a very simplified and specific application of the confidence profile method (9), in which we aggregated the impact of each separate bias rather than considering them individually and considered this aggregated effect to be either minutely small (if the study meets designated quality criteria) or overwhelmingly large (if the study does not meet designated quality criteria).

Our finding of a large tradeoff between quality of evidence and precision of modeling results is likely to be true of many other published decision analytic models. For example, the cost-effectiveness of alendronate therapy (32) was based on utility estimates from 2 studies with low external validity (utilities were assessed in a different culture and care system) (33, 34) and long-term cost estimates from 1 study with fair internal validity (incompletely controlled for potential confounders) (35). If the authors’ sensitivity analysis were to have incorporated uninformative distributions for these parameters instead of the relatively informative (narrow) distributions that they used, their results may not have been sufficiently precise to suggest a particular decision. Similarly, the cost-effectiveness of preventive strategies for women with *BRCA1* or *BRCA2* mutations (36) was also likely to have been affected by a tradeoff between quality of evidence and precision of results. The authors used studies with poor external validity (tamoxifen was evaluated on patients

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

biologically distinct from the target population) (37) or fair internal validity (surgical strategies were not evaluated in an experimental setting).

Because more than 1 data source may meet evidence criteria for a particular parameter estimate, we needed to develop decision rules to guide which source or sources to use. We elected to use the source with the most statistically precise estimate, a straightforward rule that reflects accepted modeling practice. An approach more consistent with our evidence-based ethos would have been to perform a separate, formal meta-analysis for each possible specification of qualifying evidence criteria, for each individual parameter. However, even for a model as simple as ours, we would have to have performed a prohibitively large number of separate meta-analyses.

Some of the decisions embedded in our approach may be appropriate subjects for debate. While we chose to use uniform distributions to describe parameters that did not meet qualifying evidence criteria, other types of distributions are consistent with our approach and may also be suitable. While we chose wide distributions with the aim of enhancing the conservatism and acceptability of our approach, it may be argued that this decision magnifies the effect of poor quality of evidence. Even though we used a quality evaluation method currently used by an expert panel, evaluating evidence using summary quality scores or domain-specific quality "weights" is often viewed as arbitrary and invalid (38, 39). In the end, the utility of this paper lies in neither its specification of particular evidence valuation strategies nor its probability distributions. Rather, its strength is to highlight the importance of incorporating quality evaluation within decision analytic models and to present one simple and tractable method for accomplishing this important aim.

Acknowledgments

Grant Support: By grant K23 AA14483-01 from the National Institute of Alcohol Abuse and Alcoholism.

References

1. Neumann PJ. Why don't Americans use cost-effectiveness analysis? Am J Manag Care. 2004; 10:308–12. [PubMed: 15152700]
2. Doubilet P, Begg CB, Weinstein MC, Braun P, McNeil BJ. Probabilistic sensitivity analysis using Monte Carlo simulation. A practical approach. Med Decis Making. 1985; 5:157–77. [PubMed: 3831638]
3. Claxton K, Sculpher M, McCabe C, Briggs A, Akehurst R, Buxton M, et al. Probabilistic sensitivity analysis for NICE technology assessment: not an optional extra. Health Econ. 2005; 14:339–47. [PubMed: 15736142]
4. Harris RP, Helfand M, Woolf SH, Lohr KN, Mulrow CD, Teutsch SM, et al. Current methods of the US Preventive Services Task Force: a review of the process. Am J Prev Med. 2001; 20:21–35. [PubMed: 11306229]
5. Guyatt GH, Haynes RB, Jaeschke RZ, Cook DJ, Green L, Naylor CD, et al. Users' Guides to the Medical Literature: XXV. Evidence-based medicine: principles for applying the Users' Guides to patient care. Evidence-Based Medicine Working Group. JAMA. 2000; 284:1290–6. [PubMed: 10979117]
6. Lohr KN. Rating the strength of scientific evidence: relevance for quality improvement programs. Int J Qual Health Care. 2004; 16:9–18. [PubMed: 15020556]
7. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet. 2005; 365:82–93. [PubMed: 15639683]
8. Greenland S. Multiple-bias modeling for analysis of observational data. J R Stat Soc [Ser A]. 2005; 186:267–306.
9. Eddy DM, Hasselblad V, Shachter R. An introduction to a Bayesian method for meta-analysis: the confidence profile method. Med Decis Making. 1990; 10:15–23. [PubMed: 2182960]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

10. Cooper NJ, Sutton AJ, Abrams KR. Decision analytical economic modelling within a Bayesian framework: application to prophylactic antibiotics use for caesarean section. Stat Methods Med Res. 2002; 11:491–512. [PubMed: 12516986]
11. Rothman, KJ.; Greenland, S. Modern Epidemiology. Philadelphia: Lippincott, Williams & Wilkins; 1998.
12. Arias E. United States life tables, 2002. Natl Vital Stat Rep. 2004; 53:1–38. [PubMed: 15580947]
13. Enger C, Graham N, Peng Y, Chmiel JS, Kingsley LA, Detels R, et al. Survival from early, intermediate, and late stages of HIV infection. JAMA. 1996; 275:1329–34. [PubMed: 8614118]
14. Egger M, May M, Chene G, Phillips AN, Ledergerber B, Dabis F, et al. Prognosis of HIV-1-infected patients starting highly active antiretroviral therapy: a collaborative analysis of prospective studies. Lancet. 2002; 360:119–29. [PubMed: 12126821]
15. Haubrich RH, Little SJ, Currier JS, Forthal DN, Kemper CA, Beall GN, et al. The value of patient-reported adherence to antiretroviral therapy in predicting virologic and immunologic response. California Collaborative Treatment Group. AIDS. 1999; 13:1099–107. [PubMed: 10397541]
16. Paterson DL, Swindells S, Mohr J, Brester M, Vergis EN, Squier C, et al. Adherence to protease inhibitor therapy and outcomes in patients with HIV infection. Ann Intern Med. 2000; 133:21–30. [PubMed: 10877736]
17. Howard AA, Arnsten JH, Lo Y, Vlahov D, Rich JD, Schuman P, et al. A prospective study of adherence and viral load in a large multi-center cohort of HIV-infected women. AIDS. 2002; 16:2175–82. [PubMed: 12409739]
18. Arnsten JH, Demas PA, Grant RW, Gourevitch MN, Farzadegan H, Howard AA, et al. Impact of active drug use on antiretroviral therapy adherence and viral suppression in HIV-infected drug users. J Gen Intern Med. 2002; 17:377–81. [PubMed: 12047736]
19. Altice FL, Mezger JA, Hodges J, Bruce RD, Marinovich A, Walton M, et al. Developing a directly administered antiretroviral therapy intervention for HIV-infected drug users: implications for program replication. Clin Infect Dis. 2004; 38 (Suppl 5):S376–87. [PubMed: 15156426]
20. Freedberg KA, Scharfstein JA, Seage GR 3rd, Losina E, Weinstein MC, Craven DE, et al. The cost-effectiveness of preventing AIDS-related opportunistic infections. JAMA. 1998; 279:130–6. [PubMed: 9440663]
21. Burman WJ, Dalton CB, Cohn DL, Butler JR, Reves RR. A cost-effectiveness analysis of directly observed therapy vs self-administered therapy for treatment of tuberculosis. Chest. 1997; 112:63–70. [PubMed: 9228359]
22. Moore RD, Chaulk CP, Griffiths R, Cavalcante S, Chaisson RE. Cost-effectiveness of directly observed versus self-administered therapy for tuberculosis. Am J Respir Crit Care Med. 1996; 154:1013–9. [PubMed: 8887600]
23. Palmer CS, Miller B, Halpern MT, Geiter LJ. A model of the cost-effectiveness of directly observed therapy for treatment of tuberculosis. J Public Health Manag Pract. 1998; 4:1–13. [PubMed: 10186738]
24. Bozzette SA, Joyce G, McCaffrey DF, Leibowitz AA, Morton SC, Berry SH, et al. Expenditures for the care of HIV-infected patients in the era of highly active antiretroviral therapy. N Engl J Med. 2001; 344:817–23. [PubMed: 11248159]
25. Goldie SJ, Paltiel AD, Weinstein MC, Losina E, Seage GR 3rd, Kimmel AD, et al. Projecting the cost-effectiveness of adherence interventions in persons with human immunodeficiency virus infection. Am J Med. 2003; 115:632–41. [PubMed: 14656616]
26. The Cochrane Collaboration. Review Manager 4.2 for Windows. Oxford, England: The Cochrane Collaboration; 2002.
27. Braithwaite RS, Justice AC, Chang CC, Fusco JS, Raffanti SR, Wong JB, et al. Estimating the proportion of patients infected with HIV who will die of comorbid diseases. Am J Med. 2005; 118:890–8. [PubMed: 16084183]
28. Braithwaite RS, Chang CC, Shechter S, Schaefer A, Roberts MS. Estimating the rate of accumulating drug mutations in the HIV genome. Value in Health. [In press].
29. Braithwaite RS, Shechter S, Roberts MS, Schaefer A, Bangsberg DR, Harrigan PR, et al. Explaining variability in the relationship between antiretroviral adherence and HIV mutation accumulation. J Antimicrob Chemother. 2006; 58:1036–43. [PubMed: 17023498]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

30. Gold, MR.; Siegel, JE.; Russell, LB.; Weinstein, MC. Cost-Effectiveness in Health and Medicine. New York: Oxford Univ Pr; 1996.

31. Fenwick E, Claxton K, Sculpher M. Representing uncertainty: the role of cost-effectiveness acceptability curves. Health Econ. 2001; 10:779–87. [PubMed: 11747057]

32. Schousboe JT, Nyman JA, Kane RL, Ensrud KE. Cost-effectiveness of alendronate therapy for osteopenic postmenopausal women. Ann Intern Med. 2005; 142:734–41. [PubMed: 15867405]

33. Burström K, Johannesson M, Diderichsen F. Swedish population health-related quality of life results using the EQ-5D. Qual Life Res. 2001; 10:621–35. [PubMed: 11822795]

34. Kanis JA, Johnell O, Oden A, Borgstrom F, Zethraeus N, De Laet C, et al. The risk and burden of vertebral fractures in Sweden. Osteoporos Int. 2004; 15:20–6. [PubMed: 14593450]

35. Leibson CL, Tosteson AN, Gabriel SE, Ransom JE, Melton LJ. Mortality, disability, and nursing home use for persons with and without hip fracture: a population-based study. J Am Geriatr Soc. 2002; 50:1644–50. [PubMed: 12366617]

36. Anderson K, Jacobson JS, Heitjan DF, Zivin JG, Hershman D, Neugut AI, et al. Cost-effectiveness of preventive strategies for women with a BRCA1 or a BRCA2 mutation. Ann Intern Med. 2006; 144:397–406. [PubMed: 16549852]

37. Fisher B, Costantino JP, Wickerham DL, Redmond CK, Kavanah M, Cronin WM, et al. Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. J Natl Cancer Inst. 1998; 90:1371–88. [PubMed: 9747868]

38. Spiegelhalter DJ, Best NG. Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. Stat Med. 2003; 22:3687–709. [PubMed: 14652869]

39. Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA. 1999; 282:1054–60. [PubMed: 10493204]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

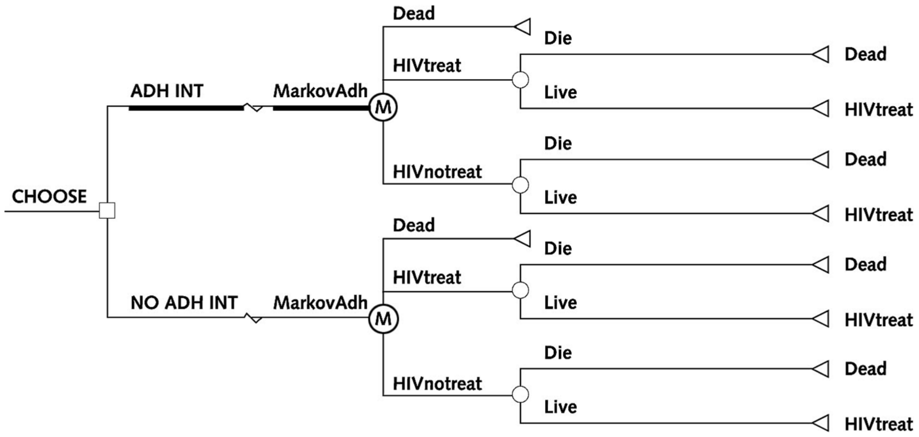

Figure 1. Schematic diagram of decision analytic model

This model was constructed with extreme parsimony solely to illustrate the approach. Squares signify choices (for example, use the adherence intervention [*ADH INT*] versus do not use the adherence intervention [*NO ADH INT*]), circles signify potential consequences (for example, live versus die), and triangles signify consequences that may endure over a significant time frame and therefore may affect quality or quantity of life (for example, treat HIV). The circle inscribed with “M” is a Markov node (*MarkovAdh*), denoting that the model has the capacity to represent changes in clinical status over time. HIVtreat = receives and adheres to HIV treatment; HIVnotreat = does not receive or does not adhere to HIV treatment.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

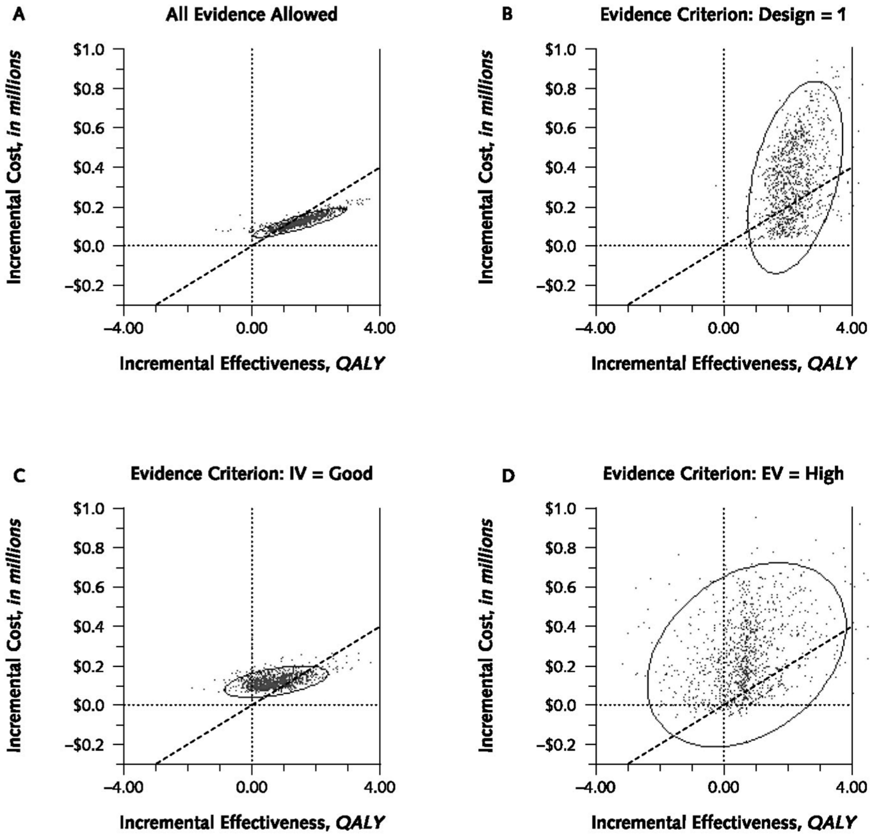

Figure 2. Cost-effectiveness plane and 95% confidence ellipses for directly observed therapy in the absence of any evidence criteria (*A*) and with quality of evidence criteria applied to research design (*B*), internal validity (*C*), and external validity (*D*)

Each point signifies the incremental cost-effectiveness of a particular model run. Points that cluster within a narrow 95% confidence ellipse (the analogue of a 95% CI) suggest great precision, whereas points that scatter throughout a wide 95% confidence ellipse suggest low precision. The location of each point on the cost-effectiveness plane indicates its cost-effectiveness. As a general guide, when points lie to the left of a decision making threshold ($100 000 per quality-adjusted life-year [*QALY*] is a commonly used threshold), incremental costs are high relative to incremental benefits, and cost-effectiveness is unfavorable. In contrast, when points lie to the right of that threshold, incremental costs are low relative to incremental benefits, and cost-effectiveness is favorable. Our results became notably less precise when stricter evidence criteria were applied, most dramatically with an external validity (*EV*) criterion. IV = internal validity.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

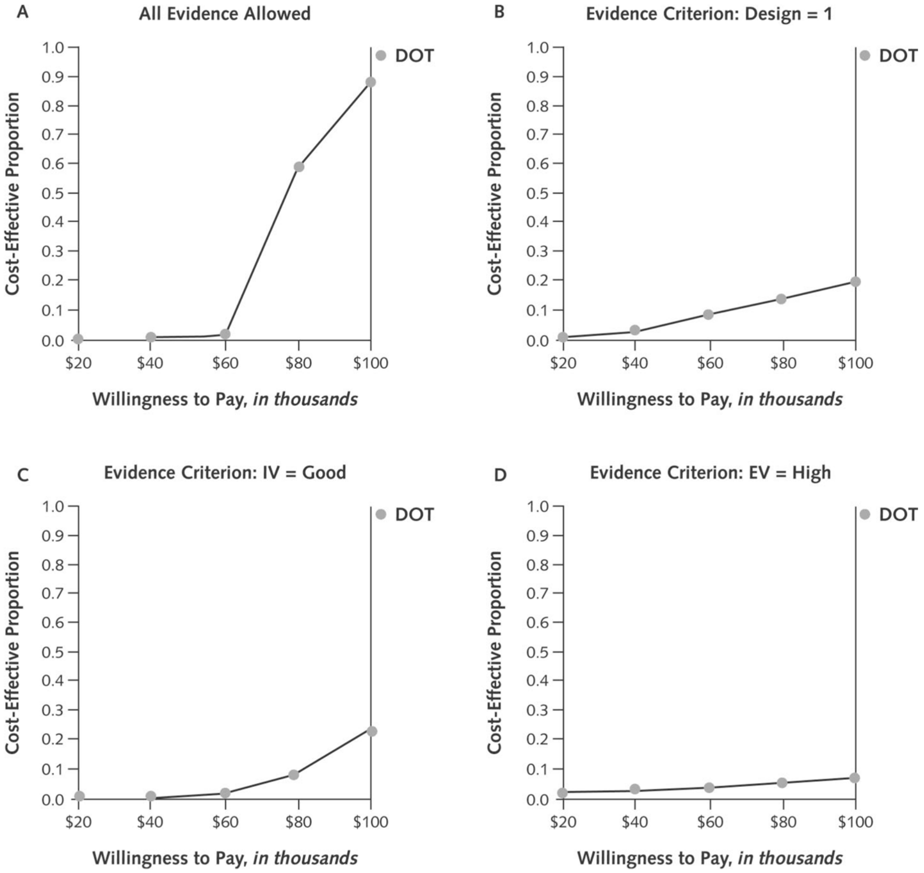

Figure 3. Acceptability curves for directly observed therapy (*DOT*) in the absence of any evidence criteria (*A*) and with strength of evidence criteria applied to research design (*B*), internal validity (*C*), and external validity (*D*)

The horizontal axis shows a range of values that society may be willing to pay for health benefits, and the curve's elevation (on the vertical axis) denotes the probability that DOT has an incremental cost-effectiveness that is more favorable than the corresponding willingness to pay. DOT became notably less cost-effective when stricter evidence criteria were applied, most dramatically with an external validity (*EV*) criterion. IV = internal validity.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 1

Strength of Evidence Hierarchy[*]

Study design

- Level 1: Evidence obtained from at least 1 properly designed randomized, controlled trial[†]
- Level 2-1: Evidence obtained from well-designed controlled trials without randomization
- Level 2-2: Evidence obtained from well-designed cohort or case–control analytic studies, preferably from more than 1 center or research group
- Level 2-3: Evidence obtained from multiple time series with or without the intervention
- Level 3: Opinions of respected authorities, based on clinical experience, descriptive studies and case reports, or reports of expert committees

Internal validity

- Good: Meets all criteria for study design
- Fair: Does not meet all criteria for study design but is judged to have no fatal flaw that invalidates its results
- Poor: Study contains a fatal flaw
- Criteria
  - Systematic reviews
    - Comprehensiveness of sources/search strategy used
    - Standard appraisal of included studies
    - Validity of conclusions
    - Recency and relevance
  - Case–control studies
    - Accurate assessment of cases
    - Nonbiased selection of cases/controls with exclusion criteria applied equally to both
    - Response rate
    - Diagnostic testing procedures applied equally to each group
    - Appropriate attention to potential confounding variables
  - Randomized, controlled trials
    - Initial assembly of comparable groups (concealment and distribution of potential confounders)
    - Maintenance of comparable groups (includes attrition, crossovers, adherence, contamination)
    - Important differential loss to follow-up or overall high loss to follow-up
    - Measurements: equal, reliable, and valid (includes masking of outcome assessment)
    - Clear definition of interventions
    - All important outcomes considered
    - Intention-to-treat analysis
    - Sample size/power
  - Cohort studies
    - Consideration of potential confounders; consideration of inception cohorts
    - Adjustment for potential confounders
    - Important differential loss to follow-up or overall high loss to follow-up
    - Sample size/power

External validity

- High: Meets criteria
- Low: Does not meet criteria
- Criteria

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Biological plausibility
Similarities of the study sample to the target population (risk factor profile, demographics, ethnicity, sex, clinical presentation, and similar factors)
Similarities of the test or intervention studied to those that would be routinely available or feasible in clinical practice
Clinical or social environmental circumstances in the studies that could modify the results from those expected in a primary care setting

[*] Modified from the U.S. Preventive Services Task Force (4).

[†] Observational studies qualify as level 1 if they are used to estimate a parameter that cannot be determined experimentally (for example, mortality rate due to age-, sex-, and race-related causes).

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 2

Parameters in Computer Simulation[*]

Parameter	Data Source (Reference)	Strength of Evidence			Does Data Source Meet Strict Evidence Criteria?				Parameter Distribution if Data Source Used	Parameter Distribution if Data Source Not Used
		Study Design	Internal Validity	External Validity	Study Design Level 1	Internal Validity Good	External Validity High	All 3 Criteria		
Mortality rate in absence of HIV	Observational; life tables (12)	1	Good	Low	Yes	Yes	No	No	Point estimates, variable	Uniform (0.5X, 1.5X estimates)
Mortality rate attributable to HIV	Observational; 1 study in similar population (13)	1	Good	Low	Yes	Yes	No	No	Normal (0.19, 0.06)	Uniform (0, 0.38)
Impact of HIV treatment on mortality	Observational; 13 studies pooled from similar populations (14)	1	Good	High	Yes	Yes	Yes	Yes	Normal (0.15, 0.02)	NA
Probability of taking HIV medications	Observational (15)	1	Fair	Low	Yes	No	No	No	Never used[†]	NA
	Observational (16)	1	Good	High	Yes	Yes	Yes	Yes	Normal (0.75, 0.02)	NA
	Observational (17)	1	Good	Low	Yes	Yes	No	No	Never used[†]	NA
	Observational (18)	1	Good	Low	Yes	Yes	No	No	Never used[†]	NA
Effectiveness of DOT	Randomized, controlled trial; 1 study in dissimilar population (19)	1	Good	Low	Yes	Yes	No	No	Normal (0.46, 0.01)	Uniform (0, 2)
Utility with HIV	Observational; 1 study in similar population (20)	1	Poor	Low	Yes	No	No	No	Normal (0.87, 0.04)	Uniform (0.5, 1.0)
Decrement in utility with HIV treatment	Observational; 1 study in similar population (unpublished)	1	Poor	Low	Yes	No	No	No	Normal (0.05, 0.01)	Uniform (0, 0.5)
Annual cost of DOT	Expert opinion	3	Poor	Low	No	No	No	No	Never used[†]	NA
	Observational; 1 study in dissimilar population (21)	2-2	Good	Low	No	Yes	No	No	Point estimate $4600	Uniform ($200, $36 500)
	Observational; 1 study in dissimilar population (22)	2-2	Poor	Low	No	No	No	No	Never used[†]	NA
	Observational; 1 study in dissimilar population (23)	2-2	Poor	Low	No	No	No	No	Never used[†]	NA
Annual cost of nondrug HIV care	Observational; 1 study in similar population (24)	1	Poor	High	Yes	No	Yes	No	Normal ($9000, $300)	Uniform ($200, $20 000)
	Observational; 1 study in similar population (25)	1	Poor	High	Yes	No	Yes	No	Never used[†]	NA
Annual cost of HIV drugs	Observational; 1 study in similar population (25)	1	Good	High	Yes	Yes	Yes	Yes	Normal ($10 300, $700)	NA

[*] Individual data sources were eligible to inform parameter estimations if their strength of evidence met or exceeded criteria in 3 separate domains (study design, internal validity, and external validity). If no data sources met strength of evidence criteria, a uniform distribution across a wide plausible range was substituted. When more than 1 study met strength of evidence criteria, the parameter's distribution was based on the study with the most precise statistical estimate. Normal = normal distribution (mean, standard deviation); uniform = uniform distribution (lower bound, higher bound). DOT = directly observed therapy; NA = not applicable because data source either was never used or met all 3 of the strict evidence criteria.

[†] "Never used" because results were statistically less precise than those of another study or studies with equal or superior grades of evidence in all 3 domains. It is also possible that a particular data source could be statistically more precise but have lower strength of evidence than alternative data sources, although this situation did not arise with our example. In accord with our decision rules, we would have used the statistically more precise data source when evidence criteria were more inclusive and the statistically less precise data source when evidence criteria were less inclusive.