International Journal of Technology Assessment in Health Care, **27:2** (2011), 133–138.

doi:10.1017/S0266462311000110

Bridging the gap between aggregate data and individual patient management: A Bayesian approach

Gert Jan van der Wilt, Hans Groenewoud, Piet van Riel
Radboud University Medical Centre

Objectives: The aim of this study was to explore whether Bayesian reasoning can be applied to therapeutic questions in a way that is similar to its application in diagnostics.
Methods: A clinically relevant, therapeutic question was formulated in accordance with Bayesian reasoning for the clinical management of patients with newly diagnosed rheumatoid arthritis (RA). Prior probability estimates of response to drug treatment (methotrexate, MTX) were obtained from the literature. As a marker of treatment response, changes in the Health Assessment Questionnaire (HAQ) scores were assessed after three months of treatment. Likelihood ratios for this marker were calculated on the basis of data from a clinical registry, using changes in the Disease Activity Score (DAS) as gold standard. Using Bayes' theorem, prior probability and likelihood ratios were combined to estimate posterior probabilities of treatment response in individual patients.
Results: On the basis of the literature, the prior probability of response of RA patients to MTX was estimated 45 percent. At 3 months follow-up, this probability increased to 80 percent or decreased to 23 percent, depending on the changes that were observed in Health Assessment Questionnaire scores.
Conclusions: Bayesian reasoning can be applied to therapeutic issues in a way that is conceptually fully compatible with its use in diagnostics. As such, it can be used to bridge the gap between aggregate data and individual patient management.

Keywords: Evidence-based medicine, Bayesian reasoning, Individual patient management, Rheumatoid arthritis, Drug treatment

A gap exists between evidence that is based on aggregate data and individual patient management. Aggregate data allow for an estimate of the probability that, on average, treatments produce a clinically important benefit in specified populations of patients. Reputedly, physicians claim that the average patient does not exist. This difference in perspective may be one of the barriers to the uptake of evidence-based medicine. With the launch of the Comparative Effectiveness Research program in the United States, the debate has gained new momentum, with opponents suggesting that the initiative poses a threat to individualized medicine (1;7;14).

To illustrate the issue, take the following example: In a double-blind, randomized study, the effect of methotrexate (MTX) was compared with placebo on disease progression in

The research that is reported in this study derives from a project, funded by the Council for Health and Health Services Research from the Netherlands, ZonMw (Title: Potential and limitations of Bayesian analyses in synthesis of evidence from multiple sources. Project Number 80-82500-98-8201. Coordinator: Professor dr. G. J. van der Wilt). The funding organization had no role in the conceptualization of the study, the collection and analysis of the data, or in the reporting of the results.

patients with probable rheumatoid arthritis (RA) (16). In the placebo group, 53 percent of patients progressed to RA; this proportion was reduced to 40 percent in the MTX group ($p < 0.05$). Now, for the practicing physician, the question is: will the treatment work *in this individual patient*? The answer is: the odds that it will work is 1.5 (0.6/(1-0.6). Whether this warrants treatment depends, of course, on several other issues, such as side-effects, costs of treatment, severity of progressing disease, and the performance of alternative treatment options, if any. Equally important, however, is the question: Once started, should the treatment be continued? In other words, does the patient belong to the group of responders or to the—slightly smaller—group of nonresponders? The relevance of the question derives from the discontinuation of a treatment that is not helpful but potentially harmful or costly, and from the benefit of early switching to a possibly more effective treatment option. For such decisions, markers of treatment benefit are crucially important. A marker provides individual patient information that, allegedly, may be interpreted as an indication of whether the patient is, or will be, responding to treatment, and that can be obtained more easily or earlier than information on actual treatment response. Such markers are ubiquitously used in diabetes, cancer, and cardiovascular and infectious disease (13). The question is: How can this patient-specific information be combined with the information at group level to provide an accurate estimate that the patient is, actually, responding to treatment?

A Bayesian approach may be conducive to this end. Bayesian analysis differs from conventional, frequentist analysis in its concept of probability. For frequentists, probability refers to the probability of an observed outcome (for instance observed difference in response rates between groups in a randomized clinical trial), under the assumption that a particular hypothesis (usually the null hypothesis of no difference) is true. Here, the data are considered stochastic variables and the hypothesis is fixed. In the Bayesian approach, probability is a measure of the strength of justified belief. Here, the data are considered as given, and the hypotheses may vary. The *p* value expresses the probability that a particular claim (for instance, "drug A leads to a greater reduction in blood pressure than drug B") is, in fact, true (9). Although the interest in Bayesian reasoning has substantially increased recently, it is not uncontroversial (8). This is not, however, the case for the application of Bayesian reasoning to diagnostic issues.

Bayesian Reasoning in Diagnostics

A key issue in diagnosis is to find out whether an individual patient has a specific disease condition. For instance, in patients presenting to a hospital's emergency department with acute pain in the right lower quadrant of the abdomen, the question is: Does the patient have acute appendicitis (AA)? Bayesian reasoning helps to estimate the probability that a specific patient does, indeed, have this condition. For this, a prior probability estimate is needed: What is the probability that individuals like this do, in fact, have AA? An estimate of this probability may be based on the prevalence of AA among such individuals, for instance from a hospital's records. For AA, a reasonable estimate would be 0.3. Such information is insufficient to guide clinical management. Hence, further data from each individual patient need to be collected to adjust this prior probability, for instance through compression ultrasonography. By how much the prior probability needs to be revised on the basis of the test result depends on the likelihood ratio of the test (12). Prior probability, test result, and likelihood ratio are combined to estimate posterior probability of AA, using Bayes' theorem (10). Likelihood ratios can be estimated on the basis of the results of diagnostic test studies. A meta-analysis of studies indicates that compression ultrasonography has a positive likelihood ratio of 4.5 and a negative likelihood ratio of 0.27 (18). Hence, the posterior probability of AA would be 0.66 and 0.12, in case of a positive and a negative test result, respectively. When these probabilities are considered still insufficient to guide clinical management, further testing is warranted, or a different diagnostic test should be used in the first place.

Application of Bayesian Analysis to Individual Patient Treatment Decisions

To be able to apply Bayesian reasoning in a similar way to individual treatment decisions, the following elements are needed: a relevant clinical question; an empirical basis for a prior probability estimate; a marker that allows for revision of this prior probability estimate in an individual patient (the "test result"); empirical evidence to estimate the likelihood ratio of this marker; a gold standard that allows for correct classification of patients; and cutoff values to guide clinical management.

The objective of our study is to develop concrete suggestions for each of these elements, using treatment of patients with newly diagnosed rheumatoid arthritis (RA) with methotrexate (MTX) as an example.

METHODS

Estimate of Prior Probability of Treatment Response

To obtain an empirical estimate of the prior probability of treatment response, a literature search was conducted in PubMed, identifying recent studies reporting responses among patients newly diagnosed with RA to MTX monotherapy at 3 months follow-up. Clinical response had to be expressed in terms of the Disease Activity Score, a valid measure of disease activity in patients with RA (6). For this purpose, "Arthritis, rheumatoid" [MeSH], methotrexate (In Ti), "Clinical Trial" [PT], "Treatment outcome" [MeSH], and "Severity of Illness Index" [MeSH] were combined with the

Table 1. Corollaries between a Bayesian Approach to Diagnostic and Therapeutic Issues

	Diagnostic context	Therapeutic context
Relevant clinical question	What is the probability that this patient has disease condition X (e.g., acute appendicitis)?	What is the probability that this patient benefits from treatment X? (e.g., the probability that a newly diagnosed patient with RA will benefit from treatment with MTX)
Empirical basis for prior probability	Prevalence of condition X among an appropriate spectrum of patients (result of diagnostic test research)	Proportion of responders to treatment X among an appropriate spectrum of patients (e.g., results of a cohort study, RCT, or meta-analysis)
Marker that serves to revise an individual's probability	Diagnostic test result (e.g., ultrasonogram suggestive of an infected appendix)	An indicator of treatment response that can be easily (preferably noninvasively) obtained, e.g., patient-reported improvement in symptom relief or improved functioning
Gold standard that serves to correctly classify patients	Gold standard to classify patients as having- and not-having AA (e.g., pathology or follow up)	A broadly accepted and validated criterion for treatment response
Cutoff values	At what probability may we accept the diagnosis, and at what probability should we reject it? (e.g., when is it justified to wait and see, when is it justified to prepare for surgery?	At what probability is it justified to continue treatment, and at what probability is it justified to discontinue or change treatment?

MTX, methotrexate; RA, rheumatoid arthritis; AA, acute appendicitis.

Boolean operator AND, with limitations English (language), and published after 2000 in Core Clinical Journals.

Identification of a Marker of Treatment Benefit and Gold Standard

As a marker of treatment benefit that can be easily and noninvasively obtained, we have used the Health Assessment Questionnaire (HAQ) score. The HAQ score is a self-reported measure of functional status, comprising disability, discomfort, and pain (4). As a means to correctly classify patients as responders and nonresponders, we used the Disease Activity Score (DAS). Calculation of the DAS requires information on the number of swollen and tender joints, the Erythrocyte Sedimentation Rate (ESR) and the patient's general health as measured by a Visual Analog Scale (VAS). The DAS has been demonstrated to be a valid measure of disease activity in patients with RA (6). This is the counterpart of the gold standard that is used to classify patients correctly in a diagnostic study.

Calculation of Likelihood Ratios

To calculate the likelihood ratios that are associated with changes in the HAQ score, we used data from our clinical registry, which contains data from over 1,000 patients with RA on patients' characteristics, treatment, side effects and course of disease, both objectively (DAS) and subjectively (HAQ) (19). From this registry, we selected first-time users of MTX with a minimal follow-up of 3 months. Patients were classified as responders (good and moderate) and nonresponders on the basis of the European League Against Rheumatism (EULAR) criteria (17). Likelihood ratios were calculated on the basis of the observed frequency of specified changes in the HAQ score among responders and nonresponders.

RESULTS

The elements that are needed for a Bayesian approach to individual patient treatment decisions are summarized in Table 1. The first element is a relevant clinical question. In our example, this question would be: "What is the probability that an individual patient is actually responding to treatment with MTX?" Note that this question is truly Bayesian, in the sense that it reflects the confidence that we may have in whether the patient is actually benefiting from treatment. It is the counterpart of the question in the diagnostic workup, asking about the probability that a patient has a disease condition of interest. This question can be raised during any follow-up visit since MTX treatment was initiated. The results that are reported below apply to follow-up of patients at 3 months.

The second element that is needed is an estimate of the prior probability of treatment response. This estimate was obtained from the literature. Our search produced nineteen hits. One of these studies reported results that could be used to estimate the prior probability of response in patients with newly diagnosed RA on MTX monotherapy at 3 months follow-up (2). In this cohort study, approximately 45 percent of patients were reported to respond to MTX, response being defined in accordance with the EULAR criteria for good and moderate response. Thus, 45 percent was used as a reasonable and evidence-based estimate of the prior probability of treatment response.

The third and fourth element that are needed, are a marker of treatment response and a gold standard to correctly classify patients as responders and nonresponders. In our example, these are the HAQ and the DAS, respectively. Together, they can be used to calculate Likelihood Ratios. In

https://doi.org/10.1017/S0266462311000110 Published online by Cambridge University Press

Table 2. Posterior Probability Estimates

Improvement in HAQ score, 3 months after start of MTX treatment (cutoff values)	Probability of the specified improvement in HAQ score among responders	Probability of the specified improvement in HAQ score among nonresponders	Positive likelihood ratio	Posterior probability that a patient with this change in HAQ score is responding to treatment
>0.05	.52	.28	1.8	.60
>0.10	.38	.20	1.9	.61
>0.15	.31	.13	2.5	.67
>0.20	.25	.09	2.8	.70
>0.25	.18	.04	5.0	.80
Deterioration in HAQ score, 3 months after start of MTX treatment (cutoff values)	**Probability of the specified deterioration in HAQ score among responders**	**Probability of the specified deterioration in HAQ score among nonresponders**	**Negative likelihood ratio**	**Posterior probability that a patient with this change in HAQ score is responding to treatment**
>0.05	.12	.31	0.38	.24
>0.10	.10	.22	0.48	.28
>0.15	.07	.14	0.54	.31
>0.20	.03	.08	0.36	.23
>0.25	.03	.06	0.53	.30

Note. Probabilities of the specified changes (improvement or deterioration) in HAQ score among responders and nonresponders to MTX treatment (according to EULAR criteria), associated likelihood ratios, and resulting posterior probabilities that patients with the specified change in HAQ score are responding to treatment (prior probability estimate of response to MTX: 0.45).
HAQ, Health Assessment Questionnaire; MTX, methotrexate.

our clinical registry, 1,652 patients with newly diagnosed RA could be identified who were first-time users of MTX with a minimal follow-up of 3 months. All patients could be classified as responders (good and moderate) and nonresponders on the basis of the EULAR criteria. Likelihood ratios were calculated for various improvements and deteriorations in HAQ score and are presented in Table 2. Negative likelihood ratio varied from 0.36 to 0.54, whereas positive likelihood ratio increased from 1.83 for moderate improvement to 5.03 for substantial improvement in HAQ, respectively.

Posterior Probability Estimates

Combining the prior probability estimate for response (45 percent) with calculated likelihood ratios resulted in posterior probability estimates as shown in Table 2. The Table shows that when, after 3 months of treatment with MTX, a patient reports substantially improved functional ability (c.q., improvement in HAQ > 0.25), this patient is almost twice as likely to be a responder to the treatment (in the objective sense) as compared to his probability when treatment was initiated (0.8 versus 0.45). When, on the other hand, a patient reports deterioration in functional ability, the probability that this patient is responding to MTX treatment may be only half of the initial value (0.24 versus 0.45). We conducted the analysis using a different drug (etanercept) and different periods of follow-up, and found similar likelihood ratios (data not shown). This suggests that, as might be expected, these likelihood ratios are a property of the marker (in this case, the change in HAQ score), rather than the drug or follow-up period.

DISCUSSION

Bayesian reasoning can be applied to therapeutic issues in a way that is conceptually fully compatible with its application to diagnostic issues. As such, it deserves broader acceptance and use. A Bayesian approach offers several advantages, including ease of interpretation (p values directly reflect confidence that we may have that a specific statement is true), relevance to clinical practice, and its flexibility (e.g., a different prior probability may be used if a clinician has reason to believe that a specific patient may respond better or poorer than average on a particular treatment). Also, a Bayesian approach offers a means of bridging the gap between aggregate evidence and individual patient management. In Bayesian reasoning, we start from an estimate of the probability that a patient like this will, generally speaking, respond to a specific treatment. This estimate should be based on aggregate data (e.g., outcomes of controlled or observational studies). This probability estimate is then revised as soon as further, specific evidence becomes available from the individual patient, using Bayes' theorem. In the diagnostic context, this patient-specific evidence is the result of a diagnostic test; in the therapeutic context, this patient-specific evidence consists of a marker of treatment response.

The Bayesian approach has been criticized for its subjective concept of probability (8). However, the p value is a measure of justified belief, and justification should be based on firm, and relevant empirical evidence. Such basis can be found, not only for diagnostic issues, but also for therapeutic issues. As such, a Bayesian estimate of the probability of treatment benefit may be used in the communication with

https://doi.org/10.1017/S0266462311000110 Published online by Cambridge University Press

the patient: what is to be expected from starting (or continuing) a particular treatment? Also, it may be used to support self-management, allowing patients to learn how improvements or deteriorations in functional ability as experienced by themselves translate into objective responses. Thus, it can help to support crucially important decisions in patients with a chronic condition regarding (dis)continuation of treatment.

To make this happen, physicians should be made more aware of the practical significance of Bayesian reasoning, not only to diagnostic issues, but to therapeutic issues as well. Calculations are greatly facilitated by Bayesian calculators that are freely available through the Web (3). The required input consists of prior odds (based on prevalence of response as observed in trials) and appropriate likelihood ratios. For the latter, more work needs to be done on identification of useful markers of disease progression for various conditions and on reporting their likelihood ratios. In terms of research design, this requires that for individual patients, data are collected on potential markers and on actual treatment response (the "gold standard") concurrently, as is done in diagnostic test research. Equally important, however, is that physicians start to reflect on cut off values: at what probability values may we safely assume that a patient actually benefits from treatment and that no change in treatment is called for, and at what level should a change in treatment be seriously considered?

In our example, we used data from an observational study, comparing response rates among patients receiving MTX with patients receiving anti-TNF alpha blockade. Two comments are in order. First, the study was a nonrandomized, open-label study and, as such, susceptible to bias and confounding. Indeed, in a randomized controlled trial comparing MTX monotherapy with combination therapy (MTX + anti-TNF alpha blockade), the response rate among patients receiving MTX monotherapy at 52 weeks follow-up was 28 percent (95 percent CI: 23–33) (5). Ideally, a prior probability estimate should be based on a meta-analysis of all available and relevant evidence, taking into account methodological quality. However, a strength of the Bayesian approach is that it easily allows for substituting initial estimates for different values, selecting those that seem to be most relevant for the particular context. Second, we used a point estimate, disregarding any uncertainty resulting from sampling variation. This too, can be easily accommodated by using a prior probability distribution, for instance in the form of a beta function. Likelihood ratios, too, can be expressed in terms of likelihood functions, giving rise to posterior probability distributions, rather than point estimates as in our example. The rationale for using point estimates was ease of interpretation. Also, the issue of uncertainty is not critical to our main argument, that Bayesian analyses can be used in individual patient treatment decisions in a way which is fully compatible with its use in establishing diagnoses in individual patients.

Translating evidence into implications for individual patient management has been of focal interest in evidence based medicine (11;15). What this study adds is the contention that a Bayesian analysis offers a formal algorithm for combining individual patient data with general evidence to arrive at individualized probability estimates of treatment benefit.

CONFLICT OF INTEREST

The authors do not report any potential conflicts of interest.

CONTACT INFORMATION

Gert Jan van der Wilt, PhD (G.vanderwilt@ebh.umcn.nl), Professor, **Hans Groenewoud**, MSc (H.Groenewoud@ebh.umcn.nl), Data Manager; Department of Epidemiology, Biostatistics, and Health Technology Assessment; **Piet van Riel**, MD, PhD (P.vanRiel@reuma.umcn.nl), Professor and Head, Department of Rheumatology, Radboud University Medical Center, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands

REFERENCES

1. Avorn J. Debate about funding comparative-effectiveness research. *N Engl J Med.* 2009;360:1927-1929.
2. Barrera P, Van Der Maas A, van Ede AE, Kiemeney BA, Laan RF, van de Putte LB, van Riel PL. Drug survival, efficacy and toxicity of monotherapy with a fully human anti-tumour necrosis factor-alpha antibody compared with methotrexate in long-standing rheumatoid arthritis. *Rheumatology (Oxford).* 2002;41:430-439.
3. Birnbaum MH. *Bayesian calculator*. 1999. http://psych.fullerton.edu/mbirnbaum/bayes/BayesCalc.htm (accessed December 10, 2010).
4. Bruce B, Fries JF. The Stanford Health Assessment Questionnaire: A review of its history, issues, progress, and documentation. *J Rheumatol.* 2003;30:167-178.
5. Emery P, Breedveld FC, Hall S, et al. Comparison of methotrexate monotherapy with a combination of methotrexate and etanercept in active, early, moderate to severe rheumatoid arthritis (COMET): A randomised, double-blind, parallel treatment trial. *Lancet.* 2008;372:375-382.
6. Fransen J, van Riel PC. The disease activity score and EULAR response criteria. *Clin Exp Rheumatol.* 2005;23 (Suppl 39):S93-S99.
7. Garber AM, Tunis SR. Does comparative-effectiveness research threaten personalized medicine? *N Engl J Med.* 2009;360:1925-1927.
8. Gelman A. Objections to Bayesian statistics. *Bayesian Analysis.* 2008;3:445-450.
9. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. *Ann Intern Med.* 1999;130:995-1004.
10. Goodman SN. Introduction to Bayesian methods I: Measuring the strength of evidence. *Clin Trials.* 2005;2:282-290.
11. Greenhalgh T, Worrall JG. From EBM to CSM: The evolution of context-sensitive medicine. *J Eval Clin Pract.* 1997;3:105-108.
12. Grimes DA, Schulz KF. Refining clinical diagnosis with likelihood ratios. *Lancet.* 2005;365:1500-1505.

13. Kuper H, Nicholson A, Kivimaki M, et al. Evaluating the causal relevance of diverse risk markers: Horizontal systematic review. *BMJ.* 2009;339:b4265.
14. Naik AD, Petersen LA. The neglected purpose of comparative-effectiveness research. *N Engl J Med.* 2009;360:1929-1931.
15. Sackett DL, Rosenberg WM, Gray JA, et al Evidence based medicine: What it is and what it isn't. *BMJ.* 1996;312: 71-72.
16. van Dongen H, van Aken J, Lard LR, et al. Efficacy of methotrexate treatment in patients with probable rheumatoid arthritis: A double-blind, randomized, placebo-controlled trial. *Arthritis Rheum.* 2007;56:1424-1432.
17. van Gestel A, Prevoo ML, van 't Hoff MA, et al. Development and validation of the European League Against Rheumatism response criteria for rheumatoid arthritis. Comparison with the preliminary American College of Rheumatology and the World Health Organization/International League Against Rheumatism Criteria. *Arthritis Rheum*. 1996;39:34-40.
18. van Randen A, Bipat S, Zwinderman AH. Acute appendicitis: Meta-analysis of diagnostic performance of CT and graded compression US related to prevalence of disease. *Radiology.* 2008;249:97-106.
19. Welsing PM, van Riel PL. The Nijmegen inception cohort of early rheumatoid arthritis. *J Rheumatol Suppl.* 2004;69:14-21.