Evidence Based Medicine
Assessing the Literature

I. Diagnostic Tests:

A. Is this evidence about a diagnostic test valid?

B. Is this evidence about a diagnostic test important?

C. Can you apply this valid, important evidence about a diagnostic test in caring for your patient?

A. Is the evidence about a diagnostic test valid?

1. Was there an independent, blind comparison with a reference “gold” standard of diagnosis?

2. Was the diagnostic test evaluated in an appropriate spectrum of patients (like those in whom it would be used in practice)?

3. Was the reference standard applied regardless of the diagnostic test result?

B. Is the evidence about this diagnostic test important:

1. How does the result of the diagnostic test change the likelihood that the disease is present?

2. Estimating the pre-test probability.
   - Literature
   - Regional variation
   - Clinical Experience

3. Finding the likelihood ratio of a test.
   - Some studies will give you this statistic.
   - Others will still give the sensitivity and specificity.
     LR for a positive test = [1 - Sensitivity/Specificity]
     LR for a negative test = Sensitivity/[1 - Specificity]

4. Use the Likelihood Ratio Nomogram to find the Post-test Probability
Can you apply this valid, important evidence about a diagnostic test in caring for your patient?
1. Is the diagnostic test available, affordable, accurate and precise in your setting?

2. Can you generate a clinically sensible estimate of your patient’s pre-test probability:
   • From practice data?
   • From personal experience?
   • From the report itself?
   • From clinical speculation?

3. Will the resulting post-test probabilities affect your management and help your patient?
   • Could it move you across a test-treatment threshold?
   • Would your patient be a willing partner in carrying it out?
   • Would the consequences of the test help your patient reach their goals in all this?
Diagnosis of acute sinus infections.

There are times when papers are so good that one wants to jump for joy, or weep because one hasn’t done it oneself. Bandolier felt this way about a beautiful demonstration of how to sort out diagnosis, in this case diagnosis of acute sinus infections in primary care [1].

Study

It was a prospective study of 357 patients intended to examine a range of symptoms and signs, and simple blood tests against computed tomography in the diagnosis of acute sinus infection. Some people couldn’t be included (pregnant women, for example), some didn’t want a scan, pressure on scan time meant that not all could be scanned. Information was therefore available on 201 people.

They had all received a diagnosis of acute sinus infection and were considered to be in need of antibiotic therapy. A host of symptoms and signs were recorded, and within two days of the clinical diagnosis they had a CT scan which included the entire nasal cavity and the paranasal sinuses. The scans were interpreted independently by two radiologists, with re-evaluation and consensus in case of disagreement. Sinusitis was defined as total opacification or fluid level in an ethmoid, sphenoid, frontal or maxillary sinus.

Altogether 17 symptoms and 10 signs were examined, together with results of ESR, C-reactive protein and white blood counts.

Results

In the 201 patients who underwent a CT scan, 127 (63%) had acute sinus infection diagnosed by the CT scan. When the signs, symptoms and blood tests were subjected to logistic regression analysis, it was found that only four were significantly associated with presence of infection. They were:

1. Purulent secretion in cavum nasi
2. Purulent rhinorrhea
3. Double sickening
4. ESR >10 mm/hr

"Double sickening" was defined as the presence of two phases in the illness history.

The likelihood ratios calculated for the presence of 0, 1, 2, 3 or 4 of these signs and symptoms in 199 of the 201 patients are shown in the Table.
## Diagnosing Acute Sinusitis

<table>
<thead>
<tr>
<th>Number of Signs and Symptoms</th>
<th>Sinusitis Present</th>
<th>Sinusitis Absent</th>
<th>Likelihood Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>43</td>
<td>1</td>
<td>25.2</td>
</tr>
<tr>
<td>3</td>
<td>41</td>
<td>13</td>
<td>1.8</td>
</tr>
<tr>
<td>2</td>
<td>32</td>
<td>23</td>
<td>0.8</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>22</td>
<td>0.2</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>14</td>
<td>0.1</td>
</tr>
<tr>
<td>Total</td>
<td>126</td>
<td>73</td>
<td>---</td>
</tr>
</tbody>
</table>

### Nomogram For Interpreting Diagnostic Test Result

[Diagram of the nomogram showing pretest probability, likelihood ratio, and post-test probability for different numbers of signs or symptoms.]
Comment

The implication of these findings are that acute sinusitis is over-diagnosed in primary care (though that has been shown before). For Norwegian physicians, a putative diagnosis of acute sinusitis is likely to be correct 63% of the time. If a patient has the four relevant signs and symptoms, that post-test probability increases to over 95%, with 3 of the four to about 80%. But if there are none, there is only a 15% chance that the patient has sinusitis.

Can this be applied generally? There is no obvious reason why not (Norwegian GPs usually have ESR and other testing available on site). In any event, this award-winning paper shows "how to do it" in sinusitis and in any other area. A pleasure and privilege to read, and a "must have" for anyone seriously interested in making evidence count in diagnosis.

References:
1. http://www.jr2.ox.ac.uk:80/bandolier/

Evidence Based Medicine  
Assessing the Literature  

II. Treatments  

A. Is this evidence about a treatment valid?  
B. Is this evidence about a treatment important?  
C. Can you apply this valid, important evidence about a treatment in caring for your patient?  

Are the results of this single study valid?  
1. Was the assignment of patients to treatments randomized and was the randomization list concealed?  
2. Were all patients who entered the trial accounted for at its conclusion, and were they analyzed in the groups to which they were randomized?  
3. Various “finer” points:  
   a) Were the patients and clinicians kept “blind” to which treatment was being received?  
   b) Aside from the experimental treatment, were the groups treated equally?  
   c) Were the groups similar at the start of the study?  

Is this evidence about a randomized trial important?  
1. Relative Risk Reduction: \( RRR = \frac{CER - EER}{CER} \)  
2. Absolute Risk Reduction \( ARR = CER - EER \)  
3. Number Needed to Treat. \( NNT = \frac{1}{ARR} \)  

Can you apply this valid, important evidence about a treatment in caring for your patient?  
1. Do these results apply to your patient?  
   • Is your patient so different from those in the trail that its results can’t help you?  
   • How great would the potential benefit of therapy actually be for your individual patient?  
2. Are your patients values and preferences satisfied by the regimen ant its consequences?  
   • Do your patient and you have a clear assessment of their values and preferences?  
   • Are they met by this regimen and its consequences?
Calculating NNTs

Using absolute risk reduction
Confidence Intervals

The NNT calculation is given here. An example calculating the NNT for oral sumatriptan from the data given on page 2 of the main issue is done on the worksheet. Methods for calculating NNTs from odds ratios and relative risk reduction were given in Bandolier 36, but we find that we don’t use these much.

The NNT calculation is given below. We need to distinguish between treatments, such as aspirin as an analgesic, and preventative measures, such as aspirin preventing further cardiac problems after myocardial infarction. Using the number outputs from systematic reviews is different depending on which you are looking at. The distinction is between treatment and prophylaxis. For prophylaxis, where fewer events occur in the group, the calculation shown will produce negative NNTs. You can use those (the number will be correct), or you can switch the active and control groups around to provide NNTs with a positive sign.

The NNT for prophylaxis is given by the equation 1/(proportion benefiting from control intervention minus the proportion benefiting from experimental intervention), and for treatment by 1/(proportion benefiting from experimental intervention minus the proportion benefiting from control intervention).

NNTs for treatment should be small. We expect large effects in small numbers of people. Because few treatments are 100% effective and because few controls - even placebo or no treatment - are without some effect, NNTs for effective treatments are usually in the range of 2 - 4. Exceptions might be antibiotics. The NNT for Helicobacter pylori eradication with triple or dual therapy, for instance, is 1.2 (Bandolier 12).

NNTs for prophylaxis will be larger, few patients affected in large populations. So the difference between treatment and control will be giving large NNTs. For instance, use of aspirin to prevent one death at five weeks after myocardial infarction had an NNT of 40 (Bandolier 17).

Using absolute risk reduction

The absolute risk reduction (ARR) is the difference between the event rate in the experimental group and the event rate in the control group. It is the denominator in the NNT calculation. Many reviews and trials provide this information, so if you have
it and convert it into a proportion, then you can get the NNT by dividing 1 by the
ARR: $\text{NNT} = \frac{1}{\text{ARR}}$

$$\text{NNT} = \frac{1}{\left(\frac{\text{IMP}_{\text{act}}}{\text{TOT}_{\text{act}}} - \frac{\text{IMP}_{\text{con}}}{\text{TOT}_{\text{con}}}\right)}$$

Where:
IMPact = number of patients given active treatment achieving the target
TOTact = total number of patients given the active treatment.
IMPcon = number of patients given control treatment achieving the target
TOTcon = total number of patients given the control treatment.

**Confidence Intervals**

The 95% confidence intervals of the NNT are an indication that 19 times out of 20 the
\textit{true} value will be in the specified range. An NNT an infinite confidence interval is
then but a point estimate; it includes the possibility of no benefit or harm. It may still
have clinical importance as a benchmark until further data permits finite confidence
intervals, but decisions must take this into account. A method for calculating
certainty intervals was given in Bandolier 18.

**L’Abbé plots**

A paper [3] by Kristen L’Abbé and colleagues written over ten years ago is regarded
by Bandolier as one of the most sensible and understandable ever written on
systematic reviews. The authors suggest a simple graphical representation of the
information from trials. Each point on a L’Abbé scatter plot is one trial in the review.
The proportion of patients achieving the outcome with the experimental intervention
is plotted against the event rate in controls. Even if a review does not show the data in
this way, you can do it yourself if the information is in the review, and that's why it's
part of Bandolier's worksheet.
For treatment, trials in which the experimental intervention was better than the control will be in the upper left of the plot, between the Y axis and the line of equality. If experimental was no better than control then the point will fall on the line of equality, and if control was better than experimental then the point will be in the lower right of the plot, between the X axis and the line of equality.

For prophylaxis this pattern will be reversed. Because prophylaxis reduces the number of bad events - such as death after myocardial infarction by the use of aspirin - we expect a smaller proportion harmed with treatment than with control. So if experimental is better than control the trial results cloud should be between the X axis and the line of equality.

These plots give a quick indication of the level of agreement among trials. If the points are in a consistent cloud, that gives some confidence that what we are seeing is a homogeneous effect. But if points are spread all over the graph, and especially if they cross the line of equality, then that should make us concerned about the intervention, or the patients being treated and their condition. This can also be called heterogeneity.

The important point about a L’Abbé plot is that it shows all of the extant data on one piece of paper. When combined with numbers in the trial, and a summary measure like NNT, it is a neat way to summarize lots of information.
Variation in treatment and control

One of the things that using systematic reviews in this way teaches you is just how variable are the effects of both treatment and control in randomized trials. It is legitimate to be surprised, but after a short time it seems that this is the norm.

The reasons for the variability are probably complex, but much will be just random chance. In many circumstances patients can have wide patterns of response to a treatment, but trial size is often relatively small. Gathering data together in systematic review and meta-analysis gives much more power than the single trial in almost all circumstances, and especially for reviews of treatments. Seeing such variability also teaches caution when faced with a single trial.
Size is everything

Take a moment to think about what you want to know about a treatment. You probably want some assurance that it works, but you really want to know how well it works. What do you mean by that? Using NNT terminology, you might want to know that the NNT is within certain limits.

Take ibuprofen. The NNT to obtain at least 50% pain relief in patients with moderate to severe pain over 4-6 hours is about 3 for 400mg ibuprofen compared with placebo. How close do you want the estimate to be? You probably wouldn’t be happy with a 95% confidence interval which went from 1 (perfect) to 10 (rotten). Would you be happy with 2 to 4, or happier still with 2.5 to 3.5?

The answer should be the last of these, but the narrower the confidence interval (the more correct you want the answer to be), the more patients you need to have studied. A mathematical but practical study [4] says that for the confidence interval to be 2.5 to 3.5 we need 500 patients taking ibuprofen and 500 taking placebo. The confidence interval with a single trial of the standard (in pain) of 40 patients per group is 1 to 10. The lesson is to beware the single trial reflex, changing practice on the basis of a single, small trial. it’s quite likely to be wrong. Random chance is in play, and has quite a big effect in small trials, which explains the scatter sometimes seen in L’Abbé plots, and is yet another reason for choosing evidence from systematic reviews.

Worthy of what?

Calculating NNTs is relatively straightforward compared with the greater complexity of deciding whether a trial is credible, or worthy only of dustbin. It is impossible to be dogmatic, as every subject has its own complexities. Here are some suggestions for your personal checklist:

Randomization: non-randomized trials over-estimate the effect of treatment. Unless there is a compelling reason, you should not believe or read non-randomized trials of treatments. Blinding: unblinded studies over-estimate the effect of treatment. Blinding may be difficult sometimes, so you should treat unblinded studies with extra caution. Withdrawals: studies with large numbers of dropouts should probably be treated with circumspection, unless it makes good sense to you.

Size: tiny studies aren’t worth your time. Large studies which are well done should carry particular weight.
Statistics: does the study do good statistical testing, like analysis of variance? If yes, then that’s good, but any study where the authors choose a single positive statistic out of many which are negative should go straight in the bin.

Statistical significance: p<0.05 isn’t that clever. It’s only 1 in 20, and you can roll two sixes with a couple of dice quite often. Weight the p<0.001 much more highly. Credible patient enrollment: just check out whether the patients at entry could demonstrate a change in whatever was being measured.

Outcomes: were the outcomes being measured at all valuable to doctors or patients, or were they just unsubstantiated surrogate measures?

Using the toolbox

These are all tools, not rules. We hope you find them useful to look at evidence that comes across your desk, especially when new treatments are being proposed. We would particularly like feedback on the NNT worksheet to know what you liked or disliked.

Bandolier’s NNT example sheet

A number needed to treat (NNT) is defined by a number of characteristics. This worksheet is designed as an aide memoir for working out NNTs from papers and systematic reviews. First fill in the answers to the questions, where appropriate, graph the data on the L’Abbé plot, and finally do the NNT calculation.

Tfelt-Hansen reviewed RCTs at that time on the benefit of Sumatriptan in treating migraine headaches. A total of 1,854 patients were given Sumatriptan 100 mg as a single dose for acute migraine headache and 1,036 were given a placebo. Of the intervention patients 1,067 had a successful outcome, while of the controls, 256 had a successful response.
<table>
<thead>
<tr>
<th>Question/Action</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>A What is the intervention (e.g. drug dose and frequency)</td>
<td>Oral Sumatriptan 100 mg, single dose</td>
</tr>
<tr>
<td>B What is the intervention for?</td>
<td>Acute migraine headache.</td>
</tr>
<tr>
<td>C What is the successful outcome (and when, or over what time did it occur?)</td>
<td>Pain mild or none at 2 hours.</td>
</tr>
<tr>
<td>D How many had the intervention</td>
<td>1854</td>
</tr>
<tr>
<td>E How many had successful outcome with the intervention?</td>
<td>1067</td>
</tr>
<tr>
<td>F Express this as a percentage (100 x E/D) and a proportion (E/D).</td>
<td>58% or 0.58</td>
</tr>
<tr>
<td>G What was the control or comparator?</td>
<td>Placebo tablet</td>
</tr>
<tr>
<td>H How many people had the control?</td>
<td>1036</td>
</tr>
<tr>
<td>I How many had successful outcome with the control?</td>
<td>256</td>
</tr>
<tr>
<td>J Express this as a percentage (100 x I/H) and a proportion (I/H).</td>
<td>25% or 0.25</td>
</tr>
</tbody>
</table>

Now graph the percentages for the trial on the graph from the percentages from F and J. This can be done for different outcomes of a trial, or individual trials in a systematic review or meta-analysis.

![L’Abbe Plot for Treatment](image)

Now calculate the NNT using the proportions from F and J.
NNT = \frac{1}{F - J} = \frac{1}{0.58 - 0.25} = \frac{1}{0.33} = 3.0

References:


