Clinical decision rules: how to use them

B Phillips

ABSTRACT
The first of this pair of papers outlined what a clinical decision rule is and how one should be created. This section examines how to use a rule, by checking that it is likely to work (examining how it has been validated), understanding what the various numbers that tell us about “accuracy” mean, and considers some practical aspects of how clinicians think and work in order to make that information usable on the front lines.

The first key steps in managing a patient’s problem consist of accurately assessing the situation, making a clear diagnosis, understanding the psychological and social impact of the illness and negotiating an effective treatment plan. The first of these, the arrival at an accurate diagnosis, if we are to undertake this using high-quality evidence, can be more difficult than we would like to admit.

The first of this pair of papers [whatever ref] outlined what a clinical decision rule (CDR) is and how one should be created. This section examines how to use a rule, by checking that it is likely to work (examining how it has been validated), understanding what the various numbers that tell us about “accuracy” mean, and considers some practical aspects of how clinicians think and work in order to make that information usable on the front lines.

VALIDATING A CDR
In order to even think about using a CDR, we should be fairly sure that it will work as we want it to. We would not use a drug that had just been developed on cell lines in our little patients, and so we should not do the equivalent with diagnostic or prognostic information (before you start to argue about the relative dangers of toxic chemicals and information, reflect on how a false positive or false negative test result has influenced your own life or medical practice). We need the CDR to have been validated—and by “validated”, I mean shown to accurately discriminate between the afflicted and the untouched—and to give accurate estimates of the proportion of patients with the affliction in each category. If we are using a “limping child” rule in the emergency department to guide our referrals to orthopaedics and early discharges, we want those with none, one, two, three or four predictors to have a meaningfully different risk of septic arthritis (ie, discriminate). We also want the no-predictor group to stay with a really low proportion, ideally zero, and the four predictor children to have a high proportion with septic arthritis, touching 95% or more, that is, retain accuracy.

It is almost always the case that the rule does not work quite as well in the validation set of patients. This is not unexpected, as the rule is derived from one particular group of patient’s data, and another dataset will be similar but not identical. There are a variety of modifications or recalibrations that can be undertaken to fine tune a rule and the modification then used more successfully in ongoing care.

What frequently happens is that the group that has developed a CDR goes on to test that it really works. This “temporal validation” provides some evidence to suggest that the rule really does work and often allows for optimisation of the rule to adjust for the overenthusiastic data-driven estimates that it initially created. Various approaches, from the complex to the straightforward, have been used to do this successfully.

The problem that remains is that the CDR might work in SuperHospital-W, where Dr X can train her residents into knowing exactly what point #2 of the CDR really means and what a positive or negative looks like, but can it work in SuperHospital-Y, where the physicians are restricted to using the data from the journal article? For some CDR, where the basis of the rule is quality-assured laboratory results and proven-to-be reproducible clinical signs or symptoms, this may be easy. For other rules, this may not be so effective (as a quick aside, it is often helpful to assess just how reproducible measurements are. There is a host of issues around this, and a very readable series of BMJ “Statistics Notes” by Profs Bland and Altman published in 1996, and their textbooks of practical medical statistics are accessible reads).

This next level of validation, seeing if the limping-child rule is rooted in only one hospital or not, looks at the CDR use in different physical locations but with similar clinical settings and has been called geographical validation. Potentially, you might want to see if it works across different clinical settings (domain validation), for example in the emergency department of secondary-care hospitals and walk-in primary-care clinics, but this may be irrelevant if you are working in the same setting as rule development occurred in.

The final step should be to demonstrate their efficacy in routine practice with multi-site randomised controlled trials. The trial does not seek to see if the rule differentiates folk, and
can be applied, but instead randomises between those who have the rule used, and those who do not, and captures in its results key patient- or health-system-important outcomes: minimising length of hospital stay or invasive testing or improved quality of life. This is hardly ever undertaken but can be a very powerful way of showing improved care with a diagnostic intervention of any sort.12

UNDERSTANDING ASSESSMENTS OF ACCURACY
Assuming you have found a rule that has been validated in some way and accepting it is unlikely to have gone through all these steps, then the next bit of information you will be wanting to understand is “how good is the rule anyway”? The information that a rule presents us with is often couched as a “result”: the rule is positive or negative (eg, Ottowa ankle rules), or the result is a value (eg, the neonatal outcome measure of CRIB score). These results are then described in the papers that examine them in terms of their diagnostic accuracy, using phrases that induce fear and occasionally nausea: sensitivity, specificity, likelihood ratios (LR) and predictive values.

Sensitivity and specificity
In its simplest form, a rule produces a positive or negative result. The group under study—take an emergency room setting for children presenting with fever—can be thought of as coming from two populations: those with the “disease”, for example pneumonia, and those without it. The proposed way to distinguish between these groups may be an invented blood test result—the “lung injury protein”, which goes up if you have a chest infection. The “test” is a “positive result” if the value is greater than a defined “threshold” (See fig 1).

The threshold then has part of populations on each side: the proportion of people with pneumonia who have a positive test is the sensitivity, and the proportion of people without pneumonia who have a negative test is the specificity.

Shifting where the threshold is drawn alters these proportions. As fig 2 shows, pushing the line for “positivity” upwards makes the test more specific (captures fewer people without the disease in the definition) but becomes less sensitive (fails to diagnose a greater number with the disease). The reverse is true when the level for positive results is reduced.

The situation has an added level of complexity with three levels of test result (eg, low, medium, high), but the core concept remains the same.
For clinical application, the important values are not the proportion of people with a pneumonia who have a positive test (sensitivity) and proportion of people without pneumonia who have a negative test (specificity) but instead, the proportion of people with a negative test result who truly did not have pneumonia (negative predictive value (NPV)) and the converse, the proportion of people with a positive test who did have pneumonia (positive predictive value (PPV)).

So, why do researchers continue to emphasise sensitivity and specificity? You might be tempted to think, as with medieval priests chanting Latin, this is to keep the information cloaked from the common doctor. You would not be entirely incorrect. However, there is a better reason. The clinically useful NPV and PPV are highly dependent on what the prevalence of the condition is in the population (also called the “baseline risk”) unlike the sensitivity and specificity, which although they are shop-floor useless, do not depend upon that value (generally, but if you take a completely different type of population, like using an adult thrombosis detection CDR in the NICU, for example, you are never going to get the same results). Take the example of testing our invented “lung injury protein” test. Figure 3 demonstrates how the positive and negative predictive values vary under different prevalence’s of disease (in the central panel, 10% prevalence, it shows how the positive test cases are made up of seven with pneumonia—the red squares—and 23 without—the blue squares. The negative test results are the three pink “missed” cases of pneumonia and 67 who really do not have pneumonia). The “correct” value of a positive test ranges from 13% to 55% depending on how prevalent pneumonia is in the population under study.

The upshot of this is that the direct translation of PPV and NPV from studies into clinical practice can only be undertaken if the prevalence of the condition is the same in your population as the study. If it is not, you need a little bit of calculation to work out what the values will be in your practice and if this would still be a viable rule.

ANOTHER APPROACH—LIKELIHOOD RATIOS
An alternative approach which captures the idea of “what a positive/negative test means” but does not rely on population prevalence is the use of LR. These values compare the proportion of patients with the disease and without the disease in for a given test result. In the LIP example where the cut-off is “93” (fig 1, Sensitivity 72.0%, Specificity 74.2%) these values would be LR+ (LR for a...
To compare the usefulness of these measures, have a look at tables 1 and 2. The tables show that with increasingly high cut-offs, the value of a positive test in diagnosing pneumonia improves, but a negative becomes less useful in ruling out the problem. It’s up to you which you find easiest to interpret.

The interpretation of these can be done arithmetically, or with a nomogram, or with a sort of “rule of thumb”:
- Likelihood ratio of $>10$ or $<0.1$ (1/10th)
- Big push towards or away from the diagnosis
- Likelihood ratio of $>5$ or $<0.2$ (1/5th)
- Smaller push towards or away from the diagnosis
- Likelihood ratio of $>2$ or $<0.5$ (0.5)
- A nudge towards or away from the diagnosis
- Likelihood ratio of closer to 1 than 2 or 0.5 (0.5)
- Very little diagnostic usefulness

To compare the usefulness of these measures, have a look at tables 1 and 2. The tables show that with increasingly high cut-offs, the value of a positive test in diagnosing pneumonia improves, but a negative becomes less useful in ruling out the problem. It’s up to you which you find easiest to interpret.

**THE PSYCHOLOGY OF DIAGNOSING**

Understanding how we physicians use a CDR, or any diagnostic information, requires a bit of knowledge about how doctors think. A number of researchers from a range of backgrounds have examined the diagnostic practices of physicians. There is a wealth of research that shows we do not generally “take a history and do a physical..."
and that we use a massive array of mental short-cuts (heuristics) which speed up our working practice, yet can lead us into dangerously wrong conclusions. In fitting CDRs into the practice of managing a clinical problem, it can be useful to take one straightforward model as to how doctors make a diagnosis.

The diagnostic process can be thought of in three stages: initiation, refinement and conclusion. The initiation stage is where a diagnosis, or differential, begins to be considered. The refinement is a working through of differentials, and the conclusion is a point where a decision to act has been made.

For most situations, initiation is the germ of a diagnostic idea: the acknowledgement that the feverish, crying baby presenting in the emergency department might have a number of causes for their illness, including bacterial meningitis. This first inklings requires further work: the refinement process. It is in this process that CDRs can help guide clinicians and lead to a diagnostic conclusion. Other approaches include a formal Bayesian analysis, pattern fitting or a stepwise rule-out of significant serious diagnoses.

The conclusion part of the process may be an actual diagnosis (“the child has pneumonia”), a rule-out (“the baby doesn’t have bacterial meningitis”) or a decision not to come to a conclusion (“we’re still not sure why your little one is poorly, and need to undertake a few more tests”).

An alternative model of diagnosis, present in situations such as seeing a toddler with trisomy 21, has a single eye-blink-fast step in it. You see the child, and you have made the diagnosis of Down’s syndrome before a thought process can be recognised. This is a pretty universal paediatric example, but there is work to suggest that the more expert a physician is in a particular area, the more rapidly they come to a diagnosis and that this in part is because of pattern recognition or fitting new cases to a mental “categorisation”. In these settings, CDRs can struggle to get a (mental) look-in to the process.

Yet, another set of theories (“dual process theory”) suggests that stepwise, rational approach and the rapid fast-heuristic conclusion approach can happen at the same time. Some clinicians favour the “rationalising” approach, other the “clinical intuition” approach, but both may be occurring. If there is a conflict, for example, between what the CDR suggests and “intuition” delivers, the person’s tendency may tip the balance of the thinking.

CDRs should help our thinking by providing good-quality guidance to avoid diagnostic errors and minimise unnecessary tests. Commonly diagnostic errors occur because of systems and cognitive errors. Such errors can include a failure to syntheses diagnostic information correctly and come to a premature diagnostic conclusion; a lack of appreciation of the value of a sign or symptom in making a diagnosis or exaggeration of the accuracy of a test finding. Other reasons for misdiagnosis would not be helped by the use of CDRs, such as the true diagnosis being rare or failure in the technical skill of the individual doctor in reading an x-ray or eliciting a physical sign.

**USING CDRS IN PRACTISE**

There remains, however, an almost emotional difficulty in turning to a CDR, when instead, we should be like House, Holmes or Throndyke in making our diagnoses from skill and knowledge. This is despite the knowledge that in many cases a CDR performs better than our in-built doctoring.

It may be that we are un-nerved by now knowing that there is a 2% chance of septic arthritis, rather than being safe in the ignorance that “children like that one” in “my clinical experience” do not have an infected joint. Chance alone could mean you send home ~150 children and not get one returning to the orthopods despite that group having a 2% chance of infection (it also raises questions about what a diagnosis really is: “the truth” or just a high enough probability of “the truth”).

There may be other issues, such as an objection to reduced professional autonomy, a perception of impracticality or a belief that the rule does not fit the population we are dealing with. It might reflect that we are just the sorts of people who prefer working with the heuristics of medical experience that we have built over time. This question, as to why we do not follow where the best evidence should steer us, is a matter of ongoing debate and research.

On a positive note, the most effective uses of CDRs seem to have been where the rule has a clear clinical usefulness, it is championed by well-respected local clinicians and it makes life better for patients and clinicians. Lots of these features can be picked up and put in place when you are trying to promote the use of a CDR. You will have decided, from an analysis based on the content of these two articles, whether the rule is technically good and likely to be true. You can assess how much of an impact it should make on your service, and you will be well placed to identify the key players who need to be “on side” in order to get the rule used (Who will it benefit? Let them know, and let them know how it will be good for them! Who will it disadvantage—with more work or nastier testing? Use case stories and hard stats together to show how overall it makes things better for the patients, who are the key cause of concern for us all).

**CONCLUSIONS**

CDRs are potentially a way of making our diagnoses more accurate and/or quicker and less unpleasant to achieve. They have not always done this and their creation and testing is a potential minefield, as these articles have examined. The essential principles that should guide us in our clinical
work, when seeking to use a CDR, should remain the same as in all other areas. We should seek to use evidence that is valid, important and makes a difference to patients. After all, that is why we became paediatricians, is it not?

Acknowledgements The author thanks to Professor Lesley Stewart for her helpful comments on an earlier draft of this manuscript.

Funding This work was funded as part of an MRC Research Training Fellowship.

Competing interests None.

Provenance and peer review Commissioned; externally peer reviewed.

REFERENCES

Clinical decision rules: how to use them

B Phillips

doi: 10.1136/adc.2009.174458

Updated information and services can be found at:
http://ep.bmj.com/content/95/3/88.full.html

These include:

References
This article cites 24 articles, 10 of which can be accessed free at:
http://ep.bmj.com/content/95/3/88.full.html#ref-list-1

Email alerting service
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/