Sensitivity and specificity: imperfect predictors of guideline utility in radiology

^{1,2}B ROUDSARI, MD, MPH, PhD, ³C MCKINNEY, PhD, MPH, ⁴D MOORE, MD and ^{1,2,5,6}J JARVIK, DM, MPH

¹Department of Radiology, ²Comparative Effectiveness, Cost and Outcomes Research Center, ³School of Public Health, University of Washington, Seattle, WA, ⁴Department of Radiology, University of Texas Southwestern Medical Center, Dallas, TX, ⁵Department of Neurological Surgery and ⁶Department of Health Services, School of Public Health, University of Washington, Seattle, WA

Objective: Many “guideline-development studies” have presented the sensitivity (SN) and specificity (SP) of a new decision tool to describe the potential improvements in utilisation of imaging techniques as a result of adopting the new guideline. However, SN and SP are measures designed to assess how well a new guideline compares with a gold standard. These measures do not evaluate how many patients with a positive test actually have the disease; nor do they evaluate how many patients with a negative test do not have the disease. To evaluate these characteristics of a decision tool, other measures, namely the positive predictive value (PPV) and negative predictive value (NPV), should be calculated. This report highlights some of the main methodological challenges in interpretation of the studies that attempt to evaluate the development of an imaging guideline and the effectiveness of an imaging guideline in real world practice.

Methods: We define four key measures of a decision tool: SN, SP, PPV and NPV. Using data from two hypothetical populations, we explain how these measures can be calculated and interpreted. We place special emphasis on the purpose of and differences between the SN–SP and PPV–NPV.

Results: Borrowing information from two studies, we demonstrate how these measures should be used in the radiology healthcare services research to evaluate decision guidelines.

Conclusion: The use of appropriate measures for the specific question at hand will ensure the guidelines are useful, safe, cost reducing and effective clinical tools.

Received 18 June 2009
Revised 12 October 2009
Accepted 28 October 2009

DOI: 10.1259/bjr/20598117

© 2011 The British Institute of
Radiology

Several studies have demonstrated a substantial increase in the use of imaging technology over the past two decades in developed countries [1–3]. This is particularly so in the United States, which has the highest per capita healthcare expenditure [4]. In a recent report to the United States Congress, the Government Accountability Office found that between 2000 and 2006, Medicare (a government programme of medical care targeting individuals ≥ 65 years old) spent more than \$14 billion on imaging technology [1]. Despite this surge in utilisation pattern, studies have not been able to demonstrate a positive correlation between higher utilisation and better patient outcome [5]. In addition, there is substantial geographical heterogeneity in the utilisation of imaging technology [2, 5]. In 1994, Olsson and Inamura [2] compared annual radiology examinations between the Scandinavian countries and Japan. They demonstrated that the Japanese receive a substantially higher number of examinations in comparison with Scandinavians (1600 per 1000 individuals per year *vs* 600 per 1000 individuals per year). The greater availability of more advanced imaging technologies such as CT in Japan (1 scanner per 17 000) compared with Scandi-

navian countries (1 scanner per 95 000) might partially explain such an observation.

These findings suggest that other factors beyond a patient's medical condition might influence inappropriate or overuse of these services. This issue has numerous potentially undesirable consequences both for patients and for healthcare systems: subsequent costly evaluations and emotional stress owing to false-positive results, exposure to unnecessary radiation and increasing the cost of healthcare reflect only the tip of this iceberg [6].

Clinical decision rules [6], appropriateness criteria [7, 8] and imaging guidelines [9, 10] are decision tools designed to minimise unnecessary utilisation of imaging techniques. These tools enable physicians to categorise patients as high or low risk for a certain disease. On the basis of these guidelines, imaging studies are requested only for at-risk patients who have a high chance of having abnormal findings in imaging evaluations. These decision tools also aim to alleviate the financial burden of unnecessary evaluations on healthcare systems and to reduce preventable radiation exposure to patients. It is essential that these tools effectively differentiate affected from non-affected individuals; ideally they should neither miss a covert disease (false-negative) nor inappropriately label an unaffected individual as diseased (false-positive).

One frequently evaluated imaging technique in radiology health services research is CT of the head for mild head

Address correspondence to: Bahman S Roudsari, MD, MPH, PhD, Assistant Professor of Epidemiology University of Texas, School of Public Health, 5323 Harry Hines, V8.112, Dallas, Texas 75390-9128, USA. E-mail: roudsari@u.washington.edu

trauma [11–13]. Although there is consensus that severe head trauma should be evaluated by head CT scan if clinically warranted, there is debate about requesting head CT scan for mild-to-moderate head trauma, which typically have a low chance of intracranial lesions. As a result, researchers have attempted to develop different imaging guidelines to differentiate between mild-to-moderate head trauma with and without intracranial lesions. For example, the National Institute of Health and Clinical Excellence (NICE) published “Head injury triage, assessment, investigation and early management of head injury in infants, children and adults, clinical guideline” that aims to improve patient care and minimise unnecessary use of head CT by triaging head trauma patients into low- or high-risk groups [11]. Only high-risk patients thought to have a greater chance of intracranial lesions are referred for evaluation by head CT scan [11]. In order to triage a patient to high-risk or low-risk, the guideline uses nine demographic and injury-related characteristics. The presence of any one of these characteristics puts the patient in the high-risk category, indicating that a CT scan should be requested. These characteristics are: Glasgow Coma Scale (GCS) <13 at the time of admission or <15 during the first 2 h after injury; retrograde amnesia more than 30 min; suspected open, depressed or basal fracture of skull; recurrent vomiting; age ≥ 65 years; any focal neural deficit; any coagulopathy; any seizure; and experiencing a “dangerous” injury mechanism [11]. An injury is considered dangerous if the patient is ejected from a motor vehicle, is a pedestrian struck by a motor vehicle or has fallen from ≥ 3 feet or ≥ 5 steps provided some loss of consciousness or amnesia has been experienced [11].

Many “guideline-development studies”, including those evaluating the NICE guideline, have presented the sensitivity (SN) and specificity (SP) of a new decision tool to describe the potential improvements in use of imaging techniques as a result of adopting the new guideline. However, SN and SP are measures designed to assess how well a new guideline compares with a gold standard (e.g. the NICE guideline *vs* head CT scan as a gold standard for diagnosis of intracranial lesions). By definition, a gold standard is a test that is considered the definitive approach for diagnosing a disease. Ideally, this test should neither miss any sick person nor falsely label a healthy person as a sick individual [14, 15].

Unfortunately, SP and SN do not evaluate how many patients with a positive test actually have the disease, nor do they evaluate how many patients with a negative test do not have the disease. To evaluate these characteristics of a decision tool other measures, namely the positive predictive value (PPV) and negative predictive value (NPV), should be calculated.

This report highlights some of the main methodological challenges in the interpretation of studies that attempt to evaluate the development of an imaging guideline and the effectiveness of an imaging guideline in real-world practice. We define four key measures of a decision tool: SN, SP, PPV and NPV. Using data from two hypothetical populations, we explain how these measures can be calculated and interpreted. We place special emphasis on the purpose of and differences between the SN and SP values and the PPV and NPV. Borrowing information from studies conducted by Stein et al [13] and Smits et al [12], we demonstrate how these measures should be used in the radiology healthcare services research to evaluate decision guidelines.

Sensitivity and specificity of a decision tool

Briefly stated, SN is the ability of a test, such as the NICE head CT guideline, to correctly identify individuals with important head injuries among all those who actually have important head injuries [14, 15]. On the other hand, SP is the ability of the test to identify non-injured individuals among those who actually do not have important head injuries (Table 1) [14, 15]. We present the SN and SP of two hypothetical studies that used the NICE guideline to identify high-risk individuals referred for CT scan for intracranial lesions relative to a gold standard (Table 2). We would like to emphasise that selection of the NICE guideline was solely for didactic purposes of this methodology-orientated article. Readers can replace this guideline with any other guideline and reach the same overarching conclusion. The populations in the two hypothetical studies differ in two aspects. Firstly, Population 1 is much smaller than Population 2 (600 *vs* 5100 individuals). Secondly, the frequency of intracranial lesions in Population 1 is almost nine times that in Population 2.

In Population 1, 95 of the 100 mild-to-moderate head trauma patients with known intracranial lesions were correctly identified by the NICE guideline as high-risk individuals. As a result, the NICE guideline in this hypothetical example had an SN of 95% (95/100). In five cases, patients were falsely categorised as low risk (false-negative results). Thus, among those who had an intracranial lesion, 95% received a head CT scan. Among 500 mild-to-moderate head trauma patients without intracranial lesions in Population 1, only 160 cases were correctly identified as low risk; the rest (340 patients) were falsely labelled according to the NICE guidelines as high risk (Table 2). As a result, in this hypothetical example, the NICE guideline had 32% (160/500) SP in Population 1. Thus, among those with no intracranial lesion, 68% were identified as high risk and would have had a CT scan

Table 1. Calculation of SN, SP, PPV and NPV of a new screening test [18]

	Gold standard test			SN (column analysis)	SP (column analysis)	PPV (row analysis)	NPV (row analysis)
	Positive	Negative	Total				
Screening test Positive	A	B	A+B	A/(A+C)	D/(B+D)	A/(A+B)	D/(B+D)
Negative	C	D	C+D				
Total	A+C	B+D	A+B+C+D				

NPV, negative predictive value; PPV, positive predictive value; SN, sensitivity; SP, specificity.

Table 2. Comparison of PPV and NPV in two hypothetical populations with different prevalence of intracranial lesions among head trauma patients

	With intracranial lesion	Without intracranial lesion	Total	Prevalence of intracranial lesion	SN	SP	PPV	NPV
Population 1								
High risk ^a	95	340	435					
Low risk ^b	5	160	165					
Total	100	500	600		95/100 (95%)	160/500 (32%)	95/435 (22%)	160/165 (97%)
				100/600 (17%)				
Population 2								
High risk ^a	95	3400	3495					
Low risk ^b	5	1600	1605					
Total	100	5000	5100		95/100 (95%)	1600/5000 (32%)	95/3495 (3%)	1600/1605 (~100%)
				100/5100 (2%)				

High-risk and low-risk were determined according to the National Institute of Health and Clinical Excellence (NICE) guideline. NPV, negative predictive value; PPV, positive predictive value; SN, sensitivity; SP, specificity. ^aCT required. ^bCT not required.

requested. The SN and SP for the NICE guideline in Population 2 were the same as for Population 1.

We calculated SN (and SP) of the NICE guideline by comparing the number identified by the decision tool as positive (or negative) to the total number of patients known to be positive (or negative) for an intracranial lesion based on some gold standard method; in this case, CT scan was the gold standard and had been conducted on all patients. Thus, the main purpose of the SN and SP calculations was evaluation of the validity of a new tool (*i.e.* the NICE guideline) in comparison with a gold standard (*i.e.* head CT scan). In this particular type of guideline-development study diagnosis was not the objective, as the disease status had already been confirmed. Rather, the purpose was to evaluate how accurately the NICE guideline identified those with and without the disease.

Positive and negative predictive values of the NICE guideline

Once evaluated, guidelines demonstrated to have satisfactory SN and SP are next applied to clinical practice and the question of interest typically changes. At this juncture, we want to know "What is the probability that a patient considered high-risk based on the NICE guideline has an intracranial lesion?" Another extremely relevant but similarly important question is "What is the probability that a patient identified as low-risk based on the NICE guideline does not have an intracranial lesion?" To address these questions we turn to two other characteristics of a decision tool, namely the PPV and the NPV.

PPV is a measure that reflects the probability of disease in the presence of a positive test result [14, 15] (Table 1). NPV is the probability of being disease-free in the presence of a negative test result [14, 15]. We revisit our hypothetical studies (Table 2), which demonstrate that the PPV of the NICE guideline for Population 1 was 22% (95/435); in Population 2, the PPV dropped to only 3% (95/3495). In other words, of the patients identified as high risk by the NICE guidelines and given a CT scan,

only 22% in Population 1 and 3% in Population 2 had intracranial lesions. The NPV of the NICE guideline was much higher for Populations 1 and 2 at 97% (160/165) and 99.7% (1600/1605), respectively. This means that if a person was identified as low risk by the NICE guidelines in Population 1, the chance of missing an intracranial lesion was less than 3%. In Population 2 this chance was less than 1%.

Using these two hypothetical populations, we demonstrated that although the SN (95%) and SP (32%) of a test was the same in both populations, the PPV and NPV of these tests differed (Table 2). This is because SN and SP are stable characteristics of a test that usually do not change with disease prevalence. By contrast, PPV and NPV will vary according to the prevalence of the disease in the population: the higher the prevalence of the disease, the higher the PPV of a test (Table 1). In our hypothetical example, the lower prevalence of intracranial lesions among head trauma patients in Population 2 resulted in a substantially lower PPV in this population. As a result, relative to Population 1, use of the NICE guideline resulted in more unnecessary head CT scans in Population 2.

Evaluation of assessments of imaging guidelines

We now evaluate information from Smits et al [12] and Stein et al [13] who compared the SN and SP of several head CT scan rules/guidelines, including the NICE guideline (Table 3).

In Smits' study, the NICE guideline had high SN (99%) and somewhat low SP (31%) for identifying any intracranial lesion [12] (Table 3). In Stein's study, the guideline had a lower SN (69%) but almost two-times higher SP (65%) for identifying "neurocranial lesions", "clinically relevant lesions" or "neurosurgical lesions" [13]. Using published data from these studies, we evaluated the PPV and NPV. Among patients with a positive NICE result who received a CT scan, only a small proportion actually had intracranial lesions. This was true in both Stein et al's and Smits et al's study:

Table 3. Comparison of the potential influence of the National Institute of Health and Clinical Excellence (NICE) guideline on utilisation of head CT scan in Stein et al [13] and Smit et al's [12] studies

	With intracranial lesion	Without intracranial lesion	Total	Prevalence of intracranial lesion	SN	SP	PPV	NPV
Stein's study								
High risk ^a	395	3151	3546					
Low risk ^b	177	5820	5997					
Total	572	8971	9543					
				572/9543 (6%)	395/572 (69%)	5820/8971 (65%)	395/3546 (11%)	5820/5997 (97%)
Smits' study								
High risk ^a	528	5121	5649					
Low risk ^b	5	2301	2306					
Total	533	7422	7955					
				533/7955 (7%)	528/533 (99%)	2301/7422 (31%)	528/5649 (9%)	2301/2306 (≈100%)

NPV, negative predictive value; PPV, positive predictive value; SN, sensitivity; SP, specificity. ^aCT required. ^bCT not required.

the PPV was 11% (395/3546) and 9% (528/5649), respectively. However, in both studies, the vast majority with a negative NICE result did not have intracranial lesions. The NPV was 97% (5820/5997) in Stein et al's study and almost 100% (2301/2306) in Smits et al's study.

The PPV is the proportion of patients characterised as high-risk (using the NICE guideline) who had a head CT scan that was necessary. As a result, a low PPV indicates that a substantial proportion of patients without intracranial lesions were tested, thereby utilising unneeded healthcare resources. The NPV reflects the proportion of low-risk patients (based on the NICE guideline) without intracranial lesions. Even when the NPV is high, the number of patients with intracranial lesions missed by the NICE guidelines need to be carefully considered; issues of missed intracranial lesions could have serious health consequences for the patient and liability issues for the provider. In the case of the Stein et al study, an estimated 30 patients with intracranial lesions will be missed per 1000 patients evaluated using the NICE guideline. The acceptability of these figures as they apply to healthcare utilisation research and clinical practice must weigh the benefits (reduced costs) and risks (radiation experienced by all if the gold standard were applied) to the limitations of the NICE guideline applied to real-world clinical practice.

Our analysis demonstrates that, although there was a substantial difference in the SN and SP of the NICE guideline between the studies of Stein et al and Smits et al, the proportion of individuals that were unnecessarily evaluated by head CT scan was similar (89% *vs* 91%), even though the absolute numbers were different (3100 *vs* 5100).

Conclusion

Our aim was to describe the way in which decision tools should be adopted in radiology health services research to evaluate their development and effectiveness in practice. Once a decision tool is developed and validated in different settings (*i.e.* SN and SP of the test are evaluated), these measures no longer address the questions of interest. Rather, we are interested in

evaluating the effectiveness of the tool as it is used in a real-world setting. For this purpose we rely on the PPV and NPV to assess the utility of the tool in minimising unnecessary utilisation of healthcare resources and risks to patients. Each of these measures is important in the evaluation of a decision tool; however, the specific question being addressed will dictate whether SN and SP or PPV and NPV are appropriate. The use of the appropriate measures for the specific question at hand will ensure that guidelines are useful, safe, cost-reducing and effective clinical tools.

Acknowledgment

This study has been funded by National Institute on Alcohol Abuse and Alcoholism (NIAA) through grant number R01AA017497.

References

1. Medicare Part B Imaging Services. Rapid spending growth and shift to physician offices indicate need for CMS to consider additional management practices. Washington DC: Government Accountability Office, 2008. Available from: <http://www.gao.gov/new.items/d08452.pdf> (Accessed 10th December 2009).
2. Olsson S, Inamura K. Diffusion of digital radiology modalities in the Nordic countries and Japan. *Comput Methods Programs Biomed* 1994;43:9–13.
3. Toms AP, Cash CJ, Linton SJ, Dixon AK. Requests for body computed tomography: increasing workload, increasing indications and increasing age. *Eur Radiol* 2001;11:2633–7.
4. Keehan S, Sisko A, Truffer C, Smith S, Cowan C, Poisal J, et al. Health spending projections through 2017: the baby-boom generation is coming to Medicare. *Health Aff (Millwood)* 2008;27:w145–55.
5. Fisher ES, Wennberg DE, Stukel TA, Gottlieb DJ, Lucas FL, Pinder EL. The implications of regional variations in Medicare spending. Part 2: health outcomes and satisfaction with care. *Ann Intern Med* 2003;138:288–98.
6. Reed MH. Clinical decision rules in radiology. *Acad Radiol* 2006;13:562–5.
7. Siström CL. The ACR appropriateness criteria: translation to practice and research. *J Am Coll Radiol* 2005;2:61–7.

8. Siström CL. In support of the ACR appropriateness criteria. *J Am Coll Radiol* 2008;5:630–5;discussion 636–7.
9. Needham G, Grimshaw J. Referral Guidelines For Imaging, Guidelines for Healthcare Professionals who prescribe Imaging Investigations involving Ionising Radiation. Aberdeen: University Court of the University of Aberdeen, 2008. Available from: http://ec.europa.eu/energy/nuclear/radio-protection/publication/doc/118_update_en.pdf (Accessed 10th December 2009).
10. RCR Working Party. Making the best use of a department of clinical radiology: guidelines for doctors 6th edn. London: The Royal College of Radiologists, 2007.
11. National Collaborating Centre for Acute Care. National Institute for Clinical Excellence, head injury triage, assessment, investigation and early management of head injury in infants, children and adults, clinical guideline 56, 2007. Available from: <http://www.nice.org.uk/nicemedia/pdf/CG56NICEGuideline.pdf> 10/5/2009 (Accessed 10th December 2009).
12. Smits M, Dippel DW, de Haan GG, Dekker HM, Vos PE, Kool DR, et al. Minor head injury: guidelines for the use of CT—a multicenter validation study. *Radiology* 2007; 245:831–8.
13. Stein SC, Fabbri A, Servadei F, Glick HA. A critical comparison of clinical decision instruments for computed tomographic scanning in mild closed traumatic brain injury in adolescents and adults. *Ann Emerg Med* 2009;53:180–8.
14. Friis R, Sellers T. Epidemiology for public health. 4th edn. Sudbury, MA: Jones and Bartlett, 2008.
15. Koepsell TD, Weiss NS. Epidemiologic methods: studying the occurrence of illness. Oxford, UK: Oxford University Press, 2003.