Diagnostic Performance of an Expert System for Interpretation of ^{99m}Tc MAG3 Scans in Suspected Renal Obstruction

Andrew Taylor[1], Ernest V. Garcia[1], Jose Nilo G. Binongo[2], Amita Manatunga[2], Raghuveer Halkar[1], Russell D. Folks[1], and Eva Dubovsky[3]

[1]Department of Radiology, Emory University School of Medicine, Atlanta, Georgia; [2]Department of Biostatistics, Emory University School of Medicine, Atlanta, Georgia; and [3]Department of Radiology, University of Alabama, Birmingham, Alabama

The purpose of the study was to compare diuresis renography scan interpretation generated by a renal expert system with the consensus interpretation of 3 expert readers. **Methods:** The expert system was evaluated in 95 randomly selected furosemide-augmented patient studies (185 kidneys) obtained for suspected obstruction; there were 55 males and 40 females with a mean age ± SD of 58.6 ± 16.5 y. Each subject had a baseline ^{99m}Tc-mercaptoacetyltriglycine (^{99m}Tc-MAG3) scan followed by furosemide administration and a separate 20-min acquisition. Quantitative parameters were automatically extracted from baseline and furosemide acquisitions and forwarded to the expert system for analysis. Three experts, unaware of clinical information, independently graded each kidney as obstructed/probably obstructed, equivocal, and probably nonobstructed/nonobstructed; experts resolved differences by a consensus reading. These 3 expert categories were compared with the obstructed, equivocal, and nonobstructed interpretations provided by the expert system. Agreement was assessed using weighted κ, and the predictive accuracy of the expert system compared with expert readers was assessed by the area under receiver-operating-characteristic (ROC curve) curves. **Results:** The expert system agreed with the consensus reading in 84% (101/120) of nonobstructed kidneys, in 92% (33/36) of obstructed kidneys, and in 45% (13/29) of equivocal kidneys. The weighted κ between the expert system and the consensus reading was 0.72 and was comparable with the weighted κ between experts. There was no significant difference in the areas under the ROC curves when the expert system was compared with each expert using the other 2 experts as the gold standard. **Conclusion:** The renal expert system showed good agreement with the expert interpretation and could be a useful educational and decision support tool to assist physicians in the diagnosis of renal obstruction. To better mirror the clinical setting, algorithms to incorporate clinical data must be designed, implemented, and tested.

Received Jul. 20, 2007; revision accepted Oct. 29, 2007.

For correspondence or reprints contact: Andrew Taylor, MD, Division of Nuclear Medicine, Emory University Hospital, 1364 Clifton Rd., NE, Atlanta, GA, 30322.

E-mail: ataylor@emory.edu

Key Words: RENEX; ^{99m}Tc-MAG3; furosemide; diuresis renography; renal obstruction

J Nucl Med 2008; 49:216–224
DOI: 10.2967/jnumed.107.045484

To evaluate suspected ureteral obstruction using diuresis renography, a recent international consensus panel recommended a protocol consisting of baseline radionuclide imaging with ^{99m}Tc-mercaptoacetyltriglycine (^{99m}Tc-MAG3) followed by furosemide administration and an additional 15 min of imaging (*1*). In spite of guidelines, it may be challenging for general nuclear medicine physicians and radiologists to acquire the experience and expertise in diuresis renography to interpret these studies with confidence. Physicians are required to assimilate a continuously expanding technical and interpretative knowledge base and apply it to specific tasks at the same time as the hours to acquire this knowledge base are steadily shrinking due to an increase in the average number of images per study as well as the pressure to increase the absolute number of studies each radiologist interprets. It is especially challenging to develop a high degree of competence for low-volume studies such as diuresis renography in which radiologists and some nuclear medicine physicians may have had limited training and experience. In fact, a large percentage of the estimated 590,000 renal scans performed annually in the United States are interpreted at sites that perform <3 studies per week (*2*). Development and implementation of decision support tools have the potential to help physicians interpret low volume studies at a faster rate and with higher levels of confidence and expertise.

To address the specific problem of diuresis renography, we have developed a renal expert system (RENEX) for detecting renal obstruction using pre- and post-furosemide ^{99m}Tc-mercaptoacetyltriglycine (^{99m}Tc-MAG3) renal scans (*3*). Briefly, RENEX consists of (a) a "parameter knowledge library" with the list of the boundary conditions necessary for transforming the values of each quantitative parameter—such as time to peak height of the renogram curve or time to

half-maximum counts (T1/2)—to a certainty factor describing the degree of abnormality or normality of that parameter; (b) a "knowledge base" of heuristic rules that uses certainty factors describing the degree of normality or abnormality of specific parameters to generate new certainty factors specifying the likelihood of obstruction (supplemental material listing rules used by RENEX is available online only at http://jnm.snmjournals.org); and (c) an "inference engine" to combine the certainty factors of the rules and parameters to reach a final certainty factor (conclusion) with regard to obstruction (*3,4*). Certainty factors range from −1.0 (no obstruction) to +1.0 (obstruction). RENEX was optimized using pilot data (*3*) and implemented so that +0.2 would specify the threshold between obstructed and equivocal and −0.2 would specify the threshold between nonobstructed and equivocal. In the pilot study (*3*), there was excellent agreement between RENEX and the consensus reading of 3 experts as to whether there was obstruction of the kidneys; however, the pilot study served only as proof of concept because the pilot data also served as the training set to develop RENEX, and it was possible that the results and the +0.2 and −0.2 thresholds applied only to that specific dataset. The goals of this present study were to (a) evaluate the overall predictive accuracy of RENEX in a randomly selected sample, (b) compare the performance of RENEX with each individual expert, and (c) confirm that the use of +0.2 as the certainty factor to represent the threshold between obstruction and equivocal and the use of −0.2 as the threshold between nonobstruction and equivocal were appropriate thresholds in a clinical setting.

MATERIALS AND METHODS

Patient Selection and Data Processing

Data collection and database use were compliant with the terms of the Health Insurance Portability and Accountability Act and followed institutional review board approval with waiver of informed consent. RENEX was evaluated in 95 randomly selected furosemide-augmented renography studies (185 kidneys) obtained because of suspected obstruction. In approximately one third of patients referred for possible obstruction, the baseline acquisition excludes obstruction and furosemide is not administered (*5*); consequently, our study was biased toward more difficult cases as obviously nonobstructed patients (no furosemide required) were excluded from the study population. There 55 males and 40 females with a mean age ± SD of 58.6 ± 16.5 y. Patients with renal transplants were excluded because our normal database included only subjects with native kidneys (*6*).

Acquisition Protocol

Patients were positioned supine, with the scintillation camera detector placed under the table. A 3-phase dynamic acquisition was begun at the time of injection of approximately 370 MBq (10 mCi) of ^{99m}Tc-MAG3. Phase 1 consisted of twenty-four 2-s frames, phase 2 was sixteen 15-s frames, and phase 3 was forty 30-s frames. All patient studies were processed using our QuantEM 2.0 renal quantification program to generate the input parameters for RENEX. The QuantEM software (licensed by Emory University to GE Healthcare), developed specifically for ^{99m}Tc-MAG3, has been validated in a multicenter trial and generates specific quantitative parameters recommended for scan interpretation as well as calculating a ^{99m}Tc-MAG3 clearance using a camera-based technique (*7,8*).

In summary, for the baseline renogram, QuantEM sums a static image from the 2- to 3-min postinjection frames. Using a filtered version of this image, whole kidney, background, and cortical regions of interest (ROIs) are automatically defined. Technologists approve or modify these automatically assigned kidney, cortical, and background ROIs. Background-subtracted whole-kidney and cortical curves are then generated, and 47 quantitative parameters are generated, including patient demographics (height, weight, age, sex, body-surface area), curve parameters (time to peak counts and 20 min to count ratio for both whole-kidney and cortical ROIs), voiding indices (baseline postvoid to maximum count ratios and furosemide prevoid to baseline maximum count ratios), relative uptake, and the ^{99m}Tc-MAG3 clearance. The ^{99m}Tc-MAG3 clearance is calculated from the 1- to 2.5-min whole-kidney uptake of ^{99m}Tc-MAG3 corrected for renal depth and attenuation and the preinjection and postinjection images of the dose syringe (*7–10*). QuantEM 2.0 generates additional parameters for input into RENEX and incorporates several quality-control procedures to improve reproducibility (*3,11*). Experts had access to all of the parameters provided to RENEX as well as the normal ranges for the standard parameters.

The furosemide component of the study was a separate acquisition consisting of forty 30-s frames. Furosemide is administered at the start of the furosemide acquisition. The time between the initial injection of ^{99m}Tc-MAG3 and the injection of furosemide varied because the baseline study was always reviewed to determine if the baseline acquisition could exclude obstruction and, consequently, if the furosemide acquisition could be omitted. When furosemide was administered, the time between the injection of ^{99m}Tc-MAG3 and the injection of furosemide was ≥30 min. The standard dose of furosemide is 40 mg but the nuclear medicine physician monitoring the study sometimes increases the dose of furosemide to 60 or 80 mg if the ^{99m}Tc-MAG3 clearance on the baseline study is reduced or if the patient is known to have an elevated creatinine level. Technologists approve or modify automatically assigned kidney and background ROIs and assign pelvic ROIs and the time interval for the calculation of the T1/2. Quantitative parameters are automatically extracted from the 2 acquisitions, placed in an XML file, and forwarded to RENEX for analysis.

Expert Readers and Scoring

The readers were defined as "expert" on the basis of the fact that each reader had >20 y experience in full-time academic nuclear medicine, had multiple publications in renal nuclear medicine, and had been invited to give renal nuclear medicine educational sessions at national radiology and national nuclear medicine meetings. Each expert independently scored each kidney for the presence of obstruction based on a 5-point scale: 1, not obstructed; 2, probably not obstructed; 3, equivocal; 4, probably obstructed; and 5, obstructed. The consensus reading was determined by majority vote; when there was substantial disagreement between 1 or more readers, a conference of the 3 readers was used to determine a consensus reading.

RENEX

The architecture of RENEX is the subject of a separate publication (*3*); it was inspired by 2 previously developed expert systems,

MYCIN (an antibiotic suffix) and PERFEX (licensed by Emory University to Syntermed; for perfusion expert) (*12,13*). MYCIN is a pioneering rule-based expert system developed in the 1970s to help physicians determine the appropriate antibiotic therapy for patients with infections (*12*). PERFEX is a commercially available imaging expert system developed to assist physicians in the interpretation of myocardial perfusion SPECT studies (*13,14*).

RENEX consists of (a) a parameter normal library, (b) a knowledge base, and (c) an inference engine. To develop the parameter normal library, QuantEM 2.0 was designed to extract 7 patient parameters, 20 left kidney parameters, and 20 corresponding right kidney parameters from each ^{99m}Tc-MAG3 scan. Normal limits were established for the kidney parameters from ^{99m}Tc-MAG3 scans of 106 potential renal donors (*6*). From these data the domain expert estimated 5 boundary conditions for each parameter: (i) definitely abnormal, (ii) probably abnormal, (iii) equivocal, (iv) probably normal, and (v) definitely normal. A sigmoidlike fit constrained to these 5 boundary conditions was then performed, creating a parameter knowledge library to be used for converting the value of any individual quantitative parameter to a certainty factor regarding normality or abnormality.

To develop the knowledge base, 60 heuristic rules (IF A THEN B) were extracted from the domain expert to serve as the knowledge base for detecting obstruction. A forward chaining inference engine was developed using the MYCIN combinatories (an approximation of Bayes theorem) to determine the need for furosemide administration. If obstruction could be excluded by the baseline study, furosemide was not administered (*5*). When obstruction could not be excluded by the baseline study and furosemide was administered, RENEX again applied a forward chaining inference engine to generate a certainty factor with regard to the presence or absence of obstruction ranging from –1.0 (definitely not obstructed) to +1.0 (definitely obstructed). The certainty factors of the heuristic rules were adjusted and implemented using the pilot data (*3*) so that a certainty factor of –0.2 would be the threshold between not obstructed and equivocal and a certainty factor of +0.2 would be the threshold between equivocal and obstructed. Any certainty factor greater than 0.2, for example, would indicate obstruction; the higher the certainty factor, the greater the confidence that the kidney was obstructed. Similarly, a certainty factor less than –0.2 would indicate that the kidney is not obstructed. RENEX has not been designed to distinguish between probably obstructed and obstructed or to distinguish between probably nonobstructed and nonobstructed. The confidence in the diagnosis simply increases as the certainty factor becomes higher or lower. In practice, the clinical response to a diagnosis of probably obstructed versus obstructed would likely be equivalent. For experts, the categories "probably obstructed" and "probably nonobstructed" allowed a qualification of confidence in the diagnosis. On the basis of these considerations and for weighted κ-analysis, the expert interpretations of obstructed and probably obstructed were considered to be obstructed and the interpretations of probably nonobstructed and nonobstructed were considered to be nonobstructed. Processing time per patient was practically instantaneous using a 3.0-GHz personal computer programmed using IDL (Research Systems, Inc.).

Statistical Analysis

To assess the performance of RENEX as a diagnostic system, we first examined the degree of agreement between RENEX and expert consensus, which was quantified by weighted κ (*15*). The degree of agreement between 2 experts can be quantified by a simple unweighted κ, which has a range from 0 to 1, with larger values indicating better reliability (*16*). A limitation of unweighted κ is that all disagreements are treated equally. For example, the degree of disagreement for the case when one expert decides "obstructed" and the other expert decides "not obstructed" is the same as that of the case when one expert decides "obstructed" and the other decides "equivocal." A weighted κ avoids this problem by assigning different weights to disagreements according to the magnitude of the discrepancy (*15*). Because we analyzed 3 categories (obstructed, equivocal, and nonobstructed), we chose to use a weighted κ for our analysis. Disagreements are more likely to be by only 1 category than by 2 categories; consequently, the weighted κ will usually be a higher value than the unweighted κ. Landis and Koch have suggested that values of $\kappa < 0.00$ indicate no agreement, 0.00–0.19 indicate poor agreement, 0.20–0.39 indicate fair agreement, 0.40–0.59 indicate moderate agreement, 0.60–0.79 indicate substantial agreement, and 0.80–1.00 indicate almost perfect agreement (*17*).

As an alternative method to assess the performance of RENEX and to confirm the +0.2 and −0.2 thresholds, receiver-operating-characteristic (ROC) curve analysis was also conducted. The area under the ROC curve, which can take values between 0 and 1, may be used as a summary measure of the predictive accuracy of a diagnostic procedure. The use of ROC analysis, however, presupposes a gold standard. ROC curves were drawn using the consensus of the 3 experts as the gold standard. ROC analysis requires a gold standard that represents 2 mutually exclusive states; in our study, the 2 mutually exclusive states were "the kidney is obstructed" and "the kidney is not obstructed." Equivocals as a third category cannot be evaluated by ROC analysis; consequently, 2 separate ROC analyses were conducted: one that treated the equivocal case as obstructed and the other that treated the equivocal case as nonobstructed.

To determine how well RENEX performed relative to an individual expert, a series of analyses were subsequently conducted. First, the agreement between RENEX and each individual expert was assessed using weighted κ. For ROC analysis, comparing the predictive accuracy of RENEX with that of an expert was possible when the ratings of the other 2 experts were used as the gold standard. As before, 2 separate analyses were conducted: one that treated the equivocal case as obstructed and the other that treated the equivocal case as nonobstructed. To test the difference between the area under the RENEX ROC curve and that of an expert, a nonparametric procedure for correlated ROC data was used (*18*).

Finally, ROC analysis was also useful in confirming if the certainty factor of 0.2 would be an acceptable threshold for distinguishing between obstruction and equivocal and if a certainty factor of −0.2 would be an acceptable threshold for distinguishing between nonobstruction and equivocal. ROC curves were generated for certainty factors ranging from −1.0 to +1.0 and the results are plotted in Figures 3 and 4. ROC curves show the sensitivity and specificity of RENEX for various cutoff values. A perfect medical test would have 100% sensitivity and 100% specificity; on the ROC curve this corresponds to the point in the upper left-hand corner (0,1). In practice, however, diagnostic tests are imperfect, and the clinician has to strike a balance between sensitivity and specificity. When the cost of a false-negative result is the same as the cost of a false-positive result, the closer the ROC curve gets to the point (0,1), the better the test is at discriminating between cases and

noncases. This criterion was used to determine the acceptability of the decision threshold scores of +0.2 and −0.2 used by RENEX. All statistical analyses were performed using SAS statistical package, version 9.1 (SAS Institute).

RESULTS

The agreement between RENEX and the consensus readings for both kidneys is shown in Table 1. Eighty-four percent (101/120) of kidneys interpreted as nonobstructed by experts were interpreted as nonobstructed by RENEX; 14 (12%) were equivocal and only 5 (4%) were interpreted as obstructed. Of the 36 kidneys considered to be obstructed by the expert readers, 33 (92%) were considered to be obstructed by RENEX, 2 (6%) were equivocal, and 1 (2%) was interpreted as nonobstructed. The 1 kidney RENEX incorrectly interpreted as nonobstructed was missed because the patient moved in the middle of the acquisition; because of patient movement, the counts in the kidney ROI decreased and RENEX interpreted the quantitative data as nonobstructed, whereas the experts noted the motion on the images and were not misled by quantitative data suggesting normal drainage. An example of an obstructed and nonobstructed kidney is shown in Figures 1 and 2 and Supplemental Figures 1 and 2.

To further quantify the degree of agreement between RENEX and the consensus reading, weighted κ-statistics were calculated. As shown in Table 2, $\kappa = 0.72$, which suggested substantial agreement between RENEX and expert consensus. Using the consensus as the gold standard, the area under the ROC curve was also obtained to assess the diagnostic performance of RENEX. When the equivocal case was treated as positive, the predictive accuracy was 94.9% (Fig. 3); when the equivocal case was treated as negative, the predictive accuracy was 93.9% (Fig. 4).

The next question was to compare RENEX with each individual expert using κ-analysis. Pairwise weighted κ-values calculated for every pair of expert readers lay between 0.65 to 0.73, which indicated substantial agreement among the experts (Table 2). The weighted κ-values comparing RENEX and an individual expert were similar, ranging from 0.61 to 0.72 (Table 2).

To compare RENEX with each individual expert using ROC analysis, the other 2 experts were used as the gold standard. Table 3 shows that there was no significant difference in the performance of RENEX and any of the experts (*P* values ranged from 0.27 to 0.82).

The use of certainty factor +0.2 as the threshold is acceptable for separating the obstructed kidneys from the equivocal/nonobstructed kidneys (Fig. 3). At this point, the sensitivity and specificity were 91% and 92%, respectively. Likewise, −0.2 provided an acceptable threshold between nonobstructed and equivocal/obstructed kidneys (Fig. 4). At this point, the sensitivity and specificity were 86% and 89%, respectively. On the ROC curves, both thresholds lie close to the point (0,1).

TABLE 1
Agreement Between RENEX and Consensus Readings ($n = 185$)

	RENEX			
Consensus	Nonobstructed kidneys	Equivocal kidneys	Obstructed kidneys	Total
Nonobstructed kidneys	101	14	5	120
Equivocal kidneys	7	13	9	29
Obstructed kidneys	1	2	33	36

DISCUSSION

Decision support systems have the potential to serve as tools to assist physicians in interpreting studies at a faster rate, with a greater level of confidence and at a higher level of expertise. Over the past several years, artificial intelligence methods such as neural networks (*19–21*) and case-based reasoning (*22*) techniques have been investigated as a way to develop such tools. In the artificial neural net approach, the net tries to emulate how human neurons perform pattern recognition tasks. Repeated recognition trials are run using sample data as input and corresponding results as output to modify the strength between the input and output nodes. In this manner, the net is trained and the input data eventually predict the output. In the case-based reasoning approach, an algorithm searches a library of patient cases to find the ones that best match those of the patient study being analyzed. Another artificial intelligence approach that has been investigated to assist diagnosticians in making clinical interpretations is the knowledge-based expert system. In expert systems, a knowledge base of heuristic rules is obtained from human experts capturing how they make their interpretations. We choose to develop a knowledge-based system because the system can not only help provide a diagnosis but—unlike neural nets or case-based reasoning—also can be queried to provide the rules and input data used to justify the diagnosis (*4*). The high level of agreement between RENEX and the consensus results indicates that RENEX is performing similarly to each of the expert readers. In fact, the κ-scores indicate that RENEX agreed with the consensus results as well as the expert readers agreed with each other.

It could be argued that a better goal would be to develop decision support systems that use the clinical outcome as the gold standard rather than expert readers. This is an attractive goal but it misses the point of an expert system, which is to interpret studies with the same level of expertise as experts. It is generally accepted that experts, as defined in this article, interpret studies in their specialty better than general nuclear medicine physicians or radiologists; this is the basis for having distinct areas of expertise within academic departments and even within private practice settings. Regardless of the type of study, clinical practice should be improved if

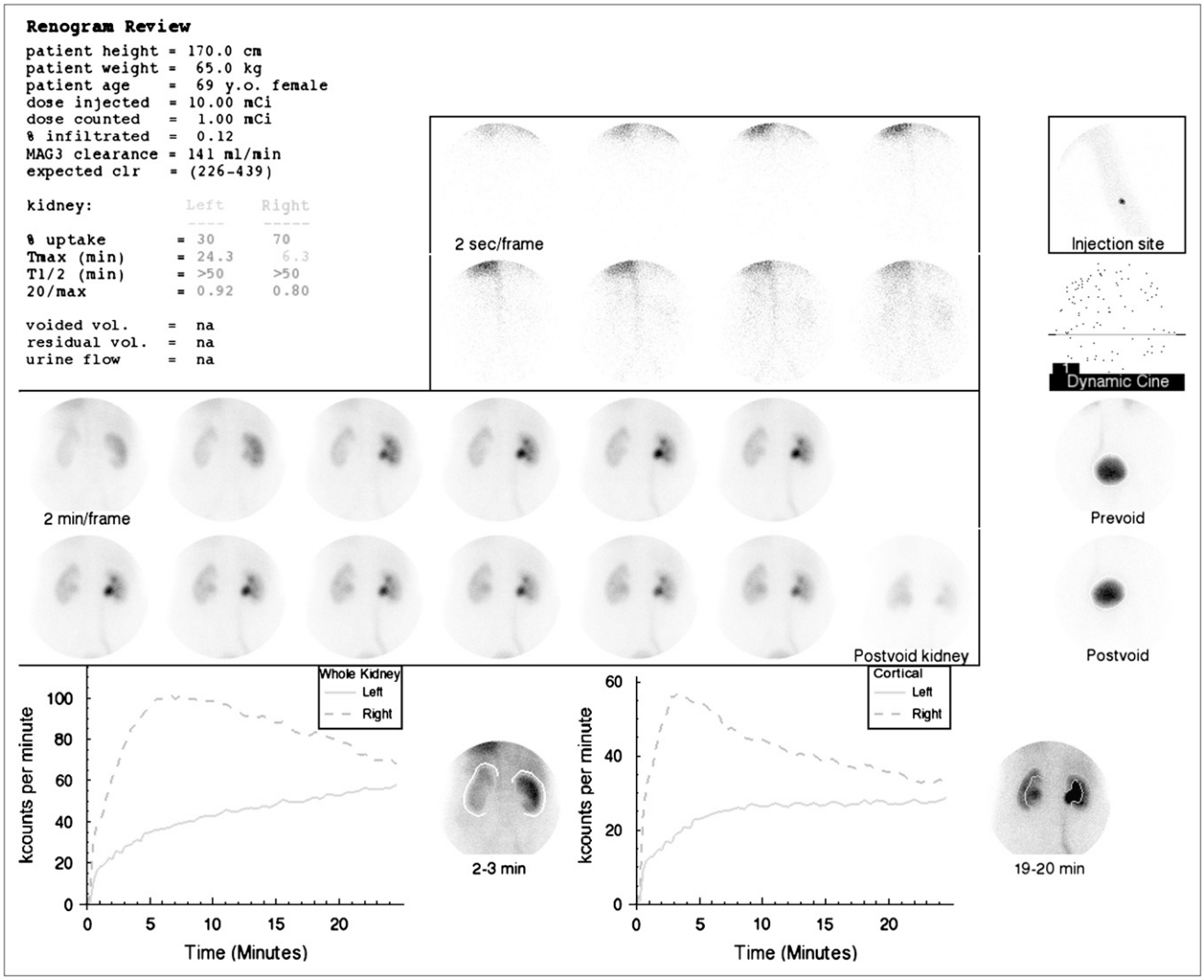

FIGURE 1. Standard display shows demographic data, dose injected, dose counted on camera, percent dose infiltrated, ^{99m}Tc-MAG3 clearance and expected ^{99m}Tc-MAG3 clearance followed by percent uptake, Tmax, T1/2, and 20 min/max ratios for whole-kidney ROI. Voided volume, postvoid residual, and urine flow rate were not measured. (Upper central panel) Two-second images at beginning of acquisition. (Upper right panel) Injection site; just beneath is a frame for viewing dynamic cine and pre- and postvoid bladder images. (Center panel) Twelve 2-min images followed by postvoid image of kidneys with patient lying on camera in same position as that for initial images. (Lower left panel) Whole-kidney ROIs and whole-kidney renogram curves. (Lower right panel) Cortical ROIs and cortical renogram curves. ^{99m}Tc-MAG3 clearance was reduced (141 mL/min/1.73m^2 compared with normal range of 226–439 mL/min/1.73 m^2). Relative uptake of left kidney was 30%. T1/2 of both kidneys was >50 min and 20 min/max ratio was bilaterally abnormal; consequently, patient received furosemide followed by a second acquisition (Fig. 2 and Supplemental Fig. 2).

imagers can provide an interpretation equivalent to that of expert readers. Outcome is certainly an important measure and it is one procedure by which experts develop expertise over time; outcome can also be useful for adjudicating disagreements between experts. However, in regard to diuresis renography, outcome as a gold standard is confounded by the fact the scan interpretation (obstruction vs. no obstruction) has a major impact on the clinical outcome (surgical intervention vs. observation); consequently, this gold standard is biased. An additional problem can be illustrated by a patient who had a pyeloplasty to relieve obstruction 1 y after a diuresis renography scan was interpreted as "no obstruction." In this illustration, did the scan miss obstruction, was the study interpreted incorrectly, did the patient become obstructed only 1 y after the scan, or did an aggressive surgeon operate on a nonobstructed kidney? Using patient outcome as a gold standard can be an important goal but interpretation of the results is not straightforward and it is not the goal of an expert system.

There are several limitations to the study. Our study addressed the diuresis renography protocol recommended by the international consensus report, in which baseline data are obtained and followed by the administration of furosemide and an additional period of imaging (*1*). There

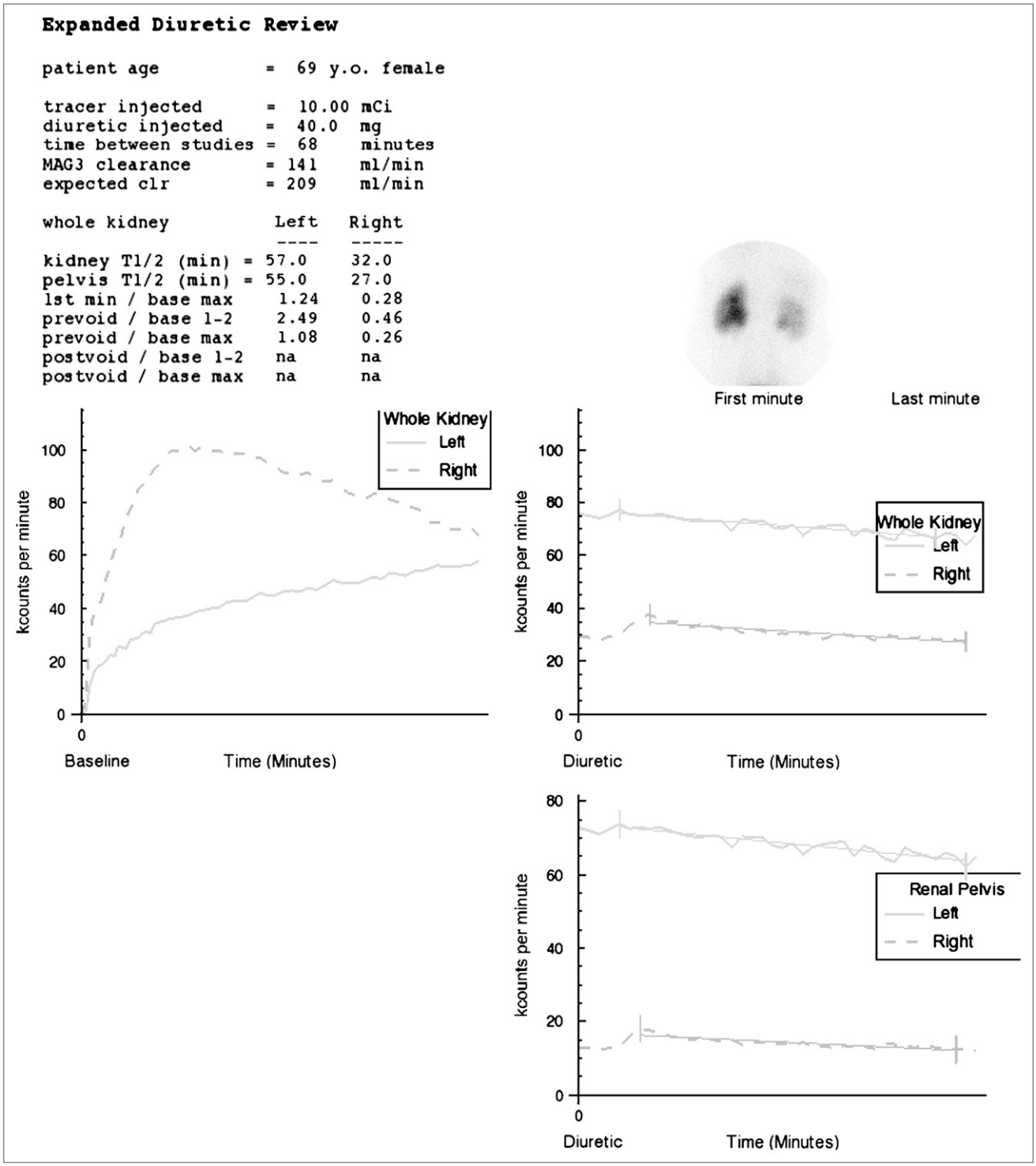

FIGURE 2. Display of baseline and furosemide whole-kidney renogram curves on same scale. Time–activity curve generated by pelvic ROI is also displayed on an expanded scale. Patient's global renal function is reduced and there is diffuse retention in right kidney compatible with reduced function; however, tracer largely washed out of the right renal pelvis, and the ratio of prevoid furosemide counts to maximal counts on baseline study was only 0.26, indicating that about 75% of maximum activity had washed out of right kidney. Experts interpreted right kidney as ''probably not obstructed.'' RENEX also interpreted right kidney as nonobstructed (certainty factor of −0.37). Experts interpreted left kidney as ''probably obstructed.'' Absolute function and relative function were reduced in left kidney; there was gradual increase in tracer activity in the left renal pelvis and minimal washout after furosemide administrations. RENEX also interpreted left kidney as obstructed (certainty factor of 0.76) (Supplemental Fig. 2).

TABLE 2
Agreement Between RENEX and Experts (n = 185)

Experts	Weighted κ	SE	95% Confidence interval
1 vs. 2	0.73	0.04	(0.64, 0.81)
1 vs. 3	0.65	0.05	(0.55, 0.75)
1 vs. RENEX	0.71	0.05	(0.62, 0.80)
2 vs. 3	0.73	0.04	(0.64, 0.82)
2 vs. RENEX	0.72	0.04	(0.63, 0.80)
3 vs. RENEX	0.61	0.04	(0.64, 0.81)
Consensus vs. RENEX	0.72	0.04	(0.64, 0.81)

are other protocols in which furosemide is given 15 min before the radiopharmaceutical, at the same time as the radiopharmaceutical, or 5–10 min later (*1,23–25*). At present, the system we describe does not apply to these protocols, although it could be adapted to evaluate data from other acquisition protocols. Another limitation is the fact that all of the patients in the test group (*3*) and in the current study were adults; the system has not been tested in a pediatric population nor has it been tested in patients with renal transplants. The ideal standard for ROC analysis is a determination of truth independent of the imaging modality to be tested. We used an expert panel as the gold standard but this gold standard is limited by the fact that the expert panel used the same data to reach its conclusion as was available to RENEX. Although ROC analysis can be used when the image analysis and the gold standard (expert or verification panel) are based on the same data (*26*), an independent determination of the presence or absence of obstruction would be a preferable standard. Unfortunately, as just discussed, an independent determination of obstruction or nonobstruction is very problematic to achieve clinically. A second problem with ROC analysis is that ROC analysis requires truth to be dichotomized but, in the clinical setting of renal scan interpretation, truth is not dichotomized; experts classify scans as obstructed, not obstructed, and indeterminate. It is possible to count indeterminate studies as "obstructed" or "nonobstructed" but this dichotomized approach does not mirror the clinical situation. In fact, to construct the ROC curves, equivocal studies had to be counted as obstructed or nonobstructed and we analyzed the data both with equivocal studies counted as obstructed and again with equivocal studies counted as nonobstructed. ROC analysis was helpful in confirming that the certainty factor values of +0.2 and −0.2 were appropriate thresholds to separate equivocal and obstructed (+0.2) and equivocal and nonobstructed (−0.2) but ROC analysis is less suitable for comparing the accuracy of RENEX with the experts. For this comparison and for comparison between experts, κ-analysis is probably superior.

The fact that the areas under the ROC curves are essentially the same (Table 3) shows that RENEX and experts performed similarly when equivocal was lumped with obstruction and or when equivocal was lumped with nonobstruction. The primary value of decision support systems, however, is not in the ability to distinguish between obviously nonobstructed and obstructed kidneys but to help less-

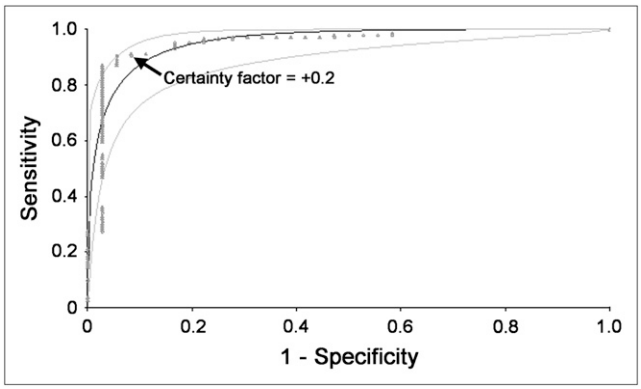

FIGURE 3. Clinical data were used to confirm that +0.2 was an acceptable certainty factor for RENEX to separate obstructed kidneys from the combined group of equivocal and nonobstructed kidneys. Kidneys considered to be equivocal or obstructed by consensus analysis were combined, and ROC curves were constructed to compare performance of RENEX with consensus interpretation with regard to distinguishing between obstructed kidneys and the combined group of nonobstructed and equivocal kidneys. ROC analysis was performed for certainty factors ranging from −1.0 to +1.0. Plot of this analysis confirms the certainty factor of +0.2 to be an acceptable threshold for separating obstructed kidneys from the combined group. Fitted ROC curve and its 95% confidence bands are shown as smooth curves. Empiric ROC curve is shown in dots.

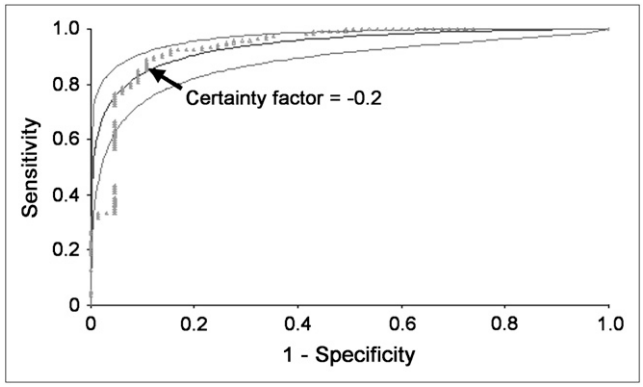

FIGURE 4. Clinical data were used to confirm that −0.2 was an acceptable certainty factor for RENEX to separate nonobstructed kidneys from the combined group of equivocal and obstructed kidneys. Kidneys considered to be equivocal or nonobstructed by consensus analysis were combined, and ROC curves were constructed to compare performance of RENEX with consensus interpretation with regard to distinguishing between nonobstructed kidneys and the combined group of obstructed and equivocal kidneys. ROC analysis was performed for certainty factors ranging from −1.0 to +1.0. Plot of this analysis confirms the certainty factor of −0.2 to be an acceptable threshold for separating nonobstructed kidneys from the combined group. Fitted ROC curve and its 95% confidence bands are shown as smooth curves. Empiric ROC curve is shown in dots.

TABLE 3
Area Under ROC Curves Comparing Each Expert with RENEX Using Consensus Reading of Other 2 Experts as Gold Standard (n = 185)

Gold standard	Expert vs. RENEX	Area under ROC curve			
		Method 1*	*P* value	Method 2†	*P* value
Expert 1 and expert 2	Expert 3	0.93	0.57	0.93	0.27
	RENEX	0.94		0.94	
Expert 1 and expert 3	Expert 2	0.93	0.71	0.93	0.42
	RENEX	0.94		0.95	
Expert 2 and expert 3	Expert 1	0.92	0.82	0.92	0.55
	RENEX	0.94		0.95	

*Method 1 gives areas under curve obtained when equivocal case was treated as obstructed.
†Method 2 gives areas under curve when equivocal case was treated as not obstructed.

experienced readers—particularly, as studies become more difficult to interpret. It is important to note that our study was biased toward more difficult interpretative cases as patients whose kidneys were obviously not obstructed on the baseline scan—about one third of patients in our experience (5)—did not receive furosemide, and patients with clearly nonobstructed kidneys were not included in the study population. Nevertheless, the majority of the disagreements between RENEX and experts were between obstruction and equivocal and between nonobstruction and equivocal, not between obstruction and nonobstruction. In fact, RENEX disagreed with the expert consensus in 55% (16/29) of the equivocal studies. These borderline cases are the most problematic for RENEX and experts, and we have preliminary data suggesting that clinical data can reduce the number of equivocal studies by 60%–70%. Although our results show that RENEX agrees with the experts and the experts agree with each other, these results do not mirror the clinical situation because the experts were unaware of clinical data so that their interpretations could be appropriately compared with the interpretations provided by RENEX, which has not yet been developed to acquire and analyze clinical data. Future research efforts will include the development of algorithms to incorporate clinical data into RENEX. Finally, the current acquisition and processing program, QuantEM 2.0, cannot detect and correct for patient motion and cannot distinguish between diffuse retention with slow washout due to impaired function and focal pelvic retention with slow washout due to a likely obstruction. Algorithms to (a) incorporate clinical data, (b) detect and correct for motion, and (c) distinguish between diffuse retention in a kidney and retention in a dilated renal collecting system will better mirror the actual clinical situation; these algorithms must be designed, implemented, and tested.

CONCLUSION

RENEX is a knowledge-based decision support system designed to interpret diuresis renography studies acquired in the baseline plus furosemide protocol. The interpretations provided by RENEX showed substantial agreement with expert readers. In fact, RENEX agreed with the consensus reading as well as the experts agreed with each other, although both RENEX and the expert readers were unaware of clinical information. Algorithms to incorporate clinical information, to detect and correct for motion, and to distinguish between diffuse retention in a kidney and retention in a dilated renal collecting system should improve the performance of RENEX and increase the level of confidence in the diagnosis.

ACKNOWLEDGMENTS

This work was funded by the National Library of Medicine grant number R01-LM007595. Some of the authors (AT, EVG, RH, RDF) receive royalties from the sale of the application software QuantEM related to the research described in this article. The terms of this arrangement have been reviewed and approved by Emory University in accordance with its conflict-of-interest practice.

REFERENCES

1. O'Reilly P, Aurell M, Britton K, Kletter K, Rosenthal L, Testa T. Consensus on diuresis renography for investigating the dilated upper urinary tract. *J Nucl Med.* 1996;37:1872–1876.
2. IMV Medical Information Division. *2003 Nuclear Medicine Census Marke Summary Report.* Des Plaines, IL: IMV, Ltd.; 2003:IV(7–11).
3. Garcia EV, Taylor A, Halkar R, et al. RENEX: an expert system for the interpretation of ^{99m}Tc-MAG3 scans to detect renal obstruction. *J Nucl Med.* 2006;47:320–329.
4. Garcia EV, Taylor A, Manatunga D, Folks R. A software engine to justify the conclusions of an expert system for detecting renal obstruction on ^{99m}Tc-MAG3 scans. *J Nucl Med.* 2007;48:463–470.
5. Taylor A, Hill A, Binongo J, et al. Evaluation of two diuresis renography decision support systems to determine the need for furosemide in patients with suspected obstruction. *AJR.* 2007;188:1395–1402.
6. Esteves FP, Taylor A, Manatunga A, Folks RD, Krishnan M, Garcia EV. ^{99m}Tc-MAG3 renography: normal values for MAG3 clearance and curve parameters, excretory parameters and residual urine volume. *AJR.* 2006;187:W610–W617.
7. Taylor A Jr, Corrigan PL, Galt J, et al. Measuring technetium-99m-MAG3 clearance with an improved camera-based method. *J Nucl Med.* 1995;36:1689–1695.
8. Taylor A, Manatunga A, Morton K, et al. Multicenter trial validation of a camera-based method to measure Tc-99m mercaptoacetyltriglycine, or Tc-99m MAG3, clearance. *Radiology.* 1997;204:47–54.
9. Taylor A, Lewis C, Giacometti A, et al. Improved formulas for the estimation of renal depth in adults. *J Nucl Med.* 1993;34:1766–1769.

10. Taylor A. Formulas to estimate renal depth in adults. *J Nucl Med.* 1994;35:2054–2055.
11. Folks RD, Garcia EV, Taylor AT. Prospective evaluation of an automated software system for quality control of quantitative ^{99m}Tc-MAG3 renal studies. *J Nucl Med Technol.* 2007;35:27–33.
12. Shortliffe EH. *Computer-Based Medical Consultations: MYCIN.* Amsterdam, The Netherlands: Elsevier Scientific Publishing Co.: 1976:264.
13. Ezquerra N, Mullick R, Cooke D, Krawczynska E, Garcia E. PERFEX: an expert system for interpreting 3D myocardial perfusion. *Expert Syst Appl.* 1993;6:459–468.
14. Garcia EV, Cooke CD, Folks RD, et al. Diagnostic performance of an expert system for the interpretation of myocardial perfusion SPECT studies. *J Nucl Med.* 2001;42:1185–1191.
15. Fleiss JL. *Statistical Methods for Rates and Proportions.* 2nd ed. New York, NY: Wiley.
16. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas.* 1960;20:37–46.
17. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159–174.
18. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics.* 1988;44:837–845.
19. Fujita H, Katafuchi T, Uehara T, Nishimura T. Application of neural network to computer-aided diagnosis of coronary artery disease in myocardial SPECT bull's-eye images. *J Nucl Med.* 1992;33:272–276.
20. Porenta G, Dorffner G, Kundrat S, Petta P, Duit-Schedlmayer J, Sochor H. Automated interpretation of planar thallium-201-dipyridamole stress-redistribution scintigrams using artificial neural networks. *J Nucl Med.* 1994;35:2041–2047.
21. Hamilton D, Riley PJ, Miola UJ, Amro AA. A feed forward neural network for classification of bull's-eye myocardial perfusion images. *Eur J Nucl Med.* 1995;22:108–115.
22. Haddad M, Adlassnig KP, Porenta G. Feasibility analysis of a case-based reasoning system for automated detection of coronary heart disease from myocardial scintigrams. *Artif Intell Med.* 1997;9:61–78.
23. Sfakianakis GN, Cohen DJ, Braunstein RH, et al. MAG3-F0 scintigraphy in decision making for emergency intervention in renal colic after helical CT positive for a urolith. *J Nucl Med.* 2000;41:1813–1822.
24. Wong DC, Rossleight MA, Farnsworth RH. F + 0 diuresis renography in infants and children. *J Nucl Med.* 1999;40:1805–1811.
25. Sundaram PS, Padma S, Bhat S, Sanjeevan KV, Rahul C. F + 10 diuretic renography protocol is better than F − 15 in followup of post pyeloplasty patients [abstract]. *J Nucl Med.* 2003;44(suppl):359P.
26. Kundel HL, Polansky M. Measurement of observer agreement. *Radiology.* 2003;228:303–308.