Health Technology Assessment

Appendices
Coronary Artery Calcium Scoring (CACS) as a Diagnostic Test for Detection of Coronary Artery Disease

Date: Monday, August 10, 2009

Health Technology Assessment Program
676 Woodland Square Loop SE
P.O. Box 42712
Olympia, WA 98504-2712
http://www.hta.hca.wa.gov

Coronary Artery Calcium Scoring (CACS) as a Diagnostic Test for Detection of Coronary Artery Disease

Provided by:

Spectrum Research, Inc.

Prepared by:

Andrea C. Skelly, PhD, MPH
ADDITIONAL AUTHORS TO BE INSERTED

This technology assessment report is based on research conducted by a contracted technology assessment center, with updates as contracted by the Washington State Health Care Authority. This report is an independent assessment of the technology question(s) described based on accepted methodological principles. The findings and conclusions contained herein are those of the investigators and authors who are responsible for the content. These findings and conclusions may not necessarily represent the views of the HCA/Agency and thus, no statement in this report shall be construed as an official position or policy of the HCA/Agency.

The information in this assessment is intended to assist health care decision makers, clinicians, patients and policy makers in making sound evidence-based decisions that may improve the quality and cost-effectiveness of health care services. Information in this report is not a substitute for sound clinical judgment. Those making decisions regarding the provision of health care services should consider this report in a manner similar to any other medical reference, integrating the information with all other pertinent information to make decisions within the context of individual patient circumstances and resource availability.

TABLE OF CONTENTS--Appendices

Appendix A: Search Strategies

Medline search strategies

Searches conducted through May, 2009 and additional safety search July-09

#	Search terms	# citations
1	"Coronary Artery Disease"[Mesh] OR "Coronary Disease"[Mesh]	151402
2	("Coronary Angiography"[Mesh] OR "Tomography, X-Ray Computed"[Mesh]) OR "Tomography, Spiral Computed"[Mesh] OR (64-slice OR EBCT OR EBT OR ultrafast OR electron beam OR electron beam tomography)	258130
3	("Sensitivity and Specificity"[Mesh] OR "Predictive Value of Tests"[Mesh]) OR "Prospective Studies"[Mesh]	508657
4	coronary artery calcium OR coronary calcium	47793
5	#1 AND #2 AND #3 AND #4 Limits: only items with abstracts, Humans, English	351
6	"Decision Making"[Mesh]	82352
7	#1 AND #2 AND #6	75
8	"Incidental Findings" [Mesh]	2218
9	#1 AND #2 AND #8	37
10	"Safety"[Mesh] OR "Equipment Safety"[Mesh]	43787
11	#1 AND #2 AND #10 Limits: only items with abstracts, Humans, English	139
12	#1 AND #2 AND #10 NOT (stent) Limits: only items with abstracts, Humans, English	66
13	"Reproducibility of Results"[Mesh] OR "Validation Studies "[Publication Type]	201398
14	coronary artery calcium OR coronary calcium Limits: only items with abstracts, Humans, English	14103
15	#1 AND #2 AND #13 AND #14	109
16	("Radiation, Ionizing"[Mesh] OR "Radiation, Ionizing/adverse effects"[Mesh])	37646
18	#1 AND #2 AND #4 AND #16	85
19	("Coronary Angiography/economics"[Mesh] OR "Tomography, X-Ray Computed/economics"[Mesh]) OR "Tomography, Spiral Computed/economics"[Mesh]	1283
20	#1 AND #4 AND #19 Limits: only items with abstracts, Humans, English	11
21	(calcium score) AND systematic[sb] Limits: only items with abstracts, Humans, English	38

Embase search 6.25.2009:

	Terms	Results
1	coronary artery disease.mp	(65508)
2	coronary stenosis.	(2126)
3	1 or 2	(66828)
4	Angiography	(94034)
5	computed tomography	(74049)
6	4 or 5	(158881)
7	(calcium or calcification)	(320543)
8	(cost or cost-effective)	(226566)
9	8 and 6 and 3 and 7	(82)
10	limit 9 to (abstracts and english language)	(65)
11	6 and 3 and 7 and 10	1527
12	(64-slice or EBCT or EBT or ultrafast or electron beam or electron beam tomography).	8410

13	11 and 12	393
14	(coronary artery calcium or coronary calcium)	985
15	13 and 14	197

Web of science search 6.26.2009:

Topic=(Coronary Angiography) AND Topic=(computed tomography) AND Topic=(64-slice OR EBCT OR EBT OR ultrafast OR electron beam OR electron beam tomography) AND Topic=(coronary artery calcium OR coronary calcium) AND Topic=(coronary artery disease)

Refined by: Languages=(ENGLISH) AND Subject Areas=(CARDIAC & CARDIOVASCULAR SYSTEMS OR RADIOLOGY, NUCLEAR MEDICINE & MEDICAL IMAGING) [Excluded Hematology and General internal Medicine to get it under 200 citations]

Timespan=All Years. Databases=SCI-EXPANDED, SSCI, A&HCI.

Economic Studies lit search

Embase search 6.25.2009:

	Terms	**Results**
1	coronary artery disease.mp	(65508)
2	coronary stenosis.	(2126)
3	1 or 2	(66828)
4	Angiography	(94034)
5	computed tomography	(74049)
6	4 or 5	(158881)
7	(calcium or calcification)	(320543)
8	(cost or cost-effective)	(226566)
9	8 and 6 and 3 and 7	(82)
10	limit 9 to (abstracts and english language)	(65)
11	**from 10 keep 1-65**	**(65)**

Web of science search 6.25.2009:

Databases=SCI-EXPANDED, SSCI, A&HCI Timespan=All Years

	Terms	Results
# 1	TS=(coronary artery disease or coronary stenosis) and TS=(angiography or computed tomography) and TS=(cost or economic)	469
# 2	#1 and TS=(calcium or calcification)	59
# 3	#2 and TS=(diagnostic or diagnosis)	28

EconLit search 6.25.2009

Search: coronary artery disease AND cost=10 articles (none include calcium/calcification)

Grey Literature Searches

AHRQ Search 5-20-09

1	coronary AND CT OR EBCT OR MDCT	310
2	coronary AND computed AND tomography	89
3	coronary AND calcium	173
4	coronary AND calcium AND computed AND tomography	32

5	coronary AND calcium AND computed AND tomography AND angiography	20
6	coronary AND calcium AND scoring	36
7	coronary AND calcium AND scoring AND computed AND tomography	7
8	coronary AND "calcium scoring" AND computed AND tomography	2

Literature Found
1 systematic evidence review on screening for asymptomatic CAD
1 recommendation statement from U.S Preventive Services Task Force on screening

NGC Search 5-20-09

1	coronary AND CT OR EBCT OR MDCT	127
2	coronary AND computed AND tomography	63
3	coronary AND calcium	115
4	coronary AND calcium AND computed AND tomography	34
5	coronary AND calcium AND computed AND tomography AND angiography	20
6	coronary AND calcium AND scoring	10
7	coronary AND calcium AND scoring AND computed AND tomography	0
8	coronary AND "calcium scoring" AND computed AND tomography	0

FDA Search 5-20-09

1	coronary AND CT OR EBCT OR MDCT	3600
2	coronary AND computed AND tomography	921
3	coronary AND calcium	4230
4	coronary AND calcium AND computed AND tomography	455
5	coronary AND calcium AND computed AND tomography AND angiography	179
6	coronary AND calcium AND scoring	508
7	coronary AND calcium AND scoring AND computed AND tomography	49
8	coronary AND "calcium scoring" AND computed AND tomography	19

INAHTA Search 5-20-09

1	coronary AND CT OR EBCT OR MDCT	8
2	coronary AND computed AND tomography	27
3	coronary AND calcium	45
4	coronary AND calcium AND computed AND tomography	28
5	coronary AND calcium AND computed AND tomography AND angiography	28
6	coronary AND calcium AND scoring	45
7	coronary AND calcium AND scoring AND computed AND tomography	28
8	coronary AND "calcium scoring" AND computed AND tomography	27

Mowat: The Effectiveness and Cost Effectiveness of Computed Tomography Screening for Coronary Artery Disease: Systematic Review Health Technol Assess 2006; 10(39). October 2006.
1 possible economic study (Dewey) in symptomatic persons

All published studies found in Cochrane and Clinical trials searches were contained in on other searches or were in asymptomatic persons
Cochrane Library Search: 5-18-09

There are **53** results out of **575975 records** for: "**Coronary calcium scoring** in **Title, Abstract or Keywords** in **Cochrane Central Register of Controlled Trials**"

There are **14** results out of **9964 records** for: "**tomography, x-ray computed** in **Title, Abstract or Keywords** and **coronary artery disease** in **Title, Abstract or Keywords** in **Database of Abstracts of Reviews of Effects**"

There are **41** results out of **575975 records** for: "**tomography, x-ray computed** and **coronary disease** in **Title, Abstract or Keywords** and **predictive value** in **Title, Abstract or Keywords** in **Cochrane Central Register of Controlled Trials**"

Clinical trials.gov search: 5-21-09

Found 34 studies with search of coronary artery AND calcium AND score

Appendix B
Exclusion of studies at the full text level of review

Screening studies or those with >80% asymptomatic patients (or data for symptomatic patients not separated out)

Halliburton SS, Stillman AE, Lieber M, Kasper JM, Kuzmiak SA, White RD. Potential clinical impact of variability in the measurement of coronary artery calcification with sequential MDCT. AJR Am J Roentgenol. Feb 2005;184(2):643-648

Rozanski A, Gransar H, Wong ND, et al. Clinical outcomes after both coronary calcium scanning and exercise myocardial perfusion scintigraphy. J Am Coll Cardiol. Mar 27 2007;49(12):1352-1361.

Mitsutake, R., H. Niimura, et al. (2006). "Clinical significance of the coronary calcification score by multidetector row computed tomography for the evaluation of coronary stenosis in Japanese patients." Circulation Journal **70**(9): 1122-1127.

Qu, W., T. T. Le, et al. (2003). "Value of coronary artery calcium scanning by computed tomography for predicting coronary heart disease in diabetic subjects." Diabetes Care 26(3): 905-10.

Agatston AS, Janowitz WR, Hildner FJ, Zusmer NR, Viamonte M, Jr., Detrano R. Quantification of coronary artery calcium using ultrafast computed tomography. J Am Coll Cardiol. Mar 15 1990;15(4):827-832.

Almeda, F. Q., R. Shah, et al. (2004). "Clinical and angiographic profile of patients with markedly elevated coronary calcium scores (>or=1000) detected by electron beam computed tomography." Cardiovasc Radiat Med **5**(3): 109-12.

Reliability studies in asymptomatic persons

Bielak, L. F., R. B. Kaufmann, et al. (1994). "Small lesions in the heart identified at electron beam CT: calcification or noise?" Radiology **192**(3): 631-6.

Callister, T. Q., B. Cooil, et al. (1998). "Coronary artery disease: improved reproducibility of calcium scoring with an electron-beam CT volumetric method." Radiology **208**(3): 807-14.

Hong C, Bae KT, Pilgram TK. Coronary artery calcium: accuracy and reproducibility of measurements with multi-detector row CT--assessment of effects of different thresholds and quantification methods. Radiology. Jun 2003;227(3):795-801.

Horiguchi, J., N. Matsuura, et al. (2008). "Variability of repeated coronary artery calcium measurements by 1.25-mm- and 2.5-mm-thickness images on prospective electrocardiograph-triggered 64-slice CT." Eur Radiol **18**(2): 209-16.

Lu, B., M. J. Budoff, et al. (2002). "Causes of interscan variability of coronary artery calcium measurements at electron-beam CT." Acad Radiol **9**(6): 654-61.

Mao SS, Pal RS, McKay CR, et al. Comparison of coronary artery calcium scores between electron beam computed tomography and 64-multidetector computed tomographic scanner. J Comput Assist Tomogr. Mar-Apr 2009;33(2):175-178.

Moser KW, Bateman TM, O'Keefe JH, Jr., McGhie AI. Interscan variability of coronary artery calcium quantification using an electrocardiographically pulsed spiral computed tomographic protocol. Am J Cardiol. May 1 2004;93(9):1153-1155.

Yamamoto H, Budoff MJ, Lu B, Takasu J, Oudiz RJ, Mao S. Reproducibility of three different scoring systems for measurement of coronary calcium. Int J Cardiovasc Imaging. Oct 2002;18(5):391-397.

Wang, S., R. C. Detrano, et al. (1996). "Detection of coronary calcification with electron-beam computed tomography: evaluation of interexamination reproducibility and comparison of three image-acquisition protocols." Am Heart J **132**(3): 550-8.

Studies with >80% patient with previous revascularization (CABG, PTCA, stent) or previous history of MI

Cordeiro MA, Miller JM, Schmidt A, et al. Non-invasive half millimetre 32 detector row computed tomography angiography accurately excludes significant stenoses in patients with advanced coronary artery disease and high calcium scores. Heart. May 2006;92(5):589-597

Mohlenkamp S, Lehmann N, Schmermund A, et al. Prognostic value of extensive coronary calcium quantities in symptomatic males--a 5-year follow-up study. Eur Heart J. May 2003;24(9):845-854.

Shemesh, J., S. Apter, et al. (1995). "Calcification of coronary arteries: detection and quantification with double-helix CT." Radiology **197**(3): 779-83.

Studies analyzing per vessel or per segment only and/or data not extractable

Chen LC, Chen JW, Wu MH, et al. Differential coronary artery calcification detected by electron beam computed tomography as an indicator of coronary stenosis among patients with stable angina pectoris. Can J Cardiol. Jun 2001;17(6):667-676.

Kitamura A, Kobayashi T, Ueda K, et al. Evaluation of coronary artery calcification by multi-detector row computed tomography for the detection of coronary artery stenosis in Japanese patients. J Epidemiol. Sep 2005;15(5):187-193.

Kajinami K, Seki H, Takekoshi N, Mabuchi H. Coronary calcification and coronary atherosclerosis: site by site comparative morphologic study of electron beam computed tomography and coronary angiography. J Am Coll Cardiol. Jun 1997;29(7):1549-1556.

Schmermund A, Bailey KR, Rumberger JA, Reed JE, Sheedy PF, 2nd, Schwartz RS. An algorithm for noninvasive identification of angiographic three-vessel and/or left main coronary artery disease in symptomatic patients on the basis of cardiac risk and electron-beam computed tomographic calcium scores. J Am Coll Cardiol. Feb 1999;33(2):444-452

Reference and test not performed within 3 months of each other

Yaghoubi S, Tang W, Wang S, et al. Offline assessment of atherosclerotic coronary calcium from electron beam tomograms. Am J Card Imaging. Oct 1995;9(4):231-236

Brown BG, Morse J, Zhao XQ, Cheung M, Marino E, Albers JJ. Electron-beam tomography coronary calcium scores are superior to Framingham risk variables for predicting the measured proximal stenosis burden. Am J Cardiol. Jul 19 2001;88(2A):23E-26E.

Studies of CT coronary angiography not using CACS as the diagnostic test

Husmann L, Schepis T, Scheffel H, et al. Comparison of diagnostic accuracy of 64-slice computed tomography coronary angiography in patients with low, intermediate, and high cardiovascular risk. Acad Radiol. Apr 2008;15(4):452-461.

White, C. S., D. Kuo, et al. (2005). "Chest pain evaluation in the emergency department: can MDCT provide a comprehensive evaluation?" AJR Am J Roentgenol **185**(2): 533-40.

Studies of technique

Qanadli SD, Mesurolle B, Aegerter P, et al. Volumetric quantification of coronary artery calcifications using dual-slice spiral CT scanner: improved reproducibility of measurements with 180 degrees linear interpolation algorithm. J Comput Assist Tomogr. Mar-Apr 2001;25(2):278-286

Shemesh, J., A. Tenenbaum, et al. (1997). "Coronary calcium measurements by double helical computed tomography. Using the average instead of peak density algorithm improves reproducibility." Invest Radiol **32**(9): 503-6.

Reports with overlap in study populations for the same outcomes

Breen, J. F., P. F. Sheedy, 2nd, et al. (1992). "Coronary artery calcification detected with ultrafast CT as an indication of coronary artery disease." Radiology **185**(2): 435-9.

Detrano, R., T. Hsiai, et al. (1996). "Prognostic value of coronary calcification and angiographic stenoses in patients undergoing coronary angiography." J Am Coll Cardiol **27**(2): 285-90.

Kaufmann RB, Peyser PA, Sheedy PF, Rumberger JA, Schwartz RS. Quantification of coronary artery calcium by electron beam computed tomography for determination of severity of angiographic coronary artery disease in younger patients. J Am Coll Cardiol. Mar 1 1995;25(3):626-632

Appendix C. Level of Evidence Determination

Introduction:
Studies which evaluate the accuracy and reliability of diagnostic tests are subject to a number of biases which may provide inaccurate assessment of its characteristics and clinical utility.[1, 2] Parameters related to diagnostic accuracy (validity) and reliability are described in Appendix H.

Methods for critical appraisal and level of evidence assessment
Spectrum Research's (SRI) methods for assessing the quality of evidence of individual studies as well as the overall quality of evidence incorporates aspects of rating scheme developed by the Oxford Centre for Evidence-based Medicine,[3] precepts outlined by the Grades of Recommendation Assessment, Development and Evaluation (GRADE) Working Group[4] and recommendations made by the Agency for Healthcare Research and Quality (AHRQ).[5] We believe that taking into account features of methodological quality and sources of bias that are important and our LoE method combines epidemiologic principles with characteristics of study design.

Our method incorporates the essential five domains and related elements delineated by AHRQ,[5]as described in the following table, in addition to considering whether the study was prospectively or retrospectively designed.

Table C1. Overview Spectrum Research's LoE Assessment Based on AHRQ Domains

AHRQ Domain	Spectrum Research LoE Assessment
1. Study Population	• Was a broad spectrum of persons with the expected condition was used?
2. Description of Test	• Are the technical features, measurements performed, planes of section, diagnostic criteria, etc. described for both the test and the reference standard with sufficient detail to permit replication?
3. Appropriate Reference Standard	• Based on the pathology/condition being evaluated, is the test compared with the current "best" standard that is likely to correctly classify patients according to disease status?
4. Blinded Comparison of Test and Reference Standard	• Interpretation of test reference standard must be done without knowledge of the results of other?
5. Avoidance of Verification	• Reference standard must be performed

Bias	independently of test?

Reproducibility studies are those that evaluate the extent to which measurements can be replicated on subject/patient. Grading the quality of evidence for reliability studies has not been well reported in the literature. SRI's method is based on epidemiologic methods for validation (degree to which measurements reflect the truth) and reliability (reproducibility) studies.[6] This system takes into consideration pertinent study design features and methods that may induce bias.

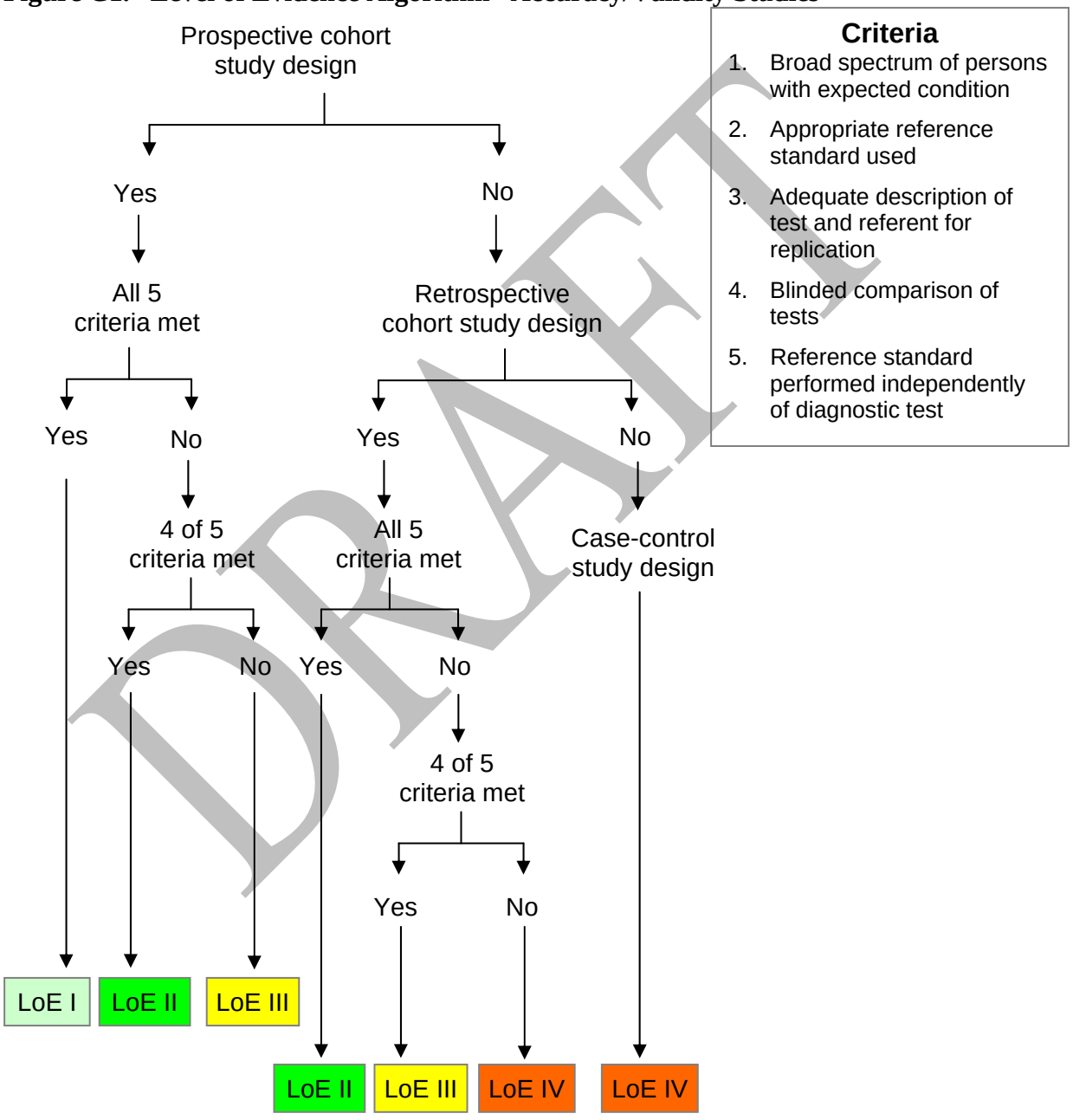
Levels of Evidence for Diagnostic Test Studies (Test Characteristics)
Table C2 and Figure C1 outline Spectrum Research's methodology for evaluating the quality of evidence for diagnostic studies and criteria used to determine the Level of Evidence (LoE). The procedure that follows describes specific considerations used to determine whether or not the various criteria were met. This method takes into account the primary sources of bias for such studies.

Each included study was evaluated independently by two investigators based on the criteria below and a LoE assigned to each article, initially at the abstract level and confirmed when the full articles were reviewed. Discrepancies in LoE determination were resolved by discussion until consensus was achieved.

Table C2. Definitions of the different levels of evidence for diagnostic test accuracy/validity studies.

Level	Study type	Criteria
I	Good quality prospective study	• Broad spectrum of persons with the expected condition • Appropriate reference standard used • Adequate description of test and reference for replication • Blinded comparison of tests with appropriate reference standard • Reference standard performed independently of diagnostic test
II	Moderate quality prospective study	• Violation of any one of the criteria for a good quality prospective study (LoE I)
	Good quality retrospective study	• Broad spectrum of persons with the expected condition • Appropriate reference standard used • Adequate description of test and reference for replication • Blinded comparison of tests with appropriate reference standard • Reference standard performed independently of diagnostic test
III	Poor quality prospective study	• Violation of any two or more of the criteria for a good quality prospective study (LoE I)
	Moderate quality retrospective study	• Violation of any one of the criteria for a good quality retrospective study (LoE II)
IV	Poor quality retrospective study	• Violation of any two or more of the criteria for a good quality retrospective study (LoE II)

Figure C1. Level of Evidence Algorithm –Accuracy/Validity Studies



Procedures for determining adherence to LoE criteria

The following describes the method for determining whether or not a given study has met the specific individual criterion used to assign the LoE. Table C3 provides a template for indicating whether the individual criterion is met or not. A blank for the criterion indicates that the criterion was not met, could not be determined or was not reported by the author.

1. Determine if the study is **prospective or retrospective.**

   Accuracy of diagnostic tests is best assessed using a prospective study of consecutive series of patients from a relevant patient population (i.e. study designed for prospective collection of data using specific protocols). Ideally, a consecutive series of patients or random selection from the relevant patient population should be prospectively studied. Retrospective collection of data or evaluation of patients who have had the diagnostic test and reference test previously may be more subject to bias.

   If it is cannot be determined whether a prospective or retrospective approach was taken, no credit will be given for this criterion having been met.

2. Was a **broad spectrum of persons with the suspected condition** used to evaluate the diagnostic test and reference standard?

   The study population must be comprised of those with a broad spectrum of suspected disease who are likely to have the test now or in the future. A broad spectrum would include patients with mild as well as more severe cases, those presenting early as well as late and those whose differential diagnosis may be commonly confused with the condition of interest. Subjects from specialty referral sources may be more likely to have a specific abnormality/condition than those presenting to a general family practice clinic. Overestimation of diagnostic accuracy may occur if a population with known disease is compared with a group of normal individuals instead of those from the relevant patient population.

   Studies providing a description of the demographic and clinical characteristics of subjects were given credit as appropriate for the type of disease under investigation.

3. Was an **appropriate reference standard** used to compare the diagnostic test being evaluated?

   Ideal reference standards are termed “gold” standards and in theory, provide the “truth” about the presence or absence of a condition or disease. Such standards provide a basis for comparing the accuracy of other tests and allow for the calculation of characteristics such as sensitivity, specificity and predictive values.

In most instances, the reference standard does not perfectly classify individuals with respect to the presence or absences of disease, but may reflect the current "best" reference and/or one that can be practically applied. It should be "likely" to classify patients according to disease status. A reference measure can be performed at the time of the testing. It may be an anatomical, physiological or pathological state or measure or a specific outcome at a later date.

The reference standard should be reproducible and the description of both the referent standard and the test should be explicit enough for replication, validation and generalization.

4. Are the **details of the test and the reference/gold standard sufficient to allow study replication**?

   Are the technical features of the test and protocols used to collect information about test results, any measurements performed, planes of section evaluated, diagnostic criteria used, etc. sufficient that other investigators could duplicate the conditions and reproduce the findings in a similar population?

5. Was there **blinded comparison of the tests with the appropriate reference standard**?

   Interpretation of the reference standard must be done without prior knowledge of the test results and the test must be interpreted without knowledge of the results of the reference test. This is necessary to avoid bias. It must be clear from the text that tests were interpreted without knowledge of the results of the other. A statement that blinding was done (for either test, preferably both) was necessary for credit.

6. Was the **reference standard performed independently of the diagnostic test**?

   The reference standard must have been applied objectively or blindly to all patients without the results of test influencing use of the reference. If the "test" affects the reference (or referral to the reference test) or is part of the reference standard, this does not constitute independent performance of the test.

Table C3. Assessment of LoE for individual studies of diagnostic test evaluation

METHODOLOGICAL PRINCIPLE	Author 1 (1999)	Author 2 (2002)	Author 3 (2004)	Author 4 (2005)
Study Design				
Prospective cohort design			■	
Retrospective cohort design	■	■		■
Case-control design				
Broad spectrum of patients with expected condition	■	■		■
Appropriate reference standard used	■	■		
Adequate description of test and reference for replication				
Blinded comparison with appropriate reference	■		■	
Reference standard performed independently of test	■	■	■	■
Evidence Level	**II**	**III**	**III**	**IV**

* Blank box indicates criterion not met, could not be determined or information not reported by author

Levels of Evidence for Diagnostic Test Studies –Reliability Studies

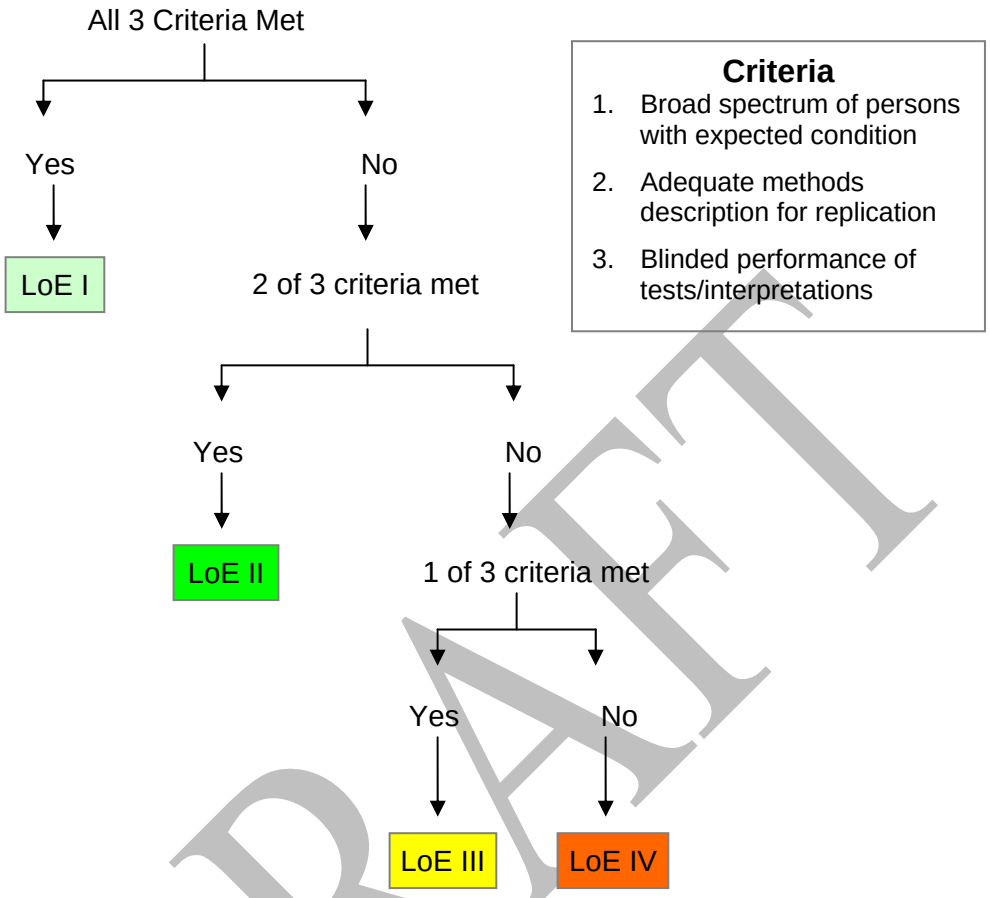
Methods for assessing the quality of evidence for reliability studies have not been well reported in the literature. Spectrum's determination of quality for such is based on epidemiologic methods for evaluating validity and reliability.[6]

The following describes the method for determining whether or not a given study has met the specific individual criterion used to assign the LoE. Table C4 provides a template for indicating whether the individual criterion is met or not. A blank for the criterion indicates that the criterion was not met, could not be determined or was not reported by the author.

Table C4. Definitions of the different levels of evidence for reliability studies

Level	Study type	Criteria
I	Good quality study	• Broad spectrum of persons with the expected condition • Adequate description of methods for replication • Blinded performance of tests, measurements or interpretation • Second test/interpretation performed independently of the first
II	Moderate quality	• Violation of any one of the criteria for a good quality study
III	Poor quality study	• Violation of any two of the criteria
IV	Very poor quality study	• Violation of all three of the criteria

Figure C2. Level of Evidence Algorithm – Reliability studies



Procedures for determining adherence to LoE criteria: Reliability studies

For these studies, the first performance or interpretation of the text is usually considered the "reference" and the second performance or interpretation the "test". Typical reliability studies are done using the same method (e.g., supine MRI) and include test-retest, inter- and intra-rater reliability. Statistical analysis is based on whether the same method or different methods are compared, the types of variables measured and the goal of the study.[6] In general, the degree (%) of concordance does not account for the role of chance agreement and is not a good index of reliability.[7] Different types of *kappa* (κ) or statistical correlation are frequently used to evaluate the role of chance.

Determination of the LoE involves evaluation of the following questions:

1. Was a **broad spectrum of persons with the suspected condition** used to determine reliability?

   The study population must be comprised of those with a broad spectrum of suspected disease who are likely to have the test now or in the future. Since differences in gender, age, body habitus and other characteristics may influence measurements and the ability to

reproduce the results, the range of patients used for reliability studies is important. Ideally a random sample of patients from the relevant clinical population would be used but may not be feasible, depending on the study. A broad spectrum would include patients with mild as well as more severe cases, those presenting early as well as late and those whose differential diagnosis may be commonly confused with the condition of interest. Reproducibility studies in a population with known disease may give different results compared with studies on a group of normal individuals and may not give an accurate picture of overall reproducibility. (If the goal of the study is to evaluate the potential for differential measurement error or bias, the separate analyses on "normal" and "diseased" populations should be done to evaluate the extent of such bias.[6] If it is a test-retest design, the test administrations should be on the same population. If it is an inter- or inter-rater reliability study the object (e.g., radiographs) should be the same for each reading/interpretation, (e.g., the same patients' radiographs are read twice).

2. Are the **details of the methods sufficient to allow study replication**?

   Is the description of the methods, i.e. the protocols used to collect information, measurements taken, planes of section, diagnostic criteria used, etc. sufficient that other investigators could duplicate the conditions and reproduce the findings in a similar population? Are the methods used for each part of the replication consistent?

3. Was there **blinded/independent performance of the repeat test administrations or interpretations**?

   The second administration of the test or second interpretation of results should be done without influence of the first test/interpretation. This is necessary to avoid bias. It must be clear from the text that both tests were interpreted without knowledge of the results of the other. Examples of when the administration would not be considered blinded or independent could include:
   - Interpretation of the second test is to be done without prior knowledge of the test results or the first interpretation.
   - The timing of the second test administration or reading/interpretation of the results is not done such that sufficient time has elapsed between them to avoid influence of the first test/interpretation on the results of the second. In the case of re-administration of the test, the timing should not be so far apart that the stage/period of disease is different from the first administration.

Table C5. Assessment of level of evidence (LoE) for reliability studies

METHODOLOGICAL PRINCIPLE	Author 1 (1999)	Author 2 (2002)	Author 3 (2004)	Author 4 (2005)
Broad spectrum of patients with expected condition	■	■		
Adequate description of methods for replication	■	■		

Blinded/independent comparison of tests/interpretations	■		■	
Evidence Level	**I**	**II**	**III**	**IV**

Determination of Overall Strength of Evidence

Following the assessment of the quality of each individual study included in the report, an overall "strength of evidence for the relevant question or topic is determined. Methods for determining the overall strength of evidence for diagnostic studies are variable across the literature and are most applicable to evaluation of therapeutic studies.

SRI's method incorporates the primary domains of quality (LoE), quantity of studies and consistency of results across studies as described by AHRQ. [5]

The following definitions are used by SRI to determine whether or not the body of evidence meets the criteria for each domain:

Table C6. Overall Strength of Evidence Domains

Domain	Definition/Criterion
Quality	• At least 80% of the studies are LoE I or II
Quantity	• There are at least three studies which are adequately powered to answer the study question
Consistency	• Study results would lead to a similar conclusion (similar values, in the same direction) in at least 70% of the studies

Based on the criteria described above, the possible scenarios that would be encountered are described below. Each scenario is ranked according to the impact that future research is likely to have on both the overall estimates of an effect and the confidence in the estimate. This ranking describes the overall "Strength of Evidence" (SoE) for the body of literature on a specific topic. The method and descriptions of overall strength are adapted for diagnostic studies from system described by the GRADE Working Group[4] for the development of clinical guidelines.

Table C7. Assessment of overall strength of evidence

SoE	Description	Further Research Impact	Domain Criterion Met		
			Quality	Quantity	Consistency
1	**High**	Very unlikely to change confidence in effect estimate	+	+	+
2	**Moderate**	Likely to have an important impact on confidence in estimate and *may* change the estimate	+	-	+
			+	+	-
3	**Low**	Very likely to have an important impact on confidence in estimate and *likely* to change the estimate	+	-	-
			-	+	+
4	**Very Low**	Any effect estimate is uncertain	-	+	-
			-	-	+
			-	-	-

The generalizability (or directness) of the study(ies) to various population is considered and addressed via narrative where applicable.

Assessment of Economic Studies

Full formal economic analyses evaluate both costs and clinical outcomes of two or more alternative interventions. The four primary types are cost minimization analysis (CMA), cost-utility analysis (CUA), cost-effectiveness analysis (CEA), and cost-benefit analyses (CBA). Each employs different methodologies, potentially complicating critical appraisal, but some common criteria can be assessed across studies.

No standard, universally accepted method of critical appraisal of economic analyses is currently in use. A number of checklists [Canadian, BMJ, AMA] are available to facilitate critique of such studies. The Quality of Health Economic Studies (QHES) instrument developed by Ofman, et al.[8] QHES embodies the primary components relevant for critical appraisal of economic studies.[8, 9] It also incorporates a weighted scoring process and which was used as one factor to assess included economic studies. This tool has not yet undergone extensive evaluation for broader use but provides a valuable starting point for critique.

In addition to assessment of criteria in the QHES, other factors are important in critical appraisal of studies from an epidemiologic perspective to assist in evaluation of generalizability and potential sources of study bias.

Such factors include:

- Are the interventions applied to similar populations (eg, with respect to age, gender, medical conditions, etc)? To what extent are the populations for each intervention comparable and are differences considered or accounted for? To what extent are population characteristics consistent with "real world" applications of the comparators?
- Are the sample sizes adequate so as to provide a reasonable representation of individuals to whom the technology would be applied?
- What types of studies form the basis for the data used in the analyses? Data (eg, complication rates) from randomized controlled trials or well-conducted, methodologically rigorous cohort studies for data collection are generally of highest quality compared with case series or studies with historical cohorts.
- Were the interventions applied in a comparable manner (eg, similar protocols, follow-up procedures, evaluation of outcomes, etc)?
- How were the data and/or patients selected or sampled (eg, a random selection of claims for the intervention from a given year/source or all claims)? What specific inclusion/exclusion criteria or processes were used?
- Were the outcomes and consequences of the interventions being compared comparable for each? (eg, were all of the relevant consequences/complications for each intervention considered or do they primarily reflect those for one intervention?)

Assessment of the overall strength of evidence for formal economic analyses does not appear to be documented in the literature. For the purposes of this HTA, overall strength was determined by:

- Quality of the individual studies: Where the majority of quality indicators described in the QHES met and were the methods related to patient/claim selection, patient population considerations and other factors listed above consistent with a high quality design?
- Number of formal analyses (3 or more)
- Consistency of findings and conclusions from analyses across studies.

QHES Instrument[8] **Study** _____

Questions	Points	Yes	No
1. Was the study objective presented in a clear, specific, and measurable manner?	**7**		
2. Were the perspective of the analysis (societal, third-party payer, etc.) and reasons for its selection stated?	**4**		
3. Were variable estimates used in the analysis from the best available source (ie, randomized controlled trial - best, expert opinion - worst)?	**8**		
4. If estimates came from a subgroup analysis, were the groups prespecified at the beginning of the study?	**1**		
5. Was uncertainty handled by (1) statistical analysis to address random events, (2) sensitivity analysis to cover a range of assumptions?	**9**		
6. Was incremental analysis performed between alternatives for resources and costs?	**6**		
7. Was the methodology for data abstraction (including the value of health states and other benefits) stated?	**5**		
8. Did the analytic horizon allow time for all relevant and important outcomes? Were benefits and costs that went beyond 1 year discounted (3% to 5%) and justification given for the discount rate?	**7**		
9. Was the measurement of costs appropriate and the methodology for the estimation of quantities and unit costs clearly described?	**8**		
10. Were the primary outcome measure(s) for the economic evaluation clearly stated and did they include the major short-term, long-term and negative outcomes included?	**6**		
11. Were the health outcomes measures/scales valid and reliable? If previously tested valid and reliable measures were not available, was justification given for the measures/scales used?	**7**		
12. Were the economic model (including structure), study methods and analysis, and the components of the numerator and denominator displayed in a clear, transparent manner?	**8**		
13. Were the choice of economic model, main assumptions, and limitations of the study stated and justified?	**7**		
14. Did the author(s) explicitly discuss direction and magnitude of potential biases?	**6**		
15. Were the conclusions/recommendations of the study justified and based on the study results?	**8**		
16. Was there a statement disclosing the source of funding for the study?	**3**		
TOTAL POINTS	**100**		

Appendix References

1. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *Jama.* Sep 15 1999;282(11):1061-1066.

2. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. *Cmaj.* Feb 14 2006;174(4):469-476.

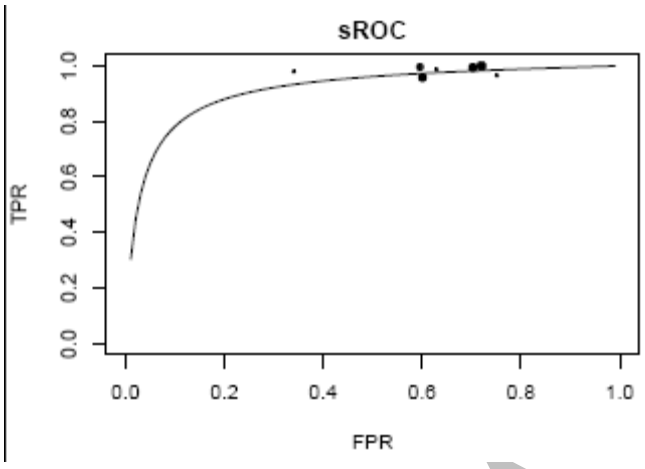
3. Phillips B, Ball C, Sackett D. Levels of evidence and grades of recommendation. Available at: Available at: http://www.cebm.net/levels_of_evidence.asp.

4. Atkins D, Best D, Briss PA, et al. Grading quality of evidence and strength of recommendations. *Bmj.* Jun 19 2004;328(7454):1490.
5. West S, King V, Carey TS, et.al. Systems to Rate the Strength of Scientific Evidence. Evidence Report/Technology Assessment No. 47 (Prepared by the Research Triangle Institute-University of North Carolina Evidence-based Practice Center, Contract No. 290-97-0011): Agency for Healthcare Research and Quality, Rockville, MD; 2002.
6. Armstrong B, White E, Saracci R. *Principles of Exposure Measurement in Epidemiology*: Oxford University Press 1992.
7. Koepsell T, Weiss N. *Epidemiologic Methods: Studying the Occurrence of Illness*. New York: Oxford University Press, Inc; 2003.
8. Ofman JJ, Sullivan SD, Neumann PJ, et al. Examining the value and quality of health economic analyses: implications of utilizing the QHES. *J Manag Care Pharm.* Jan-Feb 2003;9(1):53-61.
9. Chiou CF, Hay JW, Wallace JF, et al. Development and validation of a grading system for the quality of cost-effectiveness studies. *Med Care.* Jan 2003;41(1):32-44.

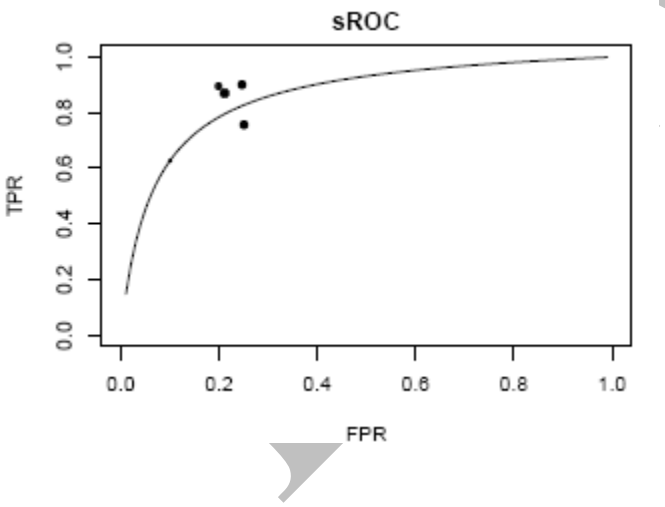
Appendix D
sROC curves for LoE I/II studies

The following sROC curves correspond to studies used in the meta-analyses.

At CACS > 0



At CACS ≥ 100



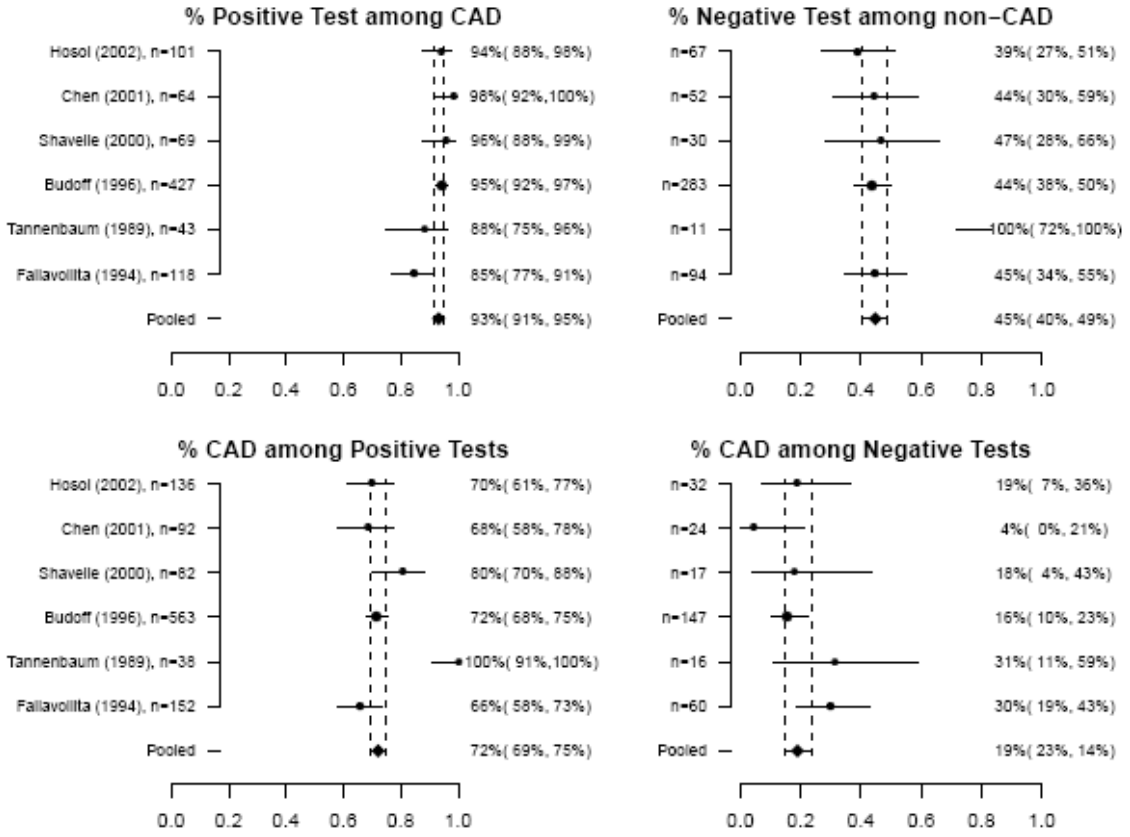
Appendix E
LoE III Studies Additional Results

Data were available from 5 LoE III studies which defined a positive test based on detection of any calcium, i.e. a threshold of > 0. The table below describes the prevalence of obstructive CAD and CACS test results based on the presence of calcium.

Author	N	TP	TN	FP	FN	n CAD	% CAD	LoE	Blinded	Independent test performance
Hosoi (2002) (nondiabetic)	181	95	26	41	6	114	63%	III	y	NR - unclear
Chen (2001)	116	63	23	29	1	64	55.2	III	y	No - only 116/163 had CCA; others refused consent
Shavelle (2000)	97	66	14	16	3	69	71%	III	y	NR - unclear
Budoff (1996)	710	404	124	159	23	427	60%	III	n	NR - unclear
Tannenbaum (1989)	54	38	11	0	5	43	80%	III	y	NR - unclear
Herzog (2004)*	38	17	4	16	1	18	47%	III	y	NR - unclear
Fallavollita (1994)	212	100	42	52	18	118	56%	III	y	NR - unclear

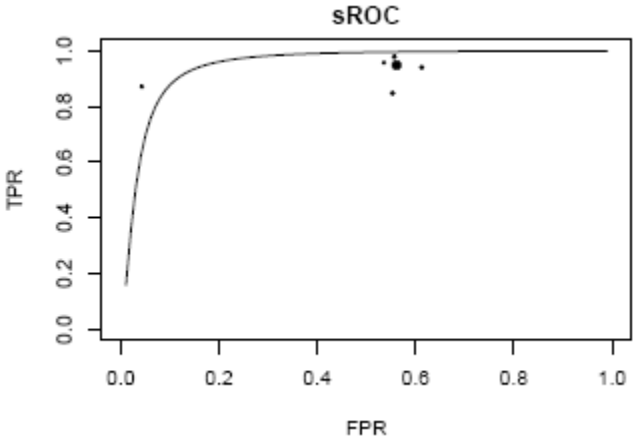
* Herzog used an angiographic cut off of 75% and is not included in the meta-analysis. All others used a ≥ 50% luminal narrowing by angiography

A summary of meta-analysis results for the same CACS test parameters using a threshold of 0 (any calcium detected) is given in the figure below. The point estimates and 95% confidence intervals for individual studies and for the pooled estimate are given.



*The upper panels show individual study and pooled estimates of sensitivity and specificity respectively. The lower left panel shows the positive predictive value and the lower right panel represents 1- negative predictive value. The sample size n refers to numbers of subjects in the denominator, that is, subjects with CAD in upper left, without CAD in upper right, with positive tests in lower left, and with negative tests in lower right.

Corresponding sROC curve



Only 1 LoE III study had data at a cut off of 100

Author	N	TP	TN	FP	FN	n CAD	% CAD	LoE	Blinded	Indep performance
Hosoi (2002) (diabetic)	100	59	9	3	29	88	88%	III	y	NR - unclear
Hosoi (2002) (nondiabetic)	181	75	56	11	39	114	63%	III	y	NR - unclear

Test characteristics at CACS > 100 for Hosoi (2002) (nondiabetic) – point estimates and 95% confidence interval

Sensitivity	65.8%	57.1%	74.5%
Specificity	83.6%	74.7%	92.5%
PPV	87.2%	80.2%	94.3%
NPV	58.9%	49.1%	68.8%

Although 2 studies had data at a cut off of 400, Hosoi used an angiographic cut off of 50% and Herzog used a cut off of 75%

Author	N	TP	TN	FP	FN	n CAD	% CAD	LoE	Blinded	Indep performance
Hosoi (2002) (diabetic)	100	43	11	1	45	88	88%	III	y	NR - unclear
Hosoi (2002) (nondiabetic)	181	50	65	2	64	114	63%	III	y	NR - unclear
Herzog (2004)*	38	12	16	4	6	18	47%	III	y	NR - indep unclear

*uses a 75% stenosis level determined by angiography

Test characteristics at CACS > 400 for Hosoi (2002) (nondiabetic) – point estimates and 95% confidence interval

Sensitivity	43.9%	34.8%	53.0%
Specificity	97.0%	92.9%	100%
PPV	96.2%	90.9%	100%
NPV	50.4%	41.8%	59.0%

Appendix F
Summary of CACS LoE IV studies

Author (year)	CACS cutoff and angiography criteria	% CAD	Sens	Spec	PPV	NPV
Konieczynska (2006) N = 340 65% male Mean age 60 years	CACS cutoff ≥56 Stenosis of ≥50% of main coronary artery by angiography n = 178 with significant stenosis	52% (178/340)	85.7%	85.3%	86.3%	93.5% in abstract 84.8% in table III (100% for women)
	CACS cutoff >0	52%	96.6%	52.4%	NR	NR
Haberl (2005) N = 133 62% male Mean age 67 years	CACS cutoff >0 Stenosis of >50% obstruction by angiography n = 53 with significant stenosis	40% (53/133)	85%	24%	46%	68%
Yao (2004) N = 73 Mean age 53 years	CACS cutoff of ≥ 130 HU Stenosis of ≥50% by angiography n = 35 with CAD	48%				
	SPECT		80.0%	92.1%	NR	NR
	EBCT		77.1%	55.3%	NR	NR
Shivastava (2003) Indian population N = 388 84% male mean age 53 years (range, 15-78 years)	CACS cutoff >0 (n = 314)	95% (n = 298)	95.5%	78.9%	94.9%	81.1%
	CACS cutoff >100 (n = 240)	98% (n = 236)	75.6%	94.7%	98.3%	51.3%
	CACS cutoff > 400 (n = 72)	100% (n = 72)	23.1%	100%	100%	24.1%
	Significant CAD defined as ≥70% stenosis by angiography	80% (n = 312)				
Bielak (2000) N = 213 76% male	CACS cutoff > 0 (adjusted for verification bias) Stenosis of ≥50% by angiography n = 112 with obstructive CAD	53%	97.0%	72.4%	NR	NR
	CACS cutoff > 0 (unadjusted)		99.1%	38.6%	NR	NR
Yao (2000) N = 64	CACS cutoff of Stenosis of ≥50% by angiography n = 45 with coronary stenosis	70%				
	Group A (n = 40, >45 years old)		SPECT: 93.6% EBCT: 90.3%	88.9% 55.6%	NR NR	NR NR
	Group B (n = 24 ≤ 45 years old)		SPECT: 92.9%	100.0%	NR	NR

			EBCT: 42.9%	100.0%	NR	NR
Baumgart (1997) N = 57 79% male Mean age 54 years	CACS > 0 Stenosis of ≥50% by angiography n = 29 with significant CAD	51%	66%	78%	39%	91%
Rumberger, Sheedy, Breen, Schwartz (1997) N = 213 71% male Mean age 50 years	CACS cutoff and angio disease optimized					
	Max % stenosis	Max % stenosis	Max % stenosis	Max % stenosis		Max % stenosis
	≥20%	≥20%	≥20%	≥20%		≥20%
	≥30%	≥30%	≥30%	≥30%		≥30%
	≥40%	≥40%	≥40%	≥40%		≥40%
	≥50%	≥50%	≥50%	≥50%		≥50%
	≥60%	≥60%	≥60%	≥60%		≥60%
	≥70%	≥70%	≥70%	≥70%		≥70%
	≥80%	≥80%	≥80%	≥80%		≥80%
	≥90%	≥90%	≥90%	≥90%		≥90%
	100%	100%	100%	100%		100%
Seese (1997) N = 120 Mean age 55 years	CACS: two adjacent pixels with CT number of at least 130 HU Stenosis of ≥50% by angiography n = 87 with significant CAD	73%	95%	72%	94%	81%
Broderick (1996) N = 101 66% male Mean age 61 years (range, 31-84 years) sensitivity, specificity, PPV, NPV values reported here were obtained using contiguous slice-step density algorithm	Stenosis of ≥50% by angiography n = 68 with significant CAD	67%				
	CACS ≥90 HU		88%	52%	79%	68%
	CACS ≥130 HU		81%	61%	81%	61%
Bielak (1994) (same pop as 2000 - reliability study) N = 160	CACS > 3.10 mm^2 area Stenosis of ≥50% by angiography n = 69 of 160 with significant CAD	43%	87%	66%	NR	NR
	CACS > 0.52 mm^2 area		100%	34%	NR	NR

CACS: coronary artery calcium score; CAD: coronary artery disease; EBCT: electron beam computed tomography; NPV: negative predictive value; NR: not reported; PPV: positive predictive value; Sens: sensitivity; Spec: specificity; SPECT: single photon emission computed tomography.

Appendix G
Safety and Clinical Decision Making Data

Safety

Radiation exposures reported in validation studies used in Spectrum Research HTA

Reference	CT Modality	Slice thickness	ECG trigger	Radiation
Calcium scoring only				
Bielak 1994	Electron beam	3 mm	ECG triggering at same phase of cardiac cycle	10 mGy [10 mSv] total dose to the skin
Fallavollita 1994	Electron beam	3 mm	Triggering at same point of RR interval	<0.005 Gy [< 5 mSv]
Budoff 1996	Electron beam	3 mm	Triggering at 80% of RR interval	< 1 rad (< 0.01 Gy) [<10 mSv]
Shavelle 2000	Electron beam	3 mm	Triggering at 80% of RR interval	< 1 rad [<10 mSv]
Budoff & Diamond 2002	Electron beam	3 mm	Triggering at 80% of RR interval	0.6 rad [6 mSv]
Lamont 2002	Electron beam	NR	NR	7 mGy [= 7mSv] to skin at back
Becker 2007	Multi-detector helical	3 mm	Retrospective gating	1.5 mSv for men, < 2 mSv for women
Calcium scoring conducted as first component with CT angiography				
Haberl 2005	Multi-detector	NR	Prospective gating	For calcium score alone: 1.2 to 1.8 mSv For CT

				angiography alone: 5.8 to 7.4 mSv for men; 7.6 to 9.8 mSv for women
Leschka 2008	Dual source	3 mm	Pulsing	For calcium score with CT angiography: 7-9 mSv

CT is computed tomography; mSv is millisieverts; Gy is Gray; NR is not reported
Radiation exposure was not reported in Kajinami 1995; Lau 2005; Nixdorff 2008; Knez 2004; Haberl 2001; Chen 2001; Almeda 2004; Hosoi 2002; Leber 2001; Tanenbaum 1989; Herzog 2004; Guerci 1998; Konieczynska 2006; Rumberber &Sheedy 1997; Bielak 2000; Yao 2000; Seese 1997; Baumgart 1997; Shrivastava S 2003; Yao 2004

Types of clinically important[1] incidental findings

	Hunold[2] 2001	Horton 2002	Schragin 2004	Law[3] 2008
Mediastinum	**12**	**1**	**2**	
Enlarged lymph nodes	6		2	
Thickened esophageal wall	3	1		
Thymus	1			
Miscellaneous	2			
Lung	**13**	**89**	**52**	**5**
Pleural scar	2			
Fibrosis	1		3	
Nodule, suspicious for tumor	8	65	46	3
Tuberculoma	1			
Infiltrate/consolidation/effusion		24	3	1
Airway dilatation				1
Miscellaneous	1			
Pulmonary mass			1	
Breast		**2**		**1**
Bones: Sclerotic lesion		**2**		
Abdomen	**16**	**9**	**1**	
Diaphragmatic hernia/ dehiscence	2			
Liver tumor	4			
Liver cyst	7			
Liver lesion, indeterminate		7	1	4
Polycystic liver		1		

Liver, miscellaneous	3		
Ascites		1	
Vascular	**9**	**1**	
Ectasia/aneursym	8		1
Dissection	1		
Heart[4]	**136**		**2**

[1] Clinically important as defined by individual study. Hunold et al and Law et al reported incidental findings with diagnostic or therapeutic consequences separately. For consistency with other studies, only findings with diagnostic consequences are reported here. Elgin et al did not report distribution of clinically important incidental findings.

Clinical Decision Making Studies – Data abstraction

Study (year)/study design	Population	Patient characteristics	Observation period	CACS cut-off	Outcomes	Author conclusion/clinical decision
Piers (2008) Retrospective	Tertiary referral center N = 598 Male: 57% Mean age: 55 years % F/U: NA	• Referred for evaluation of CAD with no prior history of CAD • Underwent additional diagnostic procedures for evaluation of ischemia as clinically relevant according to the treating cardiologist • Underwent angiography if clinically relevant	No clinical f/u beyond day of testing	< 10 10-99 100-399 ≥ 400	Angiography according to CACS (*P* < .001) • CACS < 10: 9% • CACS 10-99: 22% • CACS 100-399: 34% • CACS ≥ 400: 86% Angiography despite negative ischemia detection tests (*P* <.001) • CACS < 10: 1% • CACS 10-99: 16% • CACS 100-399: 29% • CACS ≥ 400: 88% Angiography following positive ischemia test (*P* < .05) • CACS < 10: 44% • CACS 10-99: 47% • CACS 100-399: 38% • CACS ≥ 400: 88% (14/16)	• CACS impacted whether to perform additional invasive coronary angiography or not • Referral for angiography was more likely chosen in cases of a higher CACS, especially in patients with a negative ischemia test

CACS: coronary artery calcium score; CAD: coronary artery disease; NA: not applicable.

Study (year)/ design	Population	Patient characteristics*	Observation period	CACS cut-off	Outcomes	Author conclusion/clinical decision
Esteves (2008) Unstated	Chest pain unit N = 84 Male: 39% Mean age: 62 years F/U: 100%	• Low to intermediate risk • Angina-like chest pain • Normal or nondiagnostic ECG on admission • 2 negative sets of troponin I • Underwent subsequent adenosine stress Rb-82 PET/CT myocardial perfusion imaging	No clinical f/u beyond day of testing†	0 vs >0	normal PET • CACS = 0: 100% (34/34) • CACS > 0: 74% (37/50) normal LVEF (>50%) (calculated in 72/84) • CACS = 0: 97% (30/31) • CACS > 0: 90% (37/41)	• Absence of CAC is predictive of a normal adenosine stress Rb-82 myocardial perfusion PET • myocardial perfusion imaging probably can be safely avoided in chest pain patients with a CACS = 0
Geluk (2008) Prospective Clinical decision flow diagram provided	ED N = 304 Male: 56% Mean age: 55 years (26-85) F/U: 100%	• Low risk, stable • Chest pain or other CAD symptoms • Normal ECG • Normal troponin	4 months	< 10 10-399 ≥ 400	combined endpoint of obstructive CAD (> 50% stenosis) on angiography and revascularization, MI, or cardiac death by medial records, phone contact, or general practioners • CACS < 10: 0% (0/159) • CACS 10-399: 14% (14/103) • CACS ≥ 400: 57% (24/42)	CACS may be used as a "gatekeeper" for additional invasive and noninvasive testing • CACS < 10 may facilitate safe discharge to home • CACS 10-399 suggests need for noninvasive testing and/or primary prevention • CACS ≥ 400 suggests need for coronary angiography
Georgiou (2001) Prospective enrollment retrospective chart review done for endpoints	ED N = 208 Male: 54% Mean age: 53 years F/U: 92% (n = 192)	• Low risk, stable • Age ≥ 30 years • Chest pain lasting 20 minutes or more within past 12 hours • nondiagnostic ECG • Believed to require admission to exclude MI	50 months (1-84 months)	0 1-4 5-332 > 333	"Hard events": cardiac death or nonfatal MI by hospital record review • CACS = 0: n = 0 • CACS 1-4: n = 1 • CACS 5-332: n = 10 • CACS > 333: n = 19 • Majority of patients who suffered a "hard event" had a	• CACS by EBCT allows for early discharge with a negative test • CACS by EBCT may be a sufficiently powerful tool to be used in the ED to decide the need for admission in patients presenting with chest pain and nondiagnostic ECG

					CACS in the upper range of normal (>50th percentile: 88%; >75th percentile: 64%) All cardiovascular events: cardiac death, nonfatal MI, revascularization, ischemic stroke, or angina by hospital record review • CACS = 0: n = 2 • CACS 1-4: n = 1 • CACS 5-332: n = 27 • CACS > 333: n = 27 Annualized event rate • CACS = 0: 0.6% • CACS 1-100: 6% • CACS 101-400: 10% • CACS > 400: 13.9%	• CACS = 0 implies a very low incidence of coronary events • CACS > 0 is an independent predictor of future cardiac events
Laudon (1999) Prospective	ED N = 105 Male: 54% Mean male age: 45 years (30-58) Mean female age: 51 years (40-65) F/U: 95%§	• Low risk, stable • Men age < 55 years and women age < 65 years • Angina-like chest pain requiring hospitalization • Nondiagnostic ECG • Normal cardiac enzymes	minimum of 4 months	0 vs > 0	• CACS = 0 and negative cardiac test: 53% (53/100) • CAC = 0 and positive cardiac test: 1% (1/100) • CACS > 0 and negative cardiac test: 32% (32/100) • CAC > 0 and positive cardiac test: 14% (14/100) • All patients with CACS = 0 remained free of cardiac events	• EBCT-derived CACS can possibly be used a mean of triage for patients with chest pain in the ED • Negative EBCT scan may allow early discharge of patients without further testing, with referral to PCP for outpatient care • Positive EBCT scan requires further in-hospital evaluation, whether in a CPU or as an inpatient
McLaughlin (1999) Unstated	ED N = 134 Male: 37%	• Low risk • Chest pain or to rule out MI • Normal or	30 days	> 1	Negative EBCT scan (CACS ≤ 1): 36% (48/134) • coronary event rate:	• Compared with clinical variables alone, negative EBCT may allow

EBCT done after hospital admission	Mean age: 53 years F/U: 100%	nondiagnostic ECG			2% (1/48) ‡ Positive EBCT scan (CACS > 1): 64% (86/134) • coronary event rate: 8% (7/86) • acute MI: 4/7 (57%) • CABG: 2/7 (29%) • PCI: 1/7 (14%)	for early ED discharge of patients directly from ED • EBCT scanning is a practical technique in terms of risk stratification for use in the ED in this patient population: EBCT can accurately distinguish between very low and very high risks for adverse cardiac events. • EBCT should not be used as a risk stratification method in cocaine users

CABG: coronary artery bypass graft; CACS: coronary artery calcium score; CAD: coronary artery disease; CPU: chest pain unit; CT: computed tomography; ED: emergency department; LVEF: left ventricular ejection fraction; MI: myocardial infarction; PCI: percutaneous coronary intervention; PCP: primary care provider; PET: positron emission tomography.
*All studies excluded patients if they had a prior history of CAD except for Esteves in which there is no mention of exclusion criteria.
†Normal myocardial perfusion PET is used as a proxy for good short term outcome. Authors cite two studies concluding that "myocardial perfusion PET is a widely accepted tool to exclude functionally significant coronary stenosis and is associated with a very low risk of short-term cardiac events."
§All 105 patients were initially enrolled and underwent EBCT scans within 24 hours of admission. One hundred of the patients (95%) also underwent other cardiac testing at the discretion of the attending physician. Authors use these 100 patients to compare other cardiac testing results to EBCT results.
‡This one patient was a 45-year-old male with a history of cocaine abuse and a positive toxicology screen on admission. Excluding this patient, there were no events in this patient group.

Appendix H. Evidence Tables: LoE I – III Validation Studies

LoE I- II validation studies

Author (year)	LoE	Demographics	Inclusion criteria	Exclusion criteria	Clinical profile/patient characteristics	Outcomes
Leschka (2008)	I single center	N = 74 age: 62 (± 12) years (range, 16-86 years) % male: 68	• stable clinical conditions (CCS class I-II and New York Heart Association functional class I-III)	• previous allergic reaction to iodinated contrast media • renal insufficiency	*Characteristics* • average heart rate (bpm) all: 67.7 (35-102) without CAD: 69.4 (47-102) with CAD: 65.8 (35-94) • BMI (kg/m^2) all: 27.2 (4.0) without CAD: 26.1 (3.5) with CAD: 28.3 (4.3) *P* < .05 • arterial HTN all: 59.5% (44/74) without CAD: 66.4% (25/38) with CAD: 51.4% (19/36) • DM type II all: 25.7% (19/74) without CAD: 17.7% (7/38) with CAD: 32.7% (12/36) • smoking all: 27.0% (20/74) without CAD: 13.3% (5/38) with CAD: 42.1% (15/36) *P* < .01 • hyperlipidemia	21 patients (28.4%) had known CAD (stenosis > 50% identified by previous coronary angiography) *Significant stenosis* (> 50%), n = 36, 48.6% *No. diseased vessels* • single: 10.8% (8/74) • multi: 37.8% (28/74) *Calcium scores/cut-offs* • mean all: 710 (0-4387) without CAD: 215 (0-1970) with CAD: 1253 (17-4387) *P* < .001 • 0 all: 18.9% (14/74) without CAD: 36.8% (14/38) with CAD: 0 • 1-399 all: 37.8% (28/74) with CAD: 18.4% (7/38) 58.3% (21/36) • ≥ 400 all: 43.2% (32/74) with CAD: 44.8% (17/38) without CAD: 41.7%

					all: 23.0% (17/74) with CAD: 17.7% (7/38) without CAD: 28.0% (10/36) *Presenting symptoms* • typical angina (n = 40) • atypical angina (n = 19) • pathological exercise test (n = 12) • dyspnea (n = 9)	(15/36)
Kajinami (1995) age and gender	I single center	N = 251 age: 56 (± 14) years % male: 69.3	• elective coronary angiography between May 1991 and May 1993 • chest pain on exertion or at rest or both suggesting angina pectoris • ECG findings at rest that indicated possible myocardial ischemia	• unstable condition • previous coronary interventional procedures (bypass surgery or angioplasty) • abnormal Q-waves in 2+ ECG leads (presumed known CAD)	NR	significant CAD defined as ≥ 75% densitometric stenosis calcium scoring as follows: • 1 = 130-199 HU • 2 = 200-299 HU • 3 = 300-399 HU • 4 = ≥ 400 HU log-transformed total calcium score for prediction of coronary atherosclerosis
Nixdorf (2008)	II	*ITT* N = 79 age: 62 years % male: 59 *per-protocol* N = 71	• elective coronary angiography due to symptoms suspicious of CAD • primary diagnostic procedure, i.e. no previous MI, coronary intervention, or	• severe arterial HTN • severe arrhythmia • atrial fibrillation • valve disease • contraindications to the IV application of dobutamine or X-ray contrast • stable and regional clinical condition • normal global left	NR	significant CAD defined as ≥ 70% stenosis • ITT: 43% (34/79) • per-protocol: 46% (33/71) No. diseased vessels (*ITT*) • 1-vessel, n = 26 • 2-vessel, n = 5 • 3-vessel, n = 3 average calcium score: 321 (0-

			surgery	ventricular function by echocardiography		2442) calcium cut-off score for CAD: ≥ 400
Becker (2007) Agatston and VCS	II single center	N = 1347 age: 60 (± 21) years % male: 59.6	NR	• severe arrhythmias • unstable clinical condition • documented CAD or bypass surgery • referral for a coronary intervention	*Characteristics* • BMI: 27 ± 4 kg/m2 • HTN: 62% (n = 622) • hypercholesterolemia: 40% (n = 538) • DM: 17% (n = 229) • smoking: 25% (n = 334) *Presenting symptoms* • typical angina: 49% (n = 666) • atypical angina: 35% (n = 470) • exertional dyspnea: 13% (n = 175) • heart failure: 3% (n = 40)	significant CAD defined as ≥ 50% stenosis, 53% (720/1347) *calcium scores* • mean Agatston score: 401 ± 382 (range, 0-6941) • mean volumetric score: 348 ± 299 (range, 0-5287) • *gender*: scores were higher in males vs. females across all age groups independent of angiographic status, P = .001 • *CAD*: patients with vs. without had higher mean scores, independent of age and sex: Agatston = 497 ± 987 vs. 97 ± 112; P < .01 volumetric = 483 ± 527 vs. 89 ± 201; *P* < .01 *calcium cut-off scores used* • > 0, > 10, >100 • ≥ 25th, ≥50th, ≥75th
Lau (2005)	II	N = 50 age: 62 (± 11) years male: 62 years (range,	• heart in sinus rhythm • elective conventional coronary	• previous coronary stent placement or bypass grafting • serum creatinine level higher than the normal	NR	CAD defined as ≥ 50% stenosis, 60% (30/50) *No. diseased vessels* • single-vessel: n = 14

		37-78 years); female: 61 years (range, 36-75 years) % male: 80	angiography for suspected CAD	range • allergy to iodine or IV contrast material		• multi-vessel: n = 16 *calcium cut-off scores used* • ≥ 1, ≥ 50, ≥ 400
Knez (2004) VCS (also Agatston)	II single center	N = 2115 age: 62 (± 19) years % male: 66.4	• symptomatic • referral by primary physician due to concern for possible presence of myocardial ischemia	• no prior diagnosis of CAD	*Characteristics* • HTN: 66% (n = 1422) • hypercholesterolemia: 48% (n = 1023) • DM: 22% (n = 470) • smoking: 23% (n = 486) *Presenting symptoms* • typical or atypical chest pain: 80% (n = 1697) • exertional dyspnea: 12% (n = 258) • heart failure: 8% (n = 160) • abnormal stress test: 52% (n = 1391)	CAD defined as ≥ 50% stenosis • all: 59% (1255/2115) • male: 62% (872/1404) • female: 54% (383/711) *volumetric calcium cut-off scores used* • > 0, > 10, >100 • ≥ 25th, ≥50th, ≥75th *calcium scores* • mean Agatston: 323 ± 842 (range, 0-7224) • mean volumetric: 310 ± 714 (range, 0-5490) • with and without CAD Agatston: 492 ± 1,124 vs. 76 ± 217, P < .01 volumetric: 486 ± 940 vs. 53 ± 175, *P* < .01
Budoff and Diamond (2002) (To adjust for	II multi-center	N = 1851 age: 58 (± 11) years (range, 21-86) % male: 63%	• primary physician's concern for the presence of myocardial ischemia based on	• electron beam tomography scans performed > 3 months from the angiogram • previous coronary interventional procedures	NR	CAD defined as ≥ 50% stenosis, 53% (n = 983) calcium score cutoffs used: > 0, > 20, > 80, > 100

verificiation bias, 4103 asymptomatic persons referred for CAC assessment to measure cardiovascular risk were evaluated)		training sample, n = 932* age: 58 (± 11) years % male: 65 validation sample, n = 919* age: 58 (± 11) years % male: 61	positive noninvasive stress testing, abnormal echocardiogram, or clinical history	• known CAD		
Lamont (2002)	II multi-center	N = 153 age: 58 (± 9) years % male: 76	• symptomatic patients with a positive treadmill stress test according to standard criteria who then underwent coronary angiography • all patients referred by primary physicians to evaluate the possibility of an ischemic cause for the symptoms	• known history of CAD (ie, MI, percutaneous transluminal coronary angiography, coronary artery bypass graft) • electrocardiographic evidence of prior MI • know potential causes of nonischemic ST depression that may results in a false-positive treadmill stress test result	*Characteristics* • HTN: 68% • DM: 26% • dyslipidemia: 54% • premature family history: 43% • current smoker: 26% • former smoker: 43% • concurrent medications: oral nitrate: 36%; calcium channel blocker: 20%; beta blocker: 22%; other antihypertensive: 42% • abnormal electrocardiogram: 48% *Presenting symptoms* • typical angina: 37% • atypical angina: 39% • possible non-cardiac: 24%	CAD defined as ≥ 49% stenosis • all: 73% • age > 50 years (n = 27): 56% • age 50-60 years (n = 59): 76% • age > 60 (n = 67): 78% calcium score > 0 • all: 81% • age > 50 years (n = 27): 59% • age 50-60 years (n = 59): 86% • age > 60 (n = 67): 85%

Leber (2001)	II single center	N = 93 age: 59 (± 9) years % male: 85	• suspected CAD • chest pain with an atypical pain character, an atypical pain localization , or an unusual trigger	• unstable angina pectoris • prior coronary interventions (stent implantation or bypass surgery)	• mean calcium score overall: 318 ± 464 (range, 0-2229) • mean calcium score in CAD: 592 ± 587 (range, 0-2229)	CAD defined as ≥ 50% stenosis, 47% (44/93) calcium cut-offs used • 0-45 (low), n = 31 • 46-310 (borderline), n = 31 • > 310 (high), n = 31 calcium scores • total mean: 318 ± 464 (range, 0-2229) • CAD mean: 592 ± 587 (range, 0-2229) • without CAD mean: 137 ± 210 (range, 0-775) *P* < .001
Haberl (2001)	II single center	N = 1764 age range: 20-80 years male age: 56 ± 14 years female age: 60 ± 16 years % male: 69	• typical or atypical chest pain and/or signs of myocardial ischemia on noninvasive tests (bicycle stress test in most cases) • clinical indication for cardiac catheterization	• documented CAD before cardiac catheterization • specifically referred for coronary interventions	*Characteristics* • smoking: 41% *Presenting symptoms* • "chest pain" compatible with angia: 65% • abnormal stress test: 52% (460/920)	CAD defined at ≥ 50 stenosis, 53% (940/1764) male: 60% (685/1225) female: 47% (255/539) calcium score cut-offs used: > 0, ≥ 20, ≥ 100, ≥ 75th
Kwok (2000)	II single center	N = 42 age: 55 (± 10) years % male: 79	• recent MI, unstable angina pectoris, or positive stress test	• previous cardiac interventions (coronary angioplasty, coronary bypass surgery) • renal failure	*Presenting symptoms* • MI: 19% (n = 8) • unstable angina: 40% (n = 17) • chest pain + abnormal stress test: 40% (n = 17)	CAD defined as ≥ 50 stenosis in any of the 3 major coronary arteries or their respective large branches, 76% (32/42) calciums score cut-offs used: ≥

						100 (50% sensitivity), ≥ 160 (63% sensitivity)
Khaleeli† (2001)	II	*Diabetics* N = 323 n = 168 symptomatic n = 155 asymptomatic age: 58 (± 9) years (range, 31-82 years) % male: 64	• chest pain or anginal equivalent • coronary angiography for suspicion of CAD	• declined participation • refusal to sign informed consent • previous revascularization	NR	CAD defined as ≥ 50 stenosis in any of the epicardial coronary vessels calciums cores cut- offs used: > 0, > 102

BMI: body mass index; CAD: coronary artery disease; CCS: Canadian Cardiac Society; DM: diabetes mellitus; HTN: hypertension; ITT: intention-to-treat, IV: intravenous; LoE: level of evidence; MI: myocardial infarction.
*Patients were divided into two samples by a random number generator. The training sample was used for generation of 4 different logistic progression models.
†Study in diabetic persons. Data for symptomatic non-diabetic patients is reported in Budoff 1996.

LoE III validation studies

Author (year)	Demographics	Inclusion criteria	Exclusion criteria	Clinical information	Outcomes
Herzog (2004)	N = 38 Male: 79% Mean age: 62 years (29-65)	• Symptomatic but atypical chest pain • Intermediate pretest (ie, pre-MDCT) likelihood for coronary artery disease but at same time symptomatic chest pain	• NR	• HTN: n = 31 • DM: n = 13 • Nicotine abuse > 1 pack/day: n = 29 • Familial CAD: n = 26	• significant CAD defined as ≥75% stenosis
Hosoi (2002) (diabetic)	N = 101 Male: 70% Age: 64 years	• Presented with chest pain suggestive of angina pectoris or with ambiguous symptoms but resting ECG findings suggestive of myocardial ischemia	• NR	• HTN: 66% (n = 66) • Lipidemia: 30% (n = 30) • Medications to control diabetes: 63% (n = 63)	• significant CAD defined as ≥50% stenosis
Hosoi (2002) (nondiabetic)	N = 181 Male: 72% Age: 62 years			• HTN: n = 92 • Lipidemia: n = 136	

Chen (2001) Chinese	N = 163 age: 65.6 (± 9.7) years (range, 35-84 years) % male: 85	• multiple cardiovascular risk factors or evidence of myocardial ischemia confirmed by a positive treadmill exercise test or thallium-201 myocardial scintigraphy	• no acute coronary syndrome (including Q wave or non-Q wave acute MI or unstable angina • no previous coronary revascularization, including balloon angioplasty, stenting, and coronary bypass surgery	• HTN: 58.9% (n = 96) CAD: 73.0 % (n = 54) no CAD: 33.3 % (n = 14) *P* < .001 • DM: 15.9% (n = 26) CAD: 24.3% (n = 18) no CAD: 7.1% (n = 3) • Hypercholesterolemia: 45.4% (n = 74) CAD: 54.1% (n = 40) no CAD: 33.3% (n = 14) • Smoking: 52.8% (n = 86) CAD: 60.8% (n = 45) no CAD: 53.7% (n = 22) • Family history of CAD: 31.9% (n = 52) CAD: 44.6% (n = 33) no CAD: 9.5% (n = 4) *P* < .001 • Old age: 78.5% (n = 128) CAD: 81.1% (n = 60) no CAD: 66.7% (n = 28) • History of MI: 8.6% (n = 14) CAD: 17.6% (n = 13) no CAD: 0	coronary angiograms were performed in 116 of 163 patients (71.2%) CAD defined as ≥ 50% stenosis, 55% (64/116) calcium score cut-off used: > 0 (> 5, > 75 , > 500) • CAD: 63/64 (98%) • no CAD: 29/52 (56%) • history of MI: 14/14 (100%) • no history of MI: 111/149 (74%) • age ≤ 60 years: 19/38 (50%) • age 61-70 years: 58/70 (83%) • age 71-80 years: 44/50 (88%) • age > 80 years: 4/4 (100%)
Shavelle (2000)	N = 97 Male: 69% Age: 54 years (30-73)	• Symptomatic patients • EBCT studies done within three months of the coronary angiograms • Normal baseline ECGs, without left bundle branch blocks or resting ST segment or T-wave	• History of cardiac valve replacement, coronary stent procedures, or coronary artery bypass grafting prior to the completion of all testing methods	• DM: 26% (n = 25) • HTN: 70% (n = 68) • Hypercholesterolemia: 43% (n = 42) • Family history of CAD: 49% (n = 48) • Tobacco use: 62% (n = 43) • Postmenopausal female:	• significant CAD defined as ≥50% stenosis (n = 67, 69%) • calculated sensitivity, specificity, PPV, NPV, for treadmill-ECG, technetium-stress, CAC by

		changes • At least 85% of the maximum predicted heart rate achieved during treatmill-ECG • Technetium stress testing performed at the same time as treatdmill-ECG • 90% had noninvasive testing (treatmill-ECG and technetium stress) prior to angiography		86% (n = 83) Coronary artery disease: 1-vessel: n = 25 Multivessel: n = 42	EBCT, treadmill ECG combined with EBCT • CAC thresholds of >0 and ≥80
Guerci (1998)	N = 290 Male: 71% Mean age: 59 years	• Patients scheduled for elective cardiac catheterization for clinical indications	• Prior coronary artery bypass surgery, percutaneous transluminal coronary angioplasty, MI, or previous cardiac catheterization	• NR	• obstructive coronary artery disease defined as ≥50% stenosis (n = 116, 40%)
Budoff (1996)	N = 710 Male: 64% Age: 56 years (24-86)	• Coronary angiography for suspicion of coronary artery disease or evaluation of other cardiac disease	• Ultrafast CT performed more than 3 months after angiogram	CAD: 1-vessel: n = 174 (25%) 2-vessel: n = 120 (17%) 3-vessel: n = 111 (16%) 4-vessel: n = 22 (3-vessel + left main artery	• significant CAD defined as ≥50% stenosis (n = 470, 60%)
Fallavollita (1994)	N = 106 Male: 74% Age < 50 years (range 25-49)	• Fast CT scans obtained 59 ± 29 days after angiography • Routine clinical indications for angiography	• NR	• White: 91% (n = 96) • DM: 12% (n = 13) • Family history of CAD: 63% (n = 67) • Cholesterol ≥ 200 mg/dL: 63% (n = 67) • HTN: 43% (n = 46) • Tobacco abuse: 54% (n =	• significant CAD defined as ≥50% stenosis (n = 59, 56%)

				57) CAD: 1-vessel: 26% (n = 28) 2-vessel: 13% (n = 14) 3-vessel: 16% (n = 17)	
Tannenbaum (1989)	N = 54 Male: 67% Mean age: 54 years (21-79)	• Mean interval between the two imaging studies was 36 days • Retrospectively identified as having both coronary arteriorgraphy and ultrafast CT • Performed for clinical indications other than detection of coronary artery calcium	• NR	CAD: 1-vessel: n = 11 2-vessel: n = 18 3-vessel: n = 12 Left main artery: n = 2	• significant CAD defined as ≥70% stenosis, expect for the left main artery where ≥50% was considered significant (n = 43, 80%)

CAD: coronary artery disease; CT: computed tomography; DM: diabetes mellitus; EBCT: electron-beam computed tomography; ECG: electrocardiography; HTN: hypertension; LoE: level of evidence; MDCT: multi-detector computed tomography; MI: myocardial infarction; NR: not reported.

Appendix I . Peer Reviewers

The individuals listed below have agreed to provide clinical and/or peer review.

This role should not be construed to mean that the individuals were authors or contributors to the formulation of the draft, nor does it imply endorsement, approval, or disapproval of the process or report.

Individual	Expertise/Experience
Noel S. Weiss, MD, Dr.P.H. Professor, Department of Epidemiology, University of Washington Member, Fred Hutchinson Cancer Research Center, Seattle	• MD, Stanford University • DrPH and MPH, Harvard School of Public Health (Epidemiology and Biostatistics) • Over 40 years of research in epidemiology with specific expertise in clinical epidemiology and research methods • Member, Institute of Medicine • Over 500 publications, including books on clinical epidemiology • Recipient of Outstanding Investigator Award (NCI) and Abraham Lilienfeld Award as well as awards for outstanding teaching
Rita F. Redberg, MD, MSc UCSF Division of Cardiology, Professor of Clinical Medicine	• MD, University of Pennsylvania; • ABIM-Internal medicine and Cardiovascular specialty • MSc, London School of Economics • Over 20 years of research-related and clinical experience. • Research areas include cardiovascular disease in women, cardiovascular imaging, health policy and technology assessment, evidence based-practice • Reviewer/consultation for AHRQ, USPSTF, MCAC, CDRH
Ann Derleth, PhD, MSPH Health Services Researcher, Health Economics Research Associate- University of Washington	• PhD in Health Services from University of Washington with focus on quantitative methods for outcomes and economic analysis • MSPH in biostatistics from University of Washington with specialization in statistical methods for health services research • Over 15 years experience in health services-related research and biostatistics including outcomes measures, disease severity and risk adjustment • Work with administrative data and reimbursement policy

Appendix J. Overview of Diagnostic Test Validation and Reliability

Evaluation of validity and reliability studies

The accuracy of a diagnostic test consists of two general components: the accuracy of classifying patients with respect to their disease status (validity), and the degree to which repeated measures yield the same results (reliability). However, regardless of how accurate or predictive a test may be, health policy and public health perspectives assert that a diagnostic test should only be performed if it leads to the use of interventions that, on average, are likely to improve patient outcomes or it prevents the use of interventions that are not likely to improve outcomes. [1]

Validity and test accuracy

Validation of a measure refers to comparison of that measure against the true value. Validity is the degree to which a test *accurately* measures what it is intended to measure. Technically, an error free comparison method (i.e., true gold standard) is required in order to directly measure validity. For diagnostic tests, evaluation of the test against the "truth" allows the determination of how accurately the test classifies patients with and without disease. The accuracy of classification can be expressed by first accounting for the results as described in the following 2 x 2 table:

- True Positive (TP) results (cell a) = number of individuals with a disease who test positive
- False Positive (FP) results (cell b) = number of individuals without a disease who test positive
- False Negative(FN) results (cell c) = number of individuals with a disease who test negative
- True Negative (TN) results (cell d) = number of individuals without a disease who test negative

		True Classification	
		Disease present (+)	**Disease absent (-)**
Diagnostic Test	**Disease present(+)**	a = TP	b = FP
	Disease absent(-)	c = FN	d = TN
		a + c	b + d

The number of patients who truly have the disease is given by a + c and the number who truly do not have disease is given by b + d.

A true "gold standard" should be the definitive "truth" about the presence/absence of a condition or disease. Since an error-free method is not always available, a comparison of a diagnostic test to an appropriate reference standard, which may not be error-free, is commonly done. This referent method which is not always error free, may be better termed inter-method reliability. [2] An *appropriate* reference standard should be able to correctly classify patients with respect to the presence and absence of disease and be

reproducible. However, variability in the test influences the ability to correctly classify patients according to disease status.

Sensitivity and specificity are the traditional measures of diagnostic tests used in validation to describe the accuracy of classification. They do not, however, describe the probability that a patient actually has the disease if the test is positive or does not have it if the test is negative.

Term	Definition	Calculation
Sensitivity =	% of patients with the disease who test positive =	a/(a + c) x 100
Specificity =	% of patents who do NOT have disease who test negative =	d/(b + d) x 100

The sensitivity and specificity are not fixed properties of a test. Instead, they reflect how the test performs among those with and without disease in a given population when administered in a specific manner. Sensitivity and specificity may appear to vary across populations, but do not directly depend on the prevalence of the condition.[3] Sensitivity and specificity form the basis of a receiver operating characteristic (ROC) curve which plots the relationship between the proportion of true positives (sensitivity) and the proportion of false positives (1-specificity) as a function of the diagnostic cut-off level for a disease.

When a true gold standard or appropriate reference standard is used, *and* the study population has a frequency of disease that approximates the frequency of disease in the population to which the results are to be applied (or the frequency of the disease in the population to which the test is to be applied is known), two additional measures of test accuracy can be used. These are the predictive value of a positive test (PPV) and the predictive value of a negative test (NPV) and are described as follows:

Term	Definition	Calculation
PPV =	% of patients with a positive test who have the disease =	a/(a + b) x 100
NPV =	% of patents with a negative test who do NOT have the disease =	d/(c + d) x 100

The PPV and NPV estimates are only accurate and meaningful if the actual proportion of true positives in the relevant population is represented by (a + c)/n. In other words, the actual prevalence of disease in the relevant population must be accurately estimated by the study population or it must be known for the population that is to be tested; otherwise, the predictive values are misleading.[1, 3] If the test is done in a population with a very low frequency of disease, for example, the PPV is quite low, even if the sensitivity and specificity are high.

Like PPV and NPV, most estimates of "overall accuracy" as an estimate of test validity vary with the prevalence of the disease or condition and can frequently lead to a distorted impression of a test's accuracy and validity.[1, 4] In addition, such measures do not fit into the decision making process as do PPV and PVN.[1] For these reasons, "overall accuracy" is to be avoided.

Other measures of test performance include positive and negative likelihood ratios and the area under the receiver operator (ROC) curve. These measures are based on sensitivity and specificity and do not vary with disease prevalence even though they may vary across populations.[4]
Likelihood ratios (LR) are clinically useful and can be used to consider which test may be better for ruling in or ruling out a disease. LR can also be valuable for comparing the accuracy of several tests to a gold standard. The LR is the ratio of the probability of a given test result in those with disease to the probability of that test result in people without the disease.

The likelihood ratio of a positive test (LRP) provides information about how well a positive test performs when a disease or condition is present compared with when disease is absent. The LRP describes how much the odds of disease *increase* when the test is positive. The likelihood ratio of a negative test provides information about how much the odds of disease *decrease* when the test is negative.

Term	Description	Calculation
LRP =	how much odds of disease *increase* with *positive* test =	sensitivity/(1-specificity)
LRN =	how much odds of disease *decrease* with *negative* test =	(1-sensitivity)/specificity

Likelihood ratios (LR) are combined with the pre-test odds of disease to determine the post-test odds of disease. The pre-test disease odds are based on disease prevalence, the patient population and individual patient characteristics. (The odds of disease can be determined from the probability of disease using Bayes' Theorum). The pretest odds are equal to the probability of having the disease divided by the probability of not having it.

LRs provide insight into the extent to which doing the test is worthwhile in changing the odds of disease given the pre-test odds. The post-test odds, which represents the chance that the patient has the disease, thus incorporate disease prevalence, patient population information and patient-specific risk information via the pre-test odds as well as test performance information via the likelihood ratio as follows:

Post-test odds = Pre-test odds X likelihood ratio

If the test does not change the post-test odds of disease (e.g., a LR of 1), it is not likely to be helpful for ruling in (raising the post-test odds) or ruling out (lowering the post-test odds) disease. A test with a *high LR* is best to *rule in* a disease or condition while a test with a *low LR* is best to *rule out* a disease or condition.

While LR do not rely on disease prevalence, they are based on sensitivity and specificity of the test and therefore reflect how the test performs among those with and without disease in a given population when administered in a specific manner.

In the absence of validation studies, the concordance (percent agreement) was determined. Since this calculation does not take into account agreement that may be

expected purely by chance, the kappa statistic (κ) was calculated where there were adequate data to correct for chance agreement according to the following formula:[3]

$$\kappa = (P_o - P_e) / (1 - P_e)$$

P_o is the observed concordance = (a+d)/N

P_e is the concordance expected by chance based on row and column totals:

$$P_e = \left[\frac{(a+b)(a+c)}{N} + \frac{(b+d)(c+d)}{N} \right] \Big/ N$$

Kappa describes the amount by which the observed agreement exceeds what would be expected by chance alone. While it can assist in putting results in perspective, there are several caveats that must be borne in mind. First, it is partly dependent on the true prevalence of the disease or characteristic in the population and declines as prevalence approaches 0 or 1.[5] Thus, it should not be viewed as a consistent property of the test comparison. In addition, although kappa is often used to adjust for the role of chance in studies that compare different methods (i.e. inter-method reliability studies), this is not the original intent for the application of the kappa statistic. It is most appropriately used in intra-method reliability studies described below. Guidelines for interpretation of kappa are provided by Landis and Koch. [6]

Reliability

The accuracy of a test also depends on its reliability. The purpose of reliability studies is to evaluate the reproducibility of a measure. That is, how well a measure can be replicated on, for example, a given patient, or imaging film, etc. under the same conditions. Even though a measure may be reproducible, it may still not be valid.

There are two general types of reliability studies:

Intra-method Reliability

- Test-retest reliability refers to the agreement when a test is done with the same instrument on the same subjects at two or more different times. Intra-rater reliability is test-retest reliability. This gives the upper limit of the extent to which the measure correlates with the truth or, $\rho_{XT,}$ where ρ is the correlation between the measure, X, and the truth, T.
- Inter-rater reliability refers to the agreement between two or more raters using the same instrument on the same subjects.

Inter-method reliability

- Refers to the agreement between two different instruments measuring the same underlying factor to yield similar results on the same subjects. Some refer to this as "validity" but technically, a "perfect" comparison is needed to

determine validity. In certain circumstances, inter-method reliability studies can provide some information about the validity of a measure.[2]

Analysis of Reliability Studies

The following is an overview of common and appropriate statistical methods for reliability studies. There are two basic factors to consider when determining the appropriate analysis or statistical method: the type of study (i.e., intra-method or inter-method), and the type of variable (e.g., categorical). Additional information is found in Armstrong, White and Saracci.[2]

For dichotomous or categorical measures, in an intra-method study, percent agreement, and kappa are appropriate. For inter-method or validity studies where the categorical measure is nominal, a misclassification matrix is appropriate and for dichotomous variables (assuming reasonable ability to measure the true status), sensitivity and specificity can be determined. While kappa is sometimes used for these types of studies, it may not be an appropriate use of kappa. Continuous measures in an inter-method study are evaluated using Pearson's correlation coefficient.

Ordered categorical variables used in inter-method reliability or validity studies may be evaluated by the following methods:

- Misclassification matrix
- Spearman Rank correlation coefficient
- Pearson product moment correlation coefficient
- Either Spearman or Pearson on underlying variable from which variable was created

Continuous variables in intra-method reliability studies are generally evaluated using intra-class correlation coefficients. Cohen's kappa is used for evaluation of nominal or binomial variables while weighted kappa is most appropriate for ordered categorical variable.

Appendix References

1. Weiss NS. *Clinical Epidemiology : The Study of the Outcome of Illness*. 3rd ed. ed. New York :: Oxford University Press; 2006.

2. Armstrong B, White E, Saracci R. *Principles of Exposure Measurement in Epidemiology*. Oxford University Press 1992.

3. Koepsell T, Weiss N. *Epidemiologic Methods: Studying the Occurrence of Illness*. New York: Oxford University Press, Inc; 2003.

4. Alberg AJ, Park JW, Hager BW, Brock MV, Diener-West M. The use of "overall accuracy" to evaluate the validity of screening or diagnostic tests. *J Gen Intern Med.* May 2004;19(5 Pt 1):460-465.

5. Weiss NS, Koepsell, T.D. *Epidemiologic Methods: Studying the Occurrence of Illness*. 1st Edition ed. Oxford: Oxford University Press; 2003.

6. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* Mar 1977;33(1):159-174.