

## **TUSM.MMC Evidence Based Medicine Course Outline**

### **Course Director:**

Kathleen Fairfield, MD, DrPH

[fairfk@mmc.org](mailto:fairfk@mmc.org) 661-7614

This course is meant to closely parallel the structure and content of the same course delivered to the Boston TUSM students. Course materials were developed by Allen F. Shaughnessy, PharmD, MMedEd, Professor of Family Medicine at Tufts. The study guides, individual readiness assessment tests, and final exam are identical. The in class materials and structure are unique to each site.

**Course Dates for 2013:** 5/2, 5/8, 5/16, 5/22

All classes held at the Brighton Campus

### **Course Objectives**

#### ***Overall aims of the course***

- To reinforce the biostatistics and epidemiology knowledge components, already taught, required for the USMLE Step 1 Examination.
- To demonstrate and contextualize how key concepts of biostatistics, epidemiology, and evidence-based medicine are used in the medical literature and applied to patient care.
- To help students develop the ability to analyze, synthesize, and apply knowledge.
- To prepare students to evaluate the medical literature and apply evidence to patient care in subsequent medical school courses and clerkships.
- To encourage an inquisitive, questioning attitude toward medical knowledge and medical care.
- To further inform your understanding of public health issues you will encounter in Maine.

Specific objectives are listed for each session.

### **Course Materials**

All the materials needed for this course can be found in the Maine Track section of TUSK. For each session, you will find an article to review as well a Study Guide, which is designed to be viewed electronically, with embedded links to additional readings, video, and exercises. There is not an assigned textbook. There are no powerpoint slides. Further readings, can be found at the [Center for Information Mastery](#) website.

## Class Structure:

Students will be expected to read the article for each session as well as the online Study Guides created by Dr. Allen Shaughnessy. Students are also expected to complete an Individual Readiness Assessment Test (IRAT) for each session, to be submitted on TUSK no later than the start of class.

2 hour session:

15 minutes: gathering and IRAT review

30 minutes: didactic: key concepts and link to clinical implications

45 minutes: small group critical appraisal of the article with worksheet; student lead, floating facilitators (3 for 6 groups)

30 minutes: large group discussion of content, clinical take away

## Introduction to the Course

1) **The assigned preparation *must be completed before each session*.** The study guides contains an explanation of the concepts with hyperlinks to additional readings and videos that may be helpful. They can be printed out but are designed to be read on a computer or tablet screen. If you understand the concepts by reading only the material in the study guide, there is no need to follow the hyperlinks (*i.e., there is no “gotcha” information buried in one of the hyperlinked resources*).

2) **The Individual Readiness Assessment Tests (IRATs) *must be completed on your own before the start of class*.** There will be one IRAT for each class period ( $n = 4$ ). The IRATs will be opened for completion immediately after the previous class ends and will remain open until immediately before the class starts. This step assures me, and your teammates, that you come to the class prepared. It also allows me to see where the group, as a whole, has had difficulty. These are individual (no group work) and comprise 20 percent of your grade.

Following class, the IRATs will be available again, this time with the answers provided, for your use to review before the final exam.

3) ***Class attendance is mandatory.*** Class time will be used to work in groups to complete critical appraisal of the assigned article, followed by group discussion about the clinical implications. *Classes will not be videorecorded or streamed.*

4) **There is no need for computers during class, and group work in class will not use computers.** The questions and exercises will be answered using the combined knowledge and wisdom of the group.

5) **Homework:** Your homework is to critically assess an original research study, practice guideline, or review article that you identify on your own. You will use one of the worksheets associated with the classes, determine the article's validity, come to a conclusion regarding a clinical question, and consider how you would explain the findings to a patient and their family. See Appendix 2.

## Grading

	Percent of Grade
<b>Individual Readiness Assessment Tests</b>	<b>30</b>
<b>Group Participation</b>	<b>30</b>
<b>Final Exam</b>	<b>30</b>
<b>Homework</b>	<b>10</b>

*Your participation in this group work* will be graded by your team members (see Appendix 1 at the end of this document). At the END of the course each person will divide up 50 points among the team members, awarding more points to team members who came more prepared and who participated to a greater extent, and awarding fewer points to those team members who didn't. These points will be converted into a proportion of the total available points awarded to you (to account for teams of different sizes).

Example: You are in a group of 6 students. Each student has 50 points to be distributed among the group. Let's say you receive the following points from your team members:

Student A.	10 points
Student B.	15 points
Student C.	7 points
Student D.	10 points
Student E.	9 points
Total	51 points

Since there are 5 other people in the group, the average score would be 50 (10 from each member to the group, other than yourself). Your net points =  $51 - 40 = 1$  points.

## Appendix 1.

### Peer Evaluation Form

Peer Evaluation      Name \_\_\_\_\_      Team # \_\_\_\_\_

Please assign scores that reflect how you really feel about the extent to which the other members of your team contributed to your learning and/or your team's performance. This will be your only opportunity to reward the members of your team who worked hard on your behalf. **(Note: If you give everyone pretty much the same score you will be hurting those who did the most and helping those who did the least.)**

**Preparation-** Were they prepared when they came to class?

**Contribution-** Did they contribute productively to group discussion and work?

**Respect for others' ideas-** Did they encourage others to contribute their ideas?

**Flexibility-** Were they flexible when disagreements occurred?

**Instructions:** In the space below please rate each of the other members of your team. Each member's peer evaluation score will be the average of the points they receive from the other members of the team. To complete the evaluation you should: 1.) List the name of each member of your team in the alphabetical order of their last names and 2.) assign an average of 10 points to the other members of your team (Thus, for example, you should assign a total of 50 points in a six-member team.) and, 3.) Differentiate some in your ratings; for example, you must give at least one score of 11 or higher (maximum = 15) and one score of 9 or lower.

Team Members	Score
1. _____	_____
2. _____	_____
3. _____	_____
4. _____	_____
5. _____	_____
6. _____	_____
7. _____	_____

**Total** \_\_\_\_\_

Appendix 2.

### **Homework Assignment**

**Who:** Each person will submit the assignment. This is not a group assignment.

**What:** You will critically appraise a research paper, practice guideline, or review article of your choice. You may want to choose a paper that is related to your small group mentoring work or other clinical question you encounter this month.

**When:** The homework should be submitted no later than the end of the course.

**Where:** E-mail a copy of the paper (as a pdf) and the evaluation form (in Word) to Kathleen at [fairfk@mmc.org](mailto:fairfk@mmc.org)

**How:** For your identified article, answer the questions on the appropriate worksheet (see study guides for each type of article). At the end of the questions, answer the following question:

How would you explain the findings of this article to a patient and their family? This answer should be two or three sentences.

Maine On-Line IRAT Test Schedule		Class date	Quiz open	Quiz closed	Quiz with answers posted
1	Therapy	2-May	26-Apr	May 2, 9 A	May 2, 11 A
2	Diagnosis	8-May	2-May	May 8, 10 A	May 8 12 N
3	Review	16-May	8-May	May 16, 8:30 A	May 16, 10:30
4	Guidelines	22-May	16-May	May 22, 12:30	May 22, 2:30

# Preventing Obesity Among Adolescent Girls

## *One-Year Outcomes of the Nutrition and Enjoyable Activity for Teen Girls (NEAT Girls) Cluster Randomized Controlled Trial*

David R. Lubans, PhD; Philip J. Morgan, PhD; Anthony D. Okely, EdD; Deborah Dewar, BEd; Clare E. Collins, PhD; Marijka Batterham, PhD; Robin Callister, PhD; Ronald C. Plotnikoff, PhD

**Objective:** To evaluate the impact of a 12-month multicomponent school-based obesity prevention program, Nutrition and Enjoyable Activity for Teen Girls among adolescent girls.

**Design:** Group randomized controlled trial with 12-month follow-up.

**Setting:** Twelve secondary schools in low-income communities in the Hunter and Central Coast regions of New South Wales, Australia.

**Participants:** Three hundred fifty-seven adolescent girls aged 12 to 14 years.

**Intervention:** A multicomponent school-based intervention program tailored for adolescent girls. The intervention was based on social cognitive theory and included teacher professional development, enhanced school sport sessions, interactive seminars, nutrition workshops, lunch-time physical activity sessions, handbooks and pedometers for self-monitoring, parent newsletters, and text messaging for social support.

**Main Outcome Measures:** Body mass index (BMI, calculated as weight in kilograms divided by height in me-

ters squared), BMI *z* score, body fat percentage, physical activity, screen time, dietary intake, and self-esteem.

**Results:** After 12 months, changes in BMI (adjusted mean difference,  $-0.19$ ; 95% CI,  $-0.70$  to  $0.33$ ), BMI *z* score (mean,  $-0.08$ ; 95% CI,  $-0.20$  to  $0.04$ ), and body fat percentage (mean,  $-1.09$ ; 95% CI,  $-2.88$  to  $0.70$ ) were in favor of the intervention, but they were not statistically different from those in the control group. Changes in screen time were statistically significant (mean,  $-30.67$  min/d; 95% CI,  $-62.43$  to  $-1.06$ ), but there were no group by time effects for physical activity, dietary behavior, or self-esteem.

**Conclusions:** A school-based intervention tailored for adolescent girls from schools located in low-income communities did not significantly reduce BMI gain. However, changes in body composition were of a magnitude similar to previous studies and may be associated with clinically important health outcomes.

**Trial Registration:** anzctr.org.au Identifier: 12610000330044

*Arch Pediatr Adolesc Med.* 2012;166(9):821-827.  
Published online May 7, 2012.  
doi:10.1001/archpediatrics.2012.41

**Author Affiliations:** Schools of Education (Drs Lubans, Morgan, and Plotnikoff and Mrs Dewar), Health Sciences (Dr Collins), and Biomedical Sciences and Pharmacy (Dr Callister), Priority Research Centre in Physical Activity and Nutrition, University of Newcastle, Newcastle; and Interdisciplinary Educational Research Institute (Dr Okely), and Centre for Statistical and Survey Methodology, (Dr Batterham), University of Wollongong, Wollongong, Australia.

OBESITY PREVENTION IS A global health priority<sup>1</sup> because pediatric weight status is associated with a range of adverse health outcomes<sup>2</sup> and obese youth are at an elevated risk for obesity in adulthood.<sup>3</sup> The prevalence of child and adolescent obesity has increased considerably during the past 30 years and current estimates suggest that approximately a quarter of youth in developed nations are overweight or obese.<sup>4,5</sup> Although there is evidence to suggest that levels of obesity have plateaued in recent years,<sup>6</sup> this trend has not been observed among youth living in low-income communities.<sup>7,8</sup>

Schools have been identified as important institutions for the promotion of healthy lifestyles<sup>9</sup> and provide access to populations at risk for obesity, such as adolescents living in low-income communities. Although evidence for the long-term effects of school-based obesity prevention programs is limited,<sup>10</sup> recent high-quality studies have demonstrated that these interventions can prevent unhealthy weight gain in youth.<sup>11-13</sup> Multicomponent school-based interventions targeting groups at risk for obesity can be effective, but further testing in long-term rigorously designed studies is needed.<sup>9,14</sup>

The importance of designing and implementing obesity prevention programs for

preadolescent and adolescent girls living in low-income communities has emerged in the literature.<sup>15-17</sup> The physical activity decline associated with adolescence is steeper among girls<sup>18</sup> and unhealthy weight gain is often observed in this cohort.<sup>19,20</sup> The aim of the current study is to evaluate the effects of the Nutrition and Enjoyable Activity for Teen Girls (NEAT Girls) program,<sup>21</sup> a 12-month school-based group randomized controlled trial designed to prevent unhealthy weight gain in adolescent girls living in low-income communities. This article reports the 12-month intervention effects.

## METHODS

### STUDY DESIGN AND PARTICIPANTS

Ethics approval for the study was obtained from the relevant university and school board human ethics committees. School principals, parents, and study participants provided written informed consent. The design, methods, and characteristics of participants at baseline have been reported in detail elsewhere.<sup>22</sup> In summary, NEAT Girls was a group randomized controlled trial, and the design, conduct, and reporting of the trial adhere to Consolidated Standards of Reporting Trials guidelines.<sup>23</sup> Baseline assessments were conducted in May through June 2010 and 12-month (immediate posttest) assessments were completed in May through June 2011.

The intervention was designed for adolescents from schools located in low-income communities, and the Socio-Economic Indexes for Areas of relative socioeconomic disadvantage were used to identify eligible secondary schools. The Socio-Economic Indexes for Areas (scale, 1=lowest to 10=highest) summarize the characteristics of people and households within an area. State-funded government secondary schools located in the Hunter and Central Coast areas in New South Wales, Australia, with a Socio-Economic Indexes for Areas score of 5 or less (bottom 50%) were considered eligible for inclusion. Eighteen schools in the Central Coast and Hunter regions met our eligibility criteria and all of these schools were invited to participate. Twelve secondary schools were recruited and eligible study participants were adolescent girls in grade 8 (second year of secondary school).

### SAMPLE SIZE CALCULATION AND RANDOMIZATION

The sample size calculation was based on change in body mass index (BMI, calculated as weight in kilograms divided by height in meters squared).<sup>24</sup> Assuming an  $\alpha$  of 0.05, power of 80%, and a 20% dropout rate, we calculated that we would require 30 participants from each of the 12 schools to detect a between-group difference of 1 BMI unit<sup>25</sup> using a BMI standard deviation of 1.5,<sup>12</sup> and an intraclass correlation coefficient of 0.01.<sup>26</sup> Following baseline assessments, the 12 schools were matched (ie, 6 pairs of schools) based on their geographic location, size, and demographics.<sup>27</sup> An independent researcher then randomized each pair to either the NEAT Girls intervention or control groups.

### INTERVENTION

The NEAT Girls intervention was informed by the Program X pilot study<sup>28,29</sup> and a detailed description of the intervention has been reported previously.<sup>22</sup> The intervention was guided by Bandura's social cognitive theory<sup>30</sup> and targeted evidence-based psychological (ie, self-efficacy, outcome expectations, and

outcome expectancies), behavioral (ie, goal setting and self-monitoring), and environmental (ie, teacher, family, and peer support) influences on physical activity and nutrition behavior change.<sup>31,32</sup> The intervention included the following components: enhanced school sport sessions, interactive seminars, nutrition workshops, lunch-time physical activity sessions, handbooks and pedometers for self-monitoring, parent newsletters, and text messaging for social support. To facilitate the implementation of the NEAT Girls program, school champions (ie, teachers responsible for the delivery of the program) from the intervention schools attended a 1-day training workshop at the local university. The intervention was focused on promoting lifetime physical activities, reducing sedentary behaviors, and encouraging low-cost healthy eating, and it was delivered during 4 school terms (ie, 12 months) at no additional financial cost to the school or students. All intervention schools were provided with a standard equipment pack (value=US \$1300), which consisted of a range of equipment (eg, elastic tubing resistance training devices, fitness balls, and yoga and Pilates resources) designed to support the promotion of lifetime physical activities.

NEAT Girls was based on well-defined messages designed to promote physical activity and healthy eating and reduce sedentary behavior,<sup>22</sup> which were reinforced using the intervention components. The enhanced school sport sessions (60-80 minutes) were delivered by teachers and involved a range of activities organized into 4-week units. For the first school term, the enhanced school sport sessions included an information component (10-15 minutes) delivered by teachers from the study schools. Members of the research team delivered 3 interactive seminars that focused on the benefits of physical activity and healthy eating as well as key behavioral messages. Participants were provided pedometers<sup>33</sup> and handbooks and were encouraged to use these resources to monitor their lifestyle physical activity participation.

Three practical nutrition workshops were delivered in the study schools by accredited practicing dietitians. The sessions were designed to provide students with the confidence to select, prepare, and consume healthy low-cost foods. Parents of participants were sent study newsletters at 4 periods during the 12-month intervention. The first newsletter reported their children's time spent in physical activity, sedentary behaviors, and self-reported fruit and vegetable consumption. All of the newsletters included information to increase awareness and encourage parents to support their children's physical activity and dietary behaviors. To reinforce the targeted behaviors, the girls were sent text messages weekly during the second and third terms and bi-weekly during the fourth term of the program's delivery (eg, "Sitting down for long periods of time is bad for you, but what makes it worse is that people often eat junk while sitting down in front of the TV. Try to avoid eating dinner while watching TV").

To assist in the recruitment of schools and prevent resentful demoralization or compensatory rivalry,<sup>27</sup> the control group was provided with equipment packs and a condensed version of the intervention following the completion of 24-month assessments.

### ASSESSMENTS AND MEASURES

Data collection took place in the study schools and was conducted by trained research assistants blinded to group allocation at baseline only.

#### Primary Outcome Measures

Body mass index was the primary outcome.<sup>2</sup> Weight was measured in light clothing without shoes using a portable digital



scale (Model No. UC-321PC; A&D Company Ltd) and height was measured using a portable stadiometer (Model No. PE087; Mentone Educational Centre). Body mass index weight categories were based on BMI *z* scores, which were calculated using the LMS method.<sup>34</sup> Body fat percentage was determined using the Imp SFB7 bioelectrical impedance analyzer.<sup>35</sup>

## Secondary Outcome Measures

The 90° push-up and the prone support tests<sup>36</sup> were used to provide measures of upper body muscular endurance and core abdominal isometric muscular endurance, respectively. Participants wore Actigraph accelerometers (MTI models 7164, GT1M, and GT3X)<sup>37</sup> for 7 consecutive days. Trained research assistants fitted the monitors and explained the monitoring procedures to participants.<sup>38</sup> Participant data were included in the analyses if accelerometers were worn for 600 minutes or more on 4 days or more (including 1 weekend day)<sup>39</sup> and age- and sex-specific cut points were used to categorize activity intensity.<sup>40</sup> Dietary intake was assessed using the previously validated Australian Eating Survey food frequency questionnaire and total energy (ie, total kilocalories per day and total kilocalories per kilogram per day) was presented as a summary variable to represent dietary intake.<sup>41</sup> The Adolescent Sedentary Activity Questionnaire was used to provide a self-report of screen time (ie, watching television/videos/DVDs and using computers and electronic communication).<sup>42</sup> Participants completed selected scales from Marsh's Physical Self-description Questionnaire (ie, perceived body fat, physical self-esteem, and global self-esteem).<sup>43</sup>

## Process Evaluation

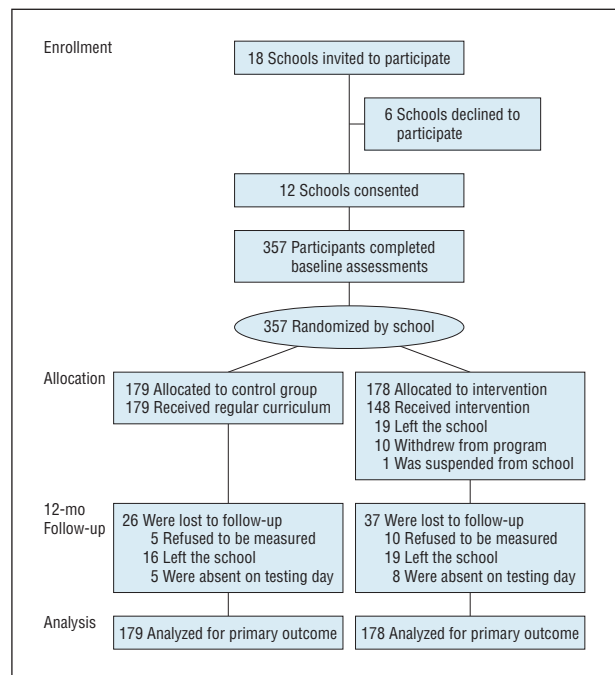
A detailed process evaluation was conducted and included attendance/reach (ie, attendance at enhanced school sport sessions, lunch-time physical activities and nutrition workshops, and percentage of students who provided postal addresses and mobile phone numbers and were sent all 4 newsletters and 58 text messages), intervention fidelity (ie, 24 randomly selected sessions were observed by a member of the research team), and program satisfaction (ie, girls completed detailed process evaluation questionnaires at the completion of the study). Although the enhanced school sport sessions were designed to be flexible in delivery, the fidelity of each session was assessed using the following criteria (yes = 1, no = 0): (1) Was there 60% or greater student attendance at the session? (2) Was the session delivered by the school champion? (3) Did the school champion deliver the session using the program handbook? (4) Did the session follow the basic structure outlined in the handbook?

## STATISTICAL ANALYSES

Differences between groups at baseline were examined using chi squares and independent sample *t* tests in PASW Statistics 17 (SPSS Inc) software and  $\alpha$  levels were set at  $P < .05$ . Statistical analyses followed the intention-to-treat principle and were conducted using mixed models, which have the advantage of being robust to the biases of missing data.<sup>44</sup> The models were specified to adjust for the clustered nature of the data and the analysis conducted using established models.<sup>27</sup> The mixed models were analyzed using the PROC MIXED statement in SAS version 9.1 (SAS Institute Inc).

## RESULTS

School and participant recruitment, enrollment, and flow are provided in the **Figure**. Twelve schools were re-



**Figure.** Flowchart of participants throughout the study.

cruited and 357 participants were assessed at baseline, representing 99.2% of the targeted sample size (**Table 1**). There were no statistically significant differences between intervention and control groups for any of the outcomes at baseline. Sixty-three girls were unavailable for 12-month assessments; 153 (85.5%) and 141 (79.2%) girls were retained in the control and intervention groups, respectively. The girls who dropped out of the study had higher baseline BMI (mean [SD], 23.81 [4.52] vs 22.39 [4.56];  $P = .03$ ) and BMI *z* score (mean [SD], 1.11 [1.06] vs 0.73 [1.15];  $P = .02$ ) values than those who completed the study.

## PRIMARY AND SECONDARY OUTCOMES

Outcomes are reported in **Table 2**. Changes in body composition were all in favor of the intervention group, but there were no statistically significant between-group differences in BMI (primary outcome), BMI *z* score, or body fat percentage. Girls in the intervention group reported significantly less screen time than girls in the control group (mean, -30.67 min/d; 95% CI, -62.43 to -1.06). Compliance with our accelerometer monitoring was poor (ie, a mean [SD] of 191 [53.5%] and 89 [24.9%] participants wore accelerometers for 600 minutes or more on 4 or more days including a weekend day at baseline and posttest, respectively) and there were no differences between groups on any of the physical activity outcomes. Muscular fitness, dietary intake, physical self-perceptions, and self-esteem remained relatively stable during the study period for both intervention and control girls with no differences between groups.

## INTERVENTION IMPLEMENTATION AND PROCESS OUTCOMES

A total of 148 girls received the intervention (83.1%). Students' mean (SD) attendance at school sport sessions was

**Table 1. Characteristics of Study Sample**

Characteristics	Control (n=179)	NEAT Girls (n=178)	Total (N=357)
Age, mean (SD), y	13.20 (0.45)	13.15 (0.44)	13.18 (0.45)
Participants born in Australia, No. (%)	174 (97.2)	175 (98.3)	349 (97.8)
English language spoken at home, No. (%)	176 (98.3)	176 (98.9)	352 (98.6)
Cultural background, No. (%) <sup>a</sup>			
Australian	153 (85.5)	152 (85.4)	305 (85.4)
Asian	1 (0.6)	3 (1.7)	4 (1.1)
European	18 (10.1)	18 (10.1)	36 (10.1)
Other	7 (4.0)	4 (2.2)	11 (3.1)
Socioeconomic position, No. (%) <sup>b</sup>			
1-2	47 (26.4)	28 (15.8)	75 (21.1)
3-4	28 (15.7)	59 (33.1)	87 (24.5)
5-6	96 (53.6)	87 (49.2)	183 (51.3)
7-8	6 (3.4)	3 (1.7)	9 (2.5)
9-10	1 (0.6)	0	1 (0.3)
Weight, mean (SD), kg	58.37 (13.78)	58.41 (14.15)	58.39 (13.95)
Height, mean (SD), m	1.61 (0.07)	1.60 (0.06)	1.60 (0.07)
BMI, mean (SD)	22.59 (4.49)	22.70 (4.68)	22.64 (4.58)
BMI z score, mean (SD) <sup>c</sup>	0.78 (1.17)	0.82 (1.12)	0.80 (1.14)
BMI category, No. (%) <sup>c</sup>			
Underweight	1 (0.6)	1 (0.6)	2 (0.6)
Healthy weight	99 (55.3)	103 (57.9)	202 (56.6)
Overweight	50 (27.9)	43 (24.2)	93 (26.1)
Obese	29 (16.2)	31 (17.4)	60 (16.8)

Abbreviation: BMI, body mass index (calculated as weight in kilograms divided by height in meters squared).

<sup>a</sup>One participant did not report her cultural background.

<sup>b</sup>Socioeconomic position by population decile using Socio-Economic Indexes for Areas of relative socioeconomic advantage and disadvantage based on home post code. 1 is the lowest and 10 the highest. Two participants did not report home post code.

<sup>c</sup>BMI z score and categories based on LMS method.

60.6% (26.0%). On average, girls attended 65.0% (25.1%) of the nutrition workshops, 24.6% (28.1%) of the optional lunch-time sessions, and completed 8.8% (25.7%) of the home physical activity and nutrition challenges. Intervention delivery fidelity was found to be 74.0%. All 4 of the parental newsletters were sent to valid addresses for 74.5% of girls in the intervention group. A total of 58 text messages were sent to 91% of girls in the intervention group. Overall, girls were satisfied with the program (mean [SD], 3.52 [1.24]; rating scale, 1=strongly disagree to 5=strongly agree). The enhanced school sport sessions (41.7%) and the nutrition workshops (38.7%) were the 2 intervention components enjoyed most by girls. No injuries or adverse effects were reported during the activity sessions or assessments.

## COMMENT

NEAT Girls is a multicomponent school-based obesity prevention program targeting adolescent girls from secondary schools located in low-income communities. The intervention effects on body composition were small and not statistically significant but have potential clinical importance. Girls in the intervention group spent 30 minutes per day less in screen-based activities than their con-

trol group peers. High levels of screen time are associated with a range of adverse health consequences,<sup>45</sup> and our findings have important implications that may help address the increasing burden of pediatric and adolescent obesity observed in areas of social and economic disadvantage.

Behaviors, attitudes, and physical morbidity that develop during adolescence have profound implications for current and future health,<sup>46</sup> yet surprisingly, few adolescent obesity prevention interventions have been designed and evaluated. The challenges of working with adolescents<sup>46</sup> may explain both the small number of studies and their modest results. Small differences can be meaningful at the population level, and the favorable changes in BMI z score (mean, -0.08; 95% CI, -0.20 to 0.04) and body fat percentage (mean, -1.09; 95% CI, -2.88 to 0.70) observed in our study may have both clinical significance and important public health implications. A recent longitudinal study<sup>47</sup> found that a 1% increase in body fat percentage was related to increases of 1.042 mg/dL and 0.621 mg/dL (to convert to millimoles per liter, multiply by 0.0259) in total cholesterol in boys and girls, respectively. Similarly, the school-based diabetes risk reduction intervention, known as the HEALTHY study, resulted in a small but statistically significant reduction in BMI z score (ie, -0.05), which was accompanied by smaller increases in fasting insulin levels. Increases in body fat during youth are consistently associated with adverse changes in plasma lipids<sup>47,48</sup> and further examination of the health implications of weight gain during this period will help to determine the clinical importance of intervention effects.

A number of recent obesity prevention interventions targeting adolescent and preadolescent girls have been evaluated in school and community settings. The New Moves intervention was similar in size and intervention design to the NEAT Girls program, but improvements in body composition were half the magnitude to those observed in our study (adjusted differences in BMI and body fat percentage were -0.10 and -0.46, respectively). The Stanford and Memphis GEMS interventions<sup>15,17</sup> were 2 well-designed obesity prevention interventions targeting unhealthy weight gain in preadolescent girls from low-income communities. The interventions resulted in positive changes in secondary outcomes (eg, reduced fasting total cholesterol levels and depressive symptoms), but there were no treatment effects for BMI. Although both schools and community settings offer promise for the prevention of obesity in youth, more work is needed to translate the strong effects typically observed in small-scale efficacy studies to large-scale effectiveness trials.

Girls in the intervention group did not increase their physical activity, but significant differences in screen time were observed during the study period. The large reductions in self-reported screen time represent one-quarter of participants' daily limit and such changes have important health implications. Young people spend 2 to 4 hours per day in screen-based recreation and 5 to 10 hours per day sedentary, both of which are associated with a range of adverse health consequences.<sup>45</sup> Targeting time spent in sedentary behavior has emerged as an effective

**Table 2. Changes in Primary and Secondary Outcomes Measures and Group Differences**

Measure	Baseline, Mean (SD)		12 Months, Mean (SD)		Adjusted Difference in Change (95% CI) <sup>a</sup>
	Control Group (n=179)	Intervention Group (n=178)	Control Group (n=153)	Intervention Group (n=141)	
BMI	22.59 (4.49)	22.70 (4.70)	23.37 (4.68)	23.30 (4.71)	-0.19 (-0.70 to 0.33) <sup>b</sup>
BMI z score	0.78 (1.16)	0.82 (1.12)	0.81 (1.17)	0.76 (1.16)	-0.08 (-0.20 to 0.04) <sup>b</sup>
Body fat (%)	28.31 (6.76)	29.58 (6.54)	32.55 (5.87)	32.72 (5.85)	-1.09 (-2.88 to 0.70) <sup>b</sup>
Push-up test, repetitions <sup>c</sup>	11 (6 to 16)	10 (6 to 16)	10 (6 to 16)	11 (7 to 19)	2.38 (-2.47 to 7.22) <sup>b</sup>
Prone support test, s <sup>c</sup>	36.8 (25.6 to 64.2)	44.0 (28.4 to 67.0)	42.8 (26.0 to 62.0)	50.0 (31.8 to 69.0)	-4.44 (-17.93 to 9.04)
Accelerometer counts/min <sup>c,d</sup>	363.0 (313.2 to 568.9)	388.6 (310.8 to 459.7)	360.1 (265.0 to 452.6)	322.1 (270.5 to 392.7)	-46.19 (-123.26 to 31.88)
MVPA, min/d <sup>c,d</sup>	32.0 (24.7 to 42.1)	33.5 (20.5 to 40.1)	25.0 (16.5 to 41.7)	21.5 (15.9 to 28.9)	-4.28 (-13.82 to 5.25)
Daily screen time, min/d <sup>c</sup>	220.7 (162.7 to 341.8)	240.0 (161.8 to 368.6)	248.6 (177.9 to 355.7)	231.4 (161.8 to 375.4)	-30.67 (-62.43 to -1.06) <sup>b</sup>
Weekday screen time, min/d <sup>c</sup>	209.0 (156.0 to 289.0)	216.0 (142.5 to 349.5)	236.0 (156.0 to 333.5)	222.0 (142.5 to 326.1)	-25.39 (-54.14 to 3.36) <sup>b</sup>
Weekend screen time, min/d <sup>c</sup>	255.0 (150.0 to 420.0)	300.0 (178.8 to 450.0)	300.0 (180.0 to 608.0)	285.0 (180.0 to 420.0)	-42.90 (-100.41 to 14.61) <sup>b</sup>
Daily energy intake, kcal/d	2241.2 (1259.8)	2598.8 (1763.6)	2233.8 (1551.9)	2524.8 (1610.0)	-62.0 (-464.2 to 340.3) <sup>b</sup>
Daily energy intake per kcal/kg/d <sup>c</sup>	36.7 (106.4 to 214.2)	35.6 (110.4 to 222.3)	33.1 (93.9 to 193.6)	35.7 (98.4 to 226.5)	-0.52 (-7.31 to 6.27) <sup>b</sup>
Perceived body fat, low=1 to high=5	3.88 (1.51)	3.75 (1.48)	3.78 (1.46)	3.84 (1.49)	0.19 (-0.10 to 0.47) <sup>b</sup>
Physical self-esteem, low=1 to high=5	3.74 (1.25)	3.71 (1.26)	3.63 (1.17)	3.75 (1.28)	0.17 (-0.15 to 0.48) <sup>b</sup>
Global self-esteem, low=1 to high=5	4.28 (1.01)	4.16 (1.09)	4.29 (0.99)	4.09 (1.10)	-0.08 (-0.30 to 0.14)

Abbreviations: BMI, body mass index calculated as weight in kilograms divided by height in meters squared; MVPA, moderate-to-vigorous physical activity.

<sup>a</sup>Adjusted mean difference and 95% confidence interval between NEAT Girls and control groups after 12 months (intervention minus control).

<sup>b</sup>Changes in favor of the intervention group.

<sup>c</sup>Data were transformed owing to non-normality; median and interquartile ranges provided.

<sup>d</sup>191 and 89 participants wore accelerometers for 600 min or more on 4 or more days including a weekend day at baseline and posttest, respectively.

strategy for preventing unhealthy weight gain in youth.<sup>49,50</sup> Screen time is associated with unhealthy dietary behaviors in youth<sup>51</sup> and the reductions in screen time observed in the intervention group may have helped to reduce energy intake. Although we did not observe clinically important changes in total energy intake, this could be owing to the lack of sensitivity in the food frequency questionnaire used in our study.

Culturally appropriate obesity prevention interventions appear to be more effective than those that disregard cultural identity.<sup>21</sup> Although NEAT Girls was not targeted toward a specific cultural group, the importance of addressing cultural uniqueness is relevant to our study and we employed a number of strategies to ensure that the intervention was tailored and relevant to the participants. For example, the intervention logo and materials were branded and tailored to appeal to adolescent girls. A variety of novel strategies were used to engage girls in the interactive seminars (eg, game show format) and participants were encouraged to bring their own music to be played on a portable digital music player in the enhanced school sports sessions. The enhanced sports sessions focused on lifetime activities that are appealing to adolescent girls and the nutrition workshops involved the preparation of inexpensive healthy snacks and meals. Both the enhanced school sports sessions and the nutrition workshops were rated favorably by girls, but the attendance at sessions was not as high as antici-

pated. NEAT Girls involved parental newsletters and home challenges to engage parents in the intervention, but we did not survey parents and cannot determine whether parental behaviors and support changed as a result of the intervention.

The strengths of this study include the group randomized controlled trial design, the monitoring of intervention compliance, the unique study population, and the high level of participant retention. However, there are some limitations that should be noted. First, despite employing a number of strategies to improve monitoring compliance, only a small number of participants provided useable accelerometer data at baseline (53.5%) and posttest (24.9%). Second, dietary intake was assessed using a food frequency questionnaire, which lacks sensitivity to detect small changes in energy intake. Third, we underestimated the school-level intraclass correlation coefficients for the body composition variables in the NEAT Girls study, which resulted in reduced statistical power. Given the higher than expected intraclass correlation coefficients and the small number of clusters, we conducted additional statistical analyses that adjusted for the clustered nature of the data but did not include time as a random effect in the statistical models. In these models, we found a significant intervention effect for body fat percentage ( $P=.02$ ) and a marginally significant effect for BMI z score ( $P=.10$ ). Finally, screen time was measured using self-report and the results may be influ-



enced by experimenter expectancies and evaluation apprehension.

In summary, the NEAT Girls intervention resulted in small improvements in body composition and large reductions in self-reported screen time. Our findings demonstrate the potential for multicomponent school-based interventions for the prevention of unhealthy weight gain in adolescent girls attending schools in low-income communities.

Accepted for Publication: January 18, 2012.

Published Online: May 7, 2012. doi:10.1001/archpediatrics.2012.41

**Correspondence:** David R. Lubans, PhD, Priority Research Centre in Physical Activity and Nutrition, School of Education, Faculty of Education and Arts, University of Newcastle, Callaghan, NSW, Australia 2308 (david.lubans@newcastle.edu.au).

**Author Contributions:** *Study concept and design:* Lubans, Morgan, Okely, Collins, Callister, and Plotnikoff. *Acquisition of data:* Lubans, Okely, and Dewar. *Analysis and interpretation of data:* Lubans, Okely, Batterham, and Plotnikoff. *Drafting of the manuscript:* Lubans, Morgan, and Collins. *Critical revision of the manuscript for important intellectual content:* Lubans, Morgan, Okely, Dewar, Batterham, Callister, and Plotnikoff. *Statistical analysis:* Lubans and Batterham. *Obtained funding:* Lubans, Morgan, Okely, Collins, Callister, and Plotnikoff. *Administrative, technical, and material support:* Lubans, Dewar, and Callister. *Study supervision:* Lubans and Okely.

**Financial Disclosure:** Dr Collins' work is funded by an Australian National Health and Medical Research Council Career Development Fellowship.

**Funding/Support:** This study is funded by grant DP1092646 from the Australian Research Council.

**Additional Contributions:** We thank Project Manager Tara Finn and the following research assistants: Sarah Costigan, Rebecca Horton, Melanie Fagg, Kayla Lawson, and Xanne Janssen. We also thank the schools, teachers, and study participants.

## REFERENCES

1. Wang LY, Chyen D, Lee S, Lowry R. The association between body mass index in adolescence and obesity in adulthood. *J Adolesc Health*. 2008;42(5):512-518.
2. Dietz WH. Health consequences of obesity in youth: childhood predictors of adult disease. *Pediatrics*. 1998;101(3, pt 2)(suppl):518-525.
3. Singh AS, Mulder C, Twisk JWR, van Mechelen W, Chinapaw MJM. Tracking of childhood overweight into adulthood: a systematic review of the literature. *Obes Rev*. 2008;9(5):474-488.
4. Lobstein T, Frelut ML. Prevalence of overweight among children in Europe. *Obes Rev*. 2003;4(4):195-200.
5. Ogden CL, Carroll MD, Curtin LR, Lamb MM, Flegal KM. Prevalence of high body mass index in US children and adolescents, 2007-2008. *JAMA*. 2010;303(3):242-249.
6. Olds TS, Tomkinson GR, Ferrar KE, Maher CA. Trends in the prevalence of childhood overweight and obesity in Australia between 1985 and 2008. *Int J Obes (Lond)*. 2010;34(1):57-66.
7. Stamatakis E, Wardle J, Cole TJ. Childhood obesity and overweight prevalence trends in England: evidence for growing socioeconomic disparities. *Int J Obes (Lond)*. 2010;34(1):41-47.
8. Hardy LL, King L, Espinel P, Cosgrove C, Bauman A. *NSW Schools Physical Activity and Nutrition Survey (SPANS) 2010: Full Report*. Sydney, Australia: New South Wales Ministry of Health; 2011.
9. Brown T, Summerbell C. Systematic review of school-based interventions that focus on changing dietary intake and physical activity levels to prevent childhood obesity: an update to the obesity guidance produced by the National Institute for Health and Clinical Excellence. *Obes Rev*. 2009;10(1):110-141.
10. Jones RA, Sinn N, Campbell KJ, et al. The importance of long-term follow-up in child and adolescent obesity prevention interventions. *Int J Pediatr Obes*. 2011;6(3-4):178-181.
11. Foster GD, Linder B, Baranowski T, et al; HEALTHY Study Group. A school-based intervention for diabetes risk reduction. *N Engl J Med*. 2010;363(5):443-453.
12. Singh AS, Chin A Paw MJM, Brug J, van Mechelen W. Dutch obesity intervention in teenagers: effectiveness of a school-based program on body composition and behavior. *Arch Pediatr Adolesc Med*. 2009;163(4):309-317.
13. Lubans DR, Morgan PJ, Aguiar EJ, Callister R. Randomized controlled trial of the Physical Activity Leaders (PALs) program for adolescent boys from disadvantaged secondary schools. *Prev Med*. 2011;52(3-4):239-246.
14. Katz DL, O'Connell M, Njike VY, Yeh MC, Nawaz H. Strategies for the prevention and control of obesity in the school setting: systematic review and meta-analysis. *Int J Obes (Lond)*. 2008;32(12):1780-1789.
15. Klesges RC, Obarzanek E, Kumanyika S, et al. The Memphis Girls' health Enrichment Multi-site Studies (GEMS): an evaluation of the efficacy of a 2-year obesity prevention program in African American girls. *Arch Pediatr Adolesc Med*. 2010;164(11):1007-1014.
16. Neumark-Sztainer DR, Friend SE, Flattum CF, et al. New moves-preventing weight-related problems in adolescent girls: a group-randomized study. *Am J Prev Med*. 2010;39(5):421-432.
17. Robinson TN, Matheson DM, Kraemer HC, et al. A randomized controlled trial of culturally tailored dance and reducing screen time to prevent weight gain in low-income African American girls: Stanford GEMS. *Arch Pediatr Adolesc Med*. 2010;164(11):995-1004.
18. Nader PR, Bradley RH, Houts RM, McRitchie SL, O'Brien M. Moderate-to-vigorous physical activity from ages 9 to 15 years. *JAMA*. 2008;300(3):295-305.
19. Berkey CS, Rockett HR, Colditz GA. Weight gain in older adolescent females: the internet, sleep, coffee, and alcohol. *J Pediatr*. 2008;153(5):635-639, 639, e1.
20. Eissa MA, Dai S, Mihalopoulos NL, Day RS, Harrist RB, Labarthe DR. Trajectories of fat mass index, fat free-mass index, and waist circumference in children: Project HeartBeat! *Am J Prev Med*. 2009;37(1)(suppl):S34-S39.
21. Wilson DK. New perspectives on health disparities and obesity interventions in youth. *J Pediatr Psychol*. 2009;34(3):231-244.
22. Lubans DR, Morgan PJ, Dewar D, et al. The Nutrition and Enjoyable Activity for Teen Girls (NEAT girls) randomized controlled trial for adolescent girls from disadvantaged secondary schools: rationale, study protocol, and baseline results. *BMC Public Health*. 2010;10(652):652. doi:10.1186/1471-2458-1110-1652.
23. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ*. 2010;340:c869. doi:10.1136/bmj.c1869.
24. Cole TJ, Faith MS, Pietrobello A, Heo M. What is the best measure of adiposity change in growing children: BMI, BMI %, BMI z-score or BMI centile? *Eur J Clin Nutr*. 2005;59(3):419-425.
25. Robinson TN, Kraemer HC, Matheson DM, et al. Stanford GEMS phase 2 obesity prevention trial for low-income African-American girls: design and sample baseline characteristics. *Contemp Clin Trials*. 2008;29(1):56-69.
26. Amorim LD, Bangdiwala SI, McMurray RG, Creighton D, Harrell J. Intraclass correlations among physiologic measures in children and adolescents. *Nurs Res*. 2007;56(5):355-360.
27. Murray DM. *Design and Analysis of Group-Randomised Trials*. New York, New York: Oxford University Press; 1998.
28. Lubans DR, Morgan PJ, Callister R, Collins CE. Effects of integrating pedometers, parental materials, and e-mail support within an extracurricular school sport intervention. *J Adolesc Health*. 2009;44(2):176-183.
29. Lubans DR, Morgan PJ, Callister R, Collins CE, Plotnikoff RC. Exploring the mechanisms of physical activity and dietary behavior change in the program x intervention for adolescents. *J Adolesc Health*. 2010;47(1):83-91.
30. Bandura A. *Social Foundations of Thought and Action: A Social Cognitive Theory*. Englewood Cliffs, New Jersey: Prentice-Hall; 1986.
31. Lubans DR, Foster C, Biddle SJH. A review of mediators of behavior in interventions to promote physical activity among children and adolescents. *Prev Med*. 2008;47(5):463-470.
32. Cerin E, Barnett A, Baranowski T. Testing theories of dietary behavior change in youth using the mediating variable model with intervention programs. *J Nutr Educ Behav*. 2009;41(5):309-318.
33. Lubans DR, Morgan PJ, Tudor-Locke C. A systematic review of studies using pedometers to promote physical activity among youth. *Prev Med*. 2009;48(4):307-315.
34. Cole TJ, Bellizzi MC, Flegal KM, Dietz WH. Establishing a standard definition for child overweight and obesity worldwide: international survey. *BMJ*. 2000;320(7244):1240-1243.

35. Lubans DR, Morgan PJ, Callister R, et al. Test-retest reliability of a battery of field-based health-related fitness measures for adolescents. *J Sports Sci*. 2011; 29(7):685-693.
36. Cooper Institute for Aerobics Research. *The Prudential FITNESSGRAM: Test Administration*. Dallas, Texas: Cooper Institute for Aerobics Research; 1992.
37. Treuth MS, Schmitz K, Catellier DJ, et al. Defining accelerometer thresholds for activity intensities in adolescent girls. *Med Sci Sports Exerc*. 2004;36(7):1259-1266.
38. Trost SG, McIver KL, Pate RR. Conducting accelerometer-based activity assessments in field-based research. *Med Sci Sports Exerc*. 2005;37(11)(suppl): S531-S543.
39. Trost SG, Pate RR, Freedson PS, Sallis JF, Taylor WC. Using objective physical activity measures with youth: how many days of monitoring are needed? *Med Sci Sports Exerc*. 2000;32(2):426-431.
40. Freedson P, Pober D, Janz KF. Calibration of accelerometer output for children. *Med Sci Sports Exerc*. 2005;37(11)(suppl):S523-S530.
41. Watson JF, Collins CE, Sibbritt DW, Dibley MJ, Garg ML. Reproducibility and comparative validity of a food frequency questionnaire for Australian children and adolescents. *Int J Behav Nutr Phys Act*. 2009;6:62.
42. Hardy LL, Booth ML, Okely AD. The reliability of the Adolescent Sedentary Activity Questionnaire (ASAQ). *Prev Med*. 2007;45(1):71-74.
43. Marsh HW, Richards GE, Johnson S, Roche L, Tremayne P. Physical self-description questionnaire: psychometric properties and a multimethod analysis of relations to existing instruments. *J Sport Exerc Psychol*. 1994;16:270-305.
44. Mallinckrodt CH, Watkin JG, Molenberghs G, Carroll RJ, Lilly E. Choice of the primary analysis in longitudinal clinical trials. *Pharm Stat*. 2004;3(3):161-169. doi:10.1002/pst.124.
45. Salmon J, Tremblay MS, Marshall SJ, Hume C. Health risks, correlates, and interventions to reduce sedentary behavior in young people. *Am J Prev Med*. 2011; 41(2):197-206.
46. Steinbeck K, Baur L, Cowell C, Pietrobelli A. Clinical research in adolescents: challenges and opportunities using obesity as a model. *Int J Obes*. 2009;33(1):2-7.
47. Dai S, Fulton JE, Harrist RB, Grunbaum JA, Steffen LM, Labarthe DR. Blood lipids in children: age-related patterns and association with body-fat indices: Project HeartBeat! *Am J Prev Med*. 2009;37(1)(suppl):S56-S64.
48. Freedman DS, Dietz WH, Srinivasan SR, Berenson GS. The relation of overweight to cardiovascular risk factors among children and adolescents: the Bogalusa Heart Study. *Pediatrics*. 1999;103(6, pt 1):1175-1182.
49. Epstein LH, Roemmich JN, Robinson JL, et al. A randomized trial of the effects of reducing television viewing and computer use on body mass index in young children. *Arch Pediatr Adolesc Med*. 2008;162(3):239-245.
50. Epstein LH, Paluch RA, Gordy CC, Dorn J. Decreasing sedentary behaviors in treating pediatric obesity. *Arch Pediatr Adolesc Med*. 2000;154(3):220-226.
51. Pearson N, Biddle SJ. Sedentary behavior and dietary intake in children, adolescents, and adults; a systematic review. *Am J Prev Med*. 2011;41(2):178-188.

#### Announcement

The *Archives of Pediatrics & Adolescent Medicine* will devote its May 2013 issue to pediatric hospital medicine. We are interested in a broad range of research related to hospital care, including clinical and comparative effectiveness research on the inpatient management of pediatric diseases. We invite all hospital-based pediatricians, including hospitalists, emergency medicine physicians, neonatologists, and intensivists, to submit manuscripts, preferably by September 15, 2012.

# **Study Guide**

## **Session 2**

### **Determining the Validity of Research on a Therapy**

**Allen F. Shaughnessy, PharmD, MMedEd**

#### **The aims of this session are to:**

- 1) Build on your previous study of epidemiology and biostatistics to apply the concepts to medical research used in clinical medicine
- 2) Introduce you to some additional issues of study validity
- 3) Practice how to quickly and accurately evaluate research methods and findings about treatments used in clinical medicine

**Specific Objectives:** By completing the initial reading and participating in class, students should be able to:

- 1) List and describe the major threats to validity in studies evaluating treatments
- 2) Interpret the results of clinical studies
- 3) Use a set of questions to quickly evaluate a study for validity and relevance.

This study guide provides an outline of the concepts necessary to meet these objectives. It contains hyperlinks to short videos, web pages, or articles that explain the concepts in other ways or in greater detail. You can follow these hyperlinks if the explanations and examples I've given you are not sufficient to help you understand and to help you complete the readiness assessment test.

### **What is evidence-based medicine, and why is it important?**

Evidence based medicine, according to the innovators of this approach, is “the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients. The practice of evidence based medicine means integrating individual clinical expertise with the best available external clinical evidence from systematic research.”\*

Further, another good explanation is that, “Evidence based medicine (EBM) has not developed a new concept of evidence; its major contribution lies in the emphasis it places on a hierarchy of evidential reliability, in which conclusions related to evidence from controlled experiments are accorded greater credibility than conclusions grounded in other sorts of evidence.”†



For a third definition, closely in line with these, watch: [EBM defined](#) (4:27)

---

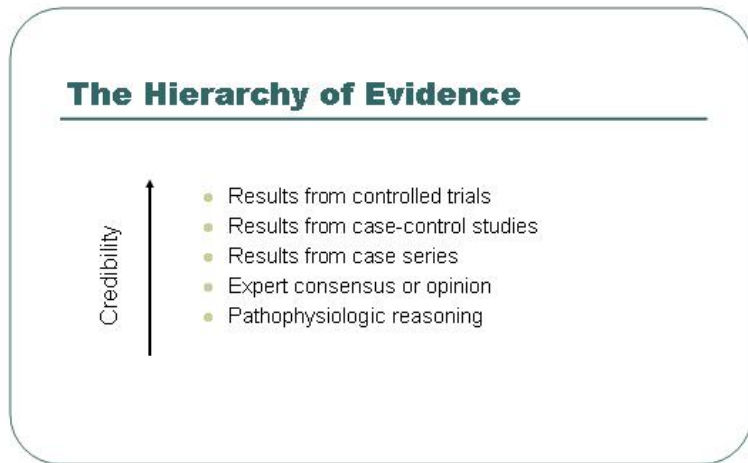
\* <http://www.bmj.com/content/312/7023/71>

† <http://www.bmj.com/content/329/7473/1024>.

So, the idea behind evidence-based medicine is that there is a [hierarchy of evidence](#), with some types of evidence being more likely to represent the truth than others are.

This idea, in turn, is based on the concept that truth is a probability rather than an absolute in medicine. A treatment may increase the likelihood of certain patients will receive benefit, but we are by no means certain that everyone will benefit.

*For example*, whenever we study a new treatment, there are various ways that a study could be designed. We might take an epidemiologic approach. Or, we might study the pharmacology of a drug. We might try the treatment on one person and publish a case report, or study a bunch of people in a case series. The credibility of the results depends on the study design.



46

## Issues of study validity

The following questions can be used to determine whether a study is sufficiently valid. There is a [worksheet](#) that can be used to keep track of the answers.



Read: "[Evaluating and Understanding Articles About Treatment](#)" (3 pages)

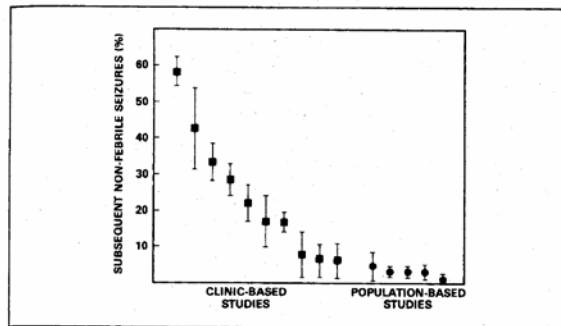
***Are the studied patients similar enough to your patients that you can apply the results in your practice?***

Another aspect of a study is whether the **population studied is similar** to your population. The severity and likelihood of illness and the response to treatment will vary based on where study subjects are found.

# Example

## *What is the risk of a non-febrile seizure in child at some point following a febrile seizure?*

This graph shows the difference in studies reporting the risk of febrile seizures in children who have had a seizure as a result of a high fever. Studies conducted in pediatric neurologist offices report a follow-up seizure rate of about 40% - 60%. However, studies in emergency departments report a follow-up seizure rate of about 5% or less. Clearly, children who end up seeing a pediatric neurologist are quite different from those simply seen in an emergency department.



**Figure 6-1.** The risk of nonfebrile seizures following febrile convulsions. ■ = studies from specialty clinics and hospitals; ● = studies from general populations. (Adapted from J. H. Ellenberg and K. B. Nelson. Sample selection and the natural history of disease. Studies of febrile seizures. *J.A.M.A.* 243:1337, 1980.)

## *Were the subjects randomly assigned?*

When it comes down to accurately evaluating any therapeutic intervention (drug, procedure, or even whether to obtain a particular diagnostic test), we generally should turn, if we can, to **a randomized control trial**.

Randomization is really the best protection that we have against being misled. In comparative studies, in which biases such as the compliance effect and the placebo effect are controlled, randomized studies are less likely than other types of studies to show that a treatment was effective when it really isn't. A longer explanation of why this is so is in the YouTube video:



[Why is randomization important? \(5:03\)](#)



# Example

*How well does warfarin work to prevent deaths in patients who have had a heart attack?*

Thomas Chalmers and colleagues compared the differences in reported rates of anticoagulation with warfarin in patients with an acute heart attack. They found that the demonstrated effectiveness dropped quite a bit when comparing results from randomized controlled studies with results from historical control (before-after) studies.

## The value of randomization

- 32 controlled trials of anticoagulation for the treatment of acute myocardial infarction
- Results by type of study:

	Relative Risk Reduction	Case fatality rate
Historical control	42%	38.3%
Controlled trial	33%	29.2%
Randomized controlled trial	31%	19.6%

A 26% drop in reported effectiveness

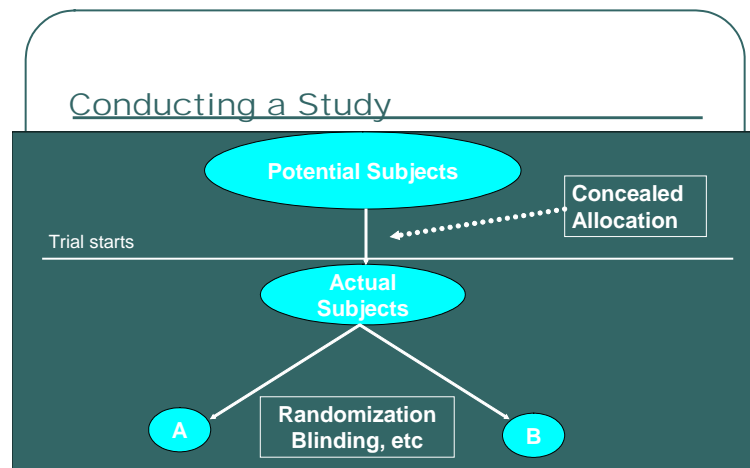
Chalmers TC, et al. N Engl J Med 1977;297:1091-6.

58

## ***Were steps taken to conceal the treatment assignment from study personnel entering patients into the study?***

The gold standard for a therapeutic trial is no longer simply just the concept of randomization, but is also whether or not **allocation assignment was concealed** from the enrolling investigator (“concealed allocation”). The definition of concealed allocation: Did the investigators know to which group a potential subject would be assigned before they were actually enrolled in this study? Concealed allocation is used in randomized studies to protect the integrity of the randomization process. Randomization is meant to prevent selection bias, and if there was no concealed allocation, the study is again susceptible to selection bias. In this case, the selection bias is occurring during the enrollment period and causes the study population to be unrepresentative of the population of potential subjects.

Trials that do not use conceal allocation consistently overestimate the benefit of the treatment by 40%.<sup>‡</sup> The recent changes in recommendations for breast cancer screening using mammography were prompted by a discovery almost 15 years ago that the lack of allocation concealment biased studies that evaluated the effectiveness of screening mammography.



*Allocation concealment is not the same thing as blinding.* Allocation concealment occurs before a study begins, during the process of selecting patients for a study. It is possible to have a study that is blinded, but does not conceal allocation. It's also possible to have a non blinded study that does conceal allocation.



Watch [“What is concealed allocation?”](#) (1:51)



Read: [Screening Mammography: Controversies and Headlines](#)

<sup>‡</sup> [JAMA. 1995 Feb 1;273\(5\):408-12.](#)



Read: [Allocation concealment](#) for examples of blinded studies without allocation concealment as well as non-blinded studies with concealed allocation.

# Example

For example, a study performed in the early 1990s, before allocation concealment was considered an important issue, evaluated the benefit of artificial surfactant for newborn premature infants in the neonatal intensive care unit. It was possible for physicians and nurses caring for infants to hold the study envelopes up to a light and determine whether or not the next baby that could potentially be enrolled in the study would either receive surfactant or placebo.



Here's a possible scenario: Let's say the investigators held study envelopes up to the light to determine whether the child would be put in the surfactant or placebo group. If they were to be put in the placebo group, children who were marginal and likely not to survive, regardless of any intervention, might have been enrolled. Children with a reasonable chance for survival may not have been enrolled in the study (but simply given surfactant outside the study).

As a result, sicker children could have been selectively enrolled into the placebo group. This selective enrollment may have greatly overestimated the benefit of surfactant because the more healthy children with the higher likelihood of survival were not enrolled in the control arm of the study. This study was published in the medical literature as a randomized, double-blind controlled trial. Since the authors did not specify that the envelopes were sealed and opaque, readers would not know that allocation was concealed.

## ***Were all patients who entered the trial properly accounted for at its conclusion?***

We want to be able to follow the history of every subject to know what happened to them. We want to avoid selective analysis of the data, which rarely happens in published work today. Most journals require a figure with a flow diagram (below) showing what happened to every patient. What percentage of patients were lost to follow-up? When more than 10 - 20% of patients are lost to follow-up it is difficult to accept the results of the study.

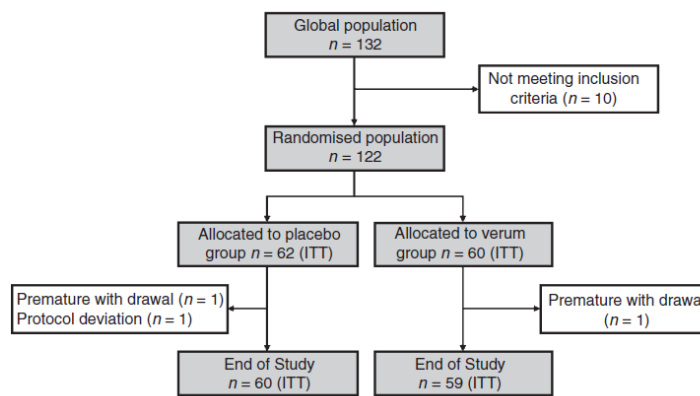


Figure 2 | Diagram of study flow. Verum, *B. bifidum* MIMBb75.

## ***Were patients analyzed in the groups to which they were randomized (“intention-to-treat” analysis)?***

Data should be analyzed by **intention-to-treat**, meaning that patients were analyzed in the group to which they were initially assigned regardless of whether they actually received the treatment. This method of analysis more accurately reflects the real world effect of an intervention where not all patients are compliant with treatment. Studies reporting results only from subjects in the intervention group proven to have taken the medicine (“on-treatment” analysis) compared with the placebo group are actually comparing compliant individuals with both compliant and non-compliant ones. Regardless of what intervention is studied, compliant patients will usually do better than those who are non-compliant, making the treatment look more effective than it actually is.



Watch “[What is intention-to-treat analysis?](#)” (1:11)



Watch “[Intention-to-treat analysis: What is it and why is it important?](#)” (4:43)

## A review of useful statistics<sup>§</sup>

After we've decided a study is valid we need to **understand the results**. Understanding just a few statistics is all that's necessary. Briefly, here are some statistics that are useful:

**P value:** "P" stands for the *probability* that the difference between two averages, rates, etc – whatever we're measuring – is due to chance. It is the likelihood that the difference we see in the numbers is not "real", in that it represents an actual difference due to one treatment over another, but that, if we did the study again, we might find the difference to no longer be present. So, a P-value of 0.05 tells us that we have a 5% risk that the difference is due to chance, or a 95% likelihood that the difference we see represents a real difference.



Watch [What is a P-Value?](#) (5:56)



*Self-test: Which of the results on this Table are not due to chance?*

Table 3| Comparison of Cohen-Mansfield agitation inventory (CMAI) total score between control and intervention (stepwise protocol for treatment of pain) groups using repeated measures analysis of covariance (ANCOVA)\*

Week	Mean (SD) CMAI total score		Effect of intervention on CMAI total†		Intraclass correlation coefficient‡
	Control group	Intervention group	Estimate (95% CI)	P value	
0	56.2 (16.1), n=177	56.5 (15.2), n=175	—	—	0.162
2	53.9 (17.0), n=161	52.0 (19.5), n=158	-3.6 (-0.5 to -6.7)	0.022	0.261
4	52.5 (16.3), n=160	49.4 (19.0), n=148	-4.1 (-0.9 to -7.4)	0.012	0.231
8	52.8 (16.8), n=157	46.9 (18.7), n=147	-7.0 (-3.7 to -10.3)	<0.001	0.226
12	52.5 (16.0), n=152	50.3 (20.3), n=142	-3.2 (0.1 to -6.4)	0.058	0.253

\*Baseline score as covariate and least squares weighted by number of patients within cluster; P value from multivariate test of intervention was 0.002, and cross effect between week and intervention was <0.001.

†Variable estimate by week of effect of intervention on CMAI score from estimated model.

‡Proportion of total variance between clusters, and measured within framework of ANCOVA.

[Answer here](#)

<sup>§</sup> To read more, see: <http://tinyurl.com/cx68dxx>

**Number needed to treat (NNT):** The NNT tells us how many people we need to treat instead of not treat, or the number that need to be treated with one therapy instead of another, for one *additional* person to benefit. It takes into account the idea that some people will benefit even without therapy, some people will benefit from another treatment, and it accounts for the fact that treatment usually is not 100% effective and thus some people who receive treatment will not benefit.

It is calculated based only on results that are presented as rates (ie, in percents). We calculate by taking the difference between two rates (the rates of cure, the rate of death, etc) and divide that number into 100.

$$\text{NNT} = \frac{100}{\begin{array}{c} \% \text{ in treatment group} - \% \text{ in control group} \\ \text{(use the absolute value; that is, don't worry about a minus sign)} \end{array}}$$



Watch [“What is number needed to treat?”](#) (0:40)



Watch [The NNT Tutorial](#) (3:57)



*Self-test: From the study results below, calculate how many additional patients would need to be treated with antibiotic rather than placebo for one additional person not to develop new lesions.*

Results: On 30-day follow-up (successful in 69% of patients), we observed fewer new lesions in the antibiotic (4/46; 9%) versus placebo (14/50; 28%) groups, difference 19%, 95% confidence interval 4% to 34%,  $P < .02$ .

[Answer here](#)

**Relative risk:** Relative risk helps us understand the difference between two rates – rates of death, rates of a side effect, etc. Depending on how the results are worded, it can represent the risk of harm or the risk of benefit. For example, a relative risk of 1.3 tells us that the likelihood of something happening is 30% higher in one group vs. the other; a relative risk of 0.7 tells us the likelihood of something happening is 30% lower in one group vs. the other. If the relative risk = 1.0, then there is no difference.

Unfortunately, the relative risk doesn't take into account the baseline risk, or the risk of no treatment, so we have to ask ourselves, 30% of what?



[Relative risk](#) (7:56)

The **confidence interval** tells us the range of possible results. A 95 percent confidence interval (95% CI) indicates that if the study were repeated 100 times, the study results would fall within this interval 95 times. In the example above for number needed to treat, the rate difference was 19% (95% CI = 4%-34%). The 95% confidence interval tells us that, if we performed the study again many times, we would find a rate difference somewhere between 4% and 34% 95 out of 100 times.



Watch ["What are Confidence Intervals?"](#) (1:29)



Watch [Interpretation of Confidence Interval](#) (1:54)

The **power** of a study is its ability to find a difference between the groups if a group truly exists. We only need to concern ourselves with power if there was not a difference between treatments – then we need to ask whether the study had sufficient power.



Watch [Power of a Study](#) (2:03)

## Using a worksheet to evaluate articles for relevance and validity

The goal of using this or other worksheets is to quickly determine whether it is worth taking the time to read a research study (or a synopsis of that study). It allows determination, based on answers to the questions, whether the information is relevant to you, and whether the study design has sufficient rigor to apply the results to your patients.

The questions focus on the study design issues described above. The first 6 questions address study “musts” – they ask about issues of relevance or validity that must be present if the study results are to be applied to clinical practice. The answers to these questions must be yes regardless of the answers to the rest of the questions.

The goal of this worksheet is **not** to determine whether a study is “good” or “bad.” Instead, we will use it to determine whether the results reported by the study are likely to occur if we use the same approach in our patients. As a result, the information is either useful to us, or not.

### Answers to the self-test questions:

*Self-test: Which of the results on this Table are not due to chance?*

All but week 12 are statistically significant.

[Back to the self-test](#)

*Self-test: From the study results below, calculate how many additional patients would need to be treated with antibiotic rather than placebo for one additional person not to develop new lesions.*

$$\text{NNT} = 100 / (28\% - 9\%) = 5.26.$$
 One additional person would not have follow-up lesions after 30 days for every 6 patients treated with antibiotic instead of placebo.

[Back to the self-test](#)



[Return to instructions](#)

## A Worksheet for Articles about Treatment

### Determine *Relevance*

*Is this article worth taking the time to read? If the answer to any of these questions is No, it may be better to read other articles first.*

#### Based on the conclusion of the abstract:

- A. Did the authors study an outcome that patients would *care* about? (Be careful to avoid results that require extrapolation to an outcome that truly matters to patients)

Yes (go on )

No (**stop**)

- B. Is the problem studied one that is *common* to your practice and the intervention feasible?

Yes (go on )

No (**stop**)

- C. Will this information, if true, require you to *change* your current practice?

Yes (go on )

No (**stop**)

### Determine *Validity*

*If the answers to all three questions above are Yes, then continued assessment of the article is mandatory.*

#### D. Population

1. Are the studied patients similar enough to your patients that you can apply the results in your practice?

Yes

No (**Stop**)

#### E. Study design

1. Was it a controlled trial?  
2. Were the subjects randomly assigned?  
3. Were steps taken to conceal the treatment assignment from study personnel entering patients into the study?  
4. Were patients, providers and outcome assessors “blind” to treatment?

Yes

No (**Stop**)

Yes

No (**Stop**)

Yes

No

Yes

No

#### F. Study conduct

1. Were all patients who entered the trial properly accounted for at its conclusion?  
a. Was follow-up complete?  
b. Were patients analyzed in the groups to which they were randomized (“intention-to-treat” analysis)?  
2. Were the intervention and control groups similar? (Table 1)

Yes

No

Yes

No

Yes

No

#### G. Study results

1. What were the results? \_\_\_\_\_

2. Are the results clinically as well as statistically significant?

Yes

No

3. If a negative trial, was the power of the study adequate?

Yes No

4. Were there other factors that might have affected the outcome?

Yes

No

5. How will it change your practice?

[Return to instructions](#)

## **TUSM.MMC Evidence Based Medicine Course Outline**

### **Course Director:**

Kathleen Fairfield, MD, DrPH

[fairfk@mmc.org](mailto:fairfk@mmc.org) 661-7614

This course is meant to closely parallel the structure and content of the same course delivered to the Boston TUSM students. Course materials were developed by Allen F. Shaughnessy, PharmD, MMedEd, Professor of Family Medicine at Tufts. The study guides, individual readiness assessment tests, and final exam are identical. The in class materials and structure are unique to each site.

**Course Dates for 2013:** 5/2, 5/8, 5/16, 5/22

All classes held at the Brighton Campus

### **Course Objectives**

#### ***Overall aims of the course***

- To reinforce the biostatistics and epidemiology knowledge components, already taught, required for the USMLE Step 1 Examination.
- To demonstrate and contextualize how key concepts of biostatistics, epidemiology, and evidence-based medicine are used in the medical literature and applied to patient care.
- To help students develop the ability to analyze, synthesize, and apply knowledge.
- To prepare students to evaluate the medical literature and apply evidence to patient care in subsequent medical school courses and clerkships.
- To encourage an inquisitive, questioning attitude toward medical knowledge and medical care.
- To further inform your understanding of public health issues you will encounter in Maine.

Specific objectives are listed for each session.

### **Course Materials**

All the materials needed for this course can be found in the Maine Track section of TUSK. For each session, you will find an article to review as well a Study Guide, which is designed to be viewed electronically, with embedded links to additional readings, video, and exercises. There is not an assigned textbook. There are no powerpoint slides. Further readings, can be found at the [Center for Information Mastery](#) website.

## Class Structure:

Students will be expected to read the article for each session as well as the online Study Guides created by Dr. Allen Shaughnessy. Students are also expected to complete an Individual Readiness Assessment Test (IRAT) for each session, to be self-graded during the discussion at the start of the session.

2 hour session:

15 minutes: gathering and IRAT review

30 minutes: didactic: key concepts

45 minutes: small group critical appraisal of the article with worksheet to complete; student lead, floating facilitators (3 for 6 groups)

30 minutes: large group discussion of content, review of worksheets, take away

## Introduction to the Course

- 1) **The assigned preparation *must be completed before each session*.** The syllabus contains an explanation of the concepts with hyperlinks to additional readings and videos that may be helpful. The syllabus can be printed out but it is designed to be read on a computer or tablet screen. If you understand the concepts by reading only the material in the syllabus, there is no need to follow the hyperlinks (*i.e., there is no “gotcha” information buried in one of the hyperlinked resources*).
- 2) **The Individual Readiness Assessment Tests (IRATs) *must be completed on your own before the start of class*.** There will be one IRAT for each class period ( $n = 4$ ). The IRATs will be opened for completion immediately after the previous class ends and will remain open until immediately before the class starts. This step assures me, and your teammates, that you come to the class prepared. It also allows me to see where the group, as a whole, has had difficulty. These are individual (no group work) and comprise 20 percent of your grade.  
  
Following class, the IRATs will be available again, this time with the answers provided, for your use to review before the final exam.
- 3) ***Class attendance is mandatory.*** Class time will be used to work in groups to complete the same questions asked on the IRAT, and then apply the ideas to an example. *Classes will not be videorecorded or streamed.*
- 4) **There is no need for computers during class, and group work in class will not use computers.** The questions and exercises will be answered using the combined knowledge and wisdom of the group.
- 5) **Homework:** Your homework is to critically assess an original research study, practice guideline, or review article that you identify on your own. You will use one of the worksheets associated with the classes, determine the article's validity, come to a conclusion regarding a clinical question, and consider how you would explain the findings to a patient and their family. See Appendix 2.

## Grading

	Percent of Grade
<b>Individual Readiness Assessment Tests (5% for each test x 4)</b>	<b>30</b>
<b>Group Participation</b>	<b>30</b>
<b>Final Exam</b>	<b>30</b>
<b>Homework</b>	<b>10</b>

*Your participation in this group work* will be graded by your team members (see Appendix 1 at the end of this document). At the END of the course each person will divide up 50 points among the team members, awarding more points to team members who came more prepared and who participated to a greater extent, and awarding fewer points to those team members who didn't. These points will be converted into a proportion of the total available points awarded to you (to account for teams of different sizes).

Example: You are in a group of 6 students. Each student has 50 points to be distributed among the group. Let's say you receive the following points from your team members:

Student A.	10 points
Student B.	15 points
Student C.	7 points
Student D.	10 points
Student E.	9 points

Total	51 points
-------	-----------

Since there are 5 other people in the group, the average score would be 50 (10 from each member to the group, other than yourself). Your net points =  $51 - 40 = 1$  points.

## Appendix 1.

### Peer Evaluation Form

Peer Evaluation      Name \_\_\_\_\_      Team # \_\_\_\_\_

Please assign scores that reflect how you really feel about the extent to which the other members of your team contributed to your learning and/or your team's performance. This will be your only opportunity to reward the members of your team who worked hard on your behalf. **(Note: If you give everyone pretty much the same score you will be hurting those who did the most and helping those who did the least.)**

**Preparation-** Were they prepared when they came to class?

**Contribution-** Did they contribute productively to group discussion and work?

**Respect for others' ideas-** Did they encourage others to contribute their ideas?

**Flexibility-** Were they flexible when disagreements occurred?

**Instructions:** In the space below please rate each of the other members of your team. Each member's peer evaluation score will be the average of the points they receive from the other members of the team. To complete the evaluation you should: 1.) List the name of each member of your team in the alphabetical order of their last names and 2.) assign an average of 10 points to the other members of your team (Thus, for example, you should assign a total of 50 points in a six-member team.) and, 3.) Differentiate some in your ratings; for example, you must give at least one score of 11 or higher (maximum = 15) and one score of 9 or lower.

Team Members	Score
1. _____	_____
2. _____	_____
3. _____	_____
4. _____	_____
5. _____	_____
6. _____	_____
7. _____	_____

**Total** \_\_\_\_\_

Appendix 2.

### **Homework Assignment**

**Who:** Each person will submit the assignment. This is not a group assignment.

**What:** You will critically appraise a research paper, practice guideline, or review article of your choice. You may want to choose a paper that is related to your small group mentoring work.

**When:** The homework should be submitted no later than the end of the course.

**Where:** E-mail a copy of the paper (as a pdf) and the evaluation form (in Word) to Kathleen at [fairfk@mmc.org](mailto:fairfk@mmc.org)

**How:** For your identified article, answer the questions on the appropriate worksheet (see each syllabus section). At the end of the questions, answer the following question:

How would you explain the findings of this article to a patient and their family? This answer should be two or three sentences.

# Individual and Combined Effects of Age, Breast Density, and Hormone Replacement Therapy Use on the Accuracy of Screening Mammography

Patricia A. Carney, PhD; Diana L. Miglioretti, PhD; Bonnie C. Yankaskas, PhD; Karla Kerlikowske, MD; Robert Rosenberg, MD; Carolyn M. Rutter, PhD; Berta M. Geller, EdD; Linn A. Abraham, MS; Steven H. Taplin, MD, MPH; Mark Dignan, PhD; Gary Cutter, PhD; and Rachel Ballard-Barbash, MD, MPH

**Background:** The relationships among breast density, age, and use of hormone replacement therapy (HRT) in breast cancer detection have not been fully evaluated.

**Objective:** To determine how breast density, age, and use of HRT individually and in combination affect the accuracy of screening mammography.

**Design:** Prospective cohort study.

**Setting:** 7 population-based mammography registries in North Carolina; New Mexico; New Hampshire; Vermont; Colorado; Seattle, Washington; and San Francisco, California.

**Participants:** 329 495 women 40 to 89 years of age who had 463 372 screening mammograms from 1996 to 1998; 2223 women received a diagnosis of breast cancer.

**Measurements:** Breast density, age, HRT use, rate of breast cancer occurrence, and sensitivity and specificity of screening mammography.

**Results:** Adjusted sensitivity ranged from 62.9% in women with extremely dense breasts to 87.0% in women with almost entirely fatty breasts; adjusted sensitivity increased with age from 68.6% in women 40 to 44 years of age to 83.3% in women 80 to 89 years of age. Adjusted specificity increased from 89.1% in women with extremely dense breasts to 96.9% in women with almost entirely fatty breasts. In women who did not use HRT, adjusted specificity increased from 91.4% in women 40 to 44 years of age to 94.4% in women 80 to 89 years of age. In women who used HRT, adjusted specificity was about 91.7% for all ages.

**Conclusions:** Mammographic breast density and age are important predictors of the accuracy of screening mammography. Although HRT use is not an independent predictor of accuracy, it probably affects accuracy by increasing breast density.

*Ann Intern Med.* 2003;138:168-175.

For author affiliations, see end of text.

[www.annals.org](http://www.annals.org)

Mammographic breast density may be the most undervalued and underused risk factor in studies investigating breast cancer occurrence (1). The risk for breast cancer is four to six times higher in women with dense breasts (2, 3). Breast density may also decrease the sensitivity and, thus, the accuracy of mammography. Radiographically dense breast tissue may obscure tumors, which increases the difficulty of detecting breast cancer. In addition, dense breast tissue may mimic breast cancer on mammography (4), which increases recall rates (4–12), reduces specificity, and compromises the benefit of screening in women with dense breasts (such as women who use HRT or who are premenopausal) (6, 8, 13). Breast density is affected by age, use of hormone replacement therapy (HRT), menstrual cycle phase, parity, body mass index, and familial or genetic tendency (4, 5, 14–21). Studies show that the sensitivity of mammography increases with age (6–8), especially in postmenopausal women whose breasts are less dense (8).

Earlier research has examined the individual effect of each factor we have described, but most studies could not adequately examine the interaction of these factors because of insufficient sample size (4–15). Studies conducted in the 1970s with data from the Breast Cancer Detection Demonstration Project (22) and New York Health Insurance Plan (23) are based on mammographic examinations that are very different from those performed using current

technology. The Mammography Quality Standards Act (24) and the standardized reporting efforts of the American College of Radiology (25) have resulted in important improvements in mammography that necessitate reexamination.

We used data from the National Cancer Institute's Breast Cancer Surveillance Consortium (BCSC) (26) on 329 495 women in the United States who had 463 372 screening mammograms, which were linked to 2223 cases of breast cancer. Our goal was to examine the individual and combined effects of age, breast density, and HRT use on mammographic accuracy. This large data set provides a unique opportunity to examine these issues in women undergoing screening mammography in the United States, especially women younger than 50 years of age and older than 80 years of age. We chose to study a sample that had been recently screened (within the previous 2 years) so that the risk for breast cancer would be similar to that in women who receive routine mammographic screening.

## METHODS

### Data Collection

Initially, we included data on women 40 to 89 years of age who underwent screening mammography between 1996 and 1998, as submitted by seven registries in the BCSC (North Carolina; New Mexico; New Hampshire;

Vermont; Colorado; Seattle, Washington; and San Francisco, California). We included women who reported having previous mammography or who had a previous mammographic examination recorded in a registry within 2 years of the index mammogram. Women with breast implants or a personal history of breast cancer were excluded. In addition, women with missing data for age (<1%), breast density (27%), or HRT use (21%) were excluded (36% of all data). Demographic characteristics, clinical characteristics, and accuracy measures for women missing any of this information were very similar to those for women with complete data. All registries obtained institutional review board approval for data collection and linkage procedures, and careful data management, processing, and security procedures were followed (27).

Consortium mammography registries and data collection procedures are described elsewhere (26). Briefly, seven institutions in seven states receive funding from the National Cancer Institute to maintain mammography registries that cover complete or contiguous portions of each state. Data are collected similarly at each registry. Demographic and history information is collected from women at the time of mammography by using a self-administered survey or face-to-face interview methods. Variables include date of birth, history of previous mammography, race or ethnicity, current use of HRT (prescription medication used to treat perimenopausal and postmenopausal symptoms), and menopausal status. We assumed that women 55 years of age and older were perimenopausal or postmenopausal. For women 40 to 54 years of age, premenopausal status was defined as having regular menstrual periods with no HRT use; perimenopausal or postmenopausal status was defined as either removal of both ovaries or uncertainty about whether periods had stopped permanently. This latter category was further classified into HRT users and non-users. These definitions recognize that HRT users with intact uteri may have menstrual-like bleeding.

Additional data, including mammographic breast density, mammographic assessment, and recommended follow-up (based on the American College of Radiology Breast Imaging Reporting and Data System [BI-RADS]), are collected from the technologist and radiologist at the time of mammography (25). Pathology data are collected from one or more sources: regional Surveillance, Epidemiology, and End Results (SEER) programs, state cancer registries, or pathology laboratories.

## Design

We included all screening examinations for women who met the described criteria and who had at least one screening mammogram in 1996, 1997, or 1998. These years were chosen to ensure 1-year follow-up for cancer reporting and to account for routine reporting schedules in obtaining data from SEER and state cancer registries. We classified mammography as screening if a radiologist indicated that the examination was a bilateral, two-view

## Context

High breast density increases breast cancer risk and the difficulty of reading mammograms. Breast density decreases with age and increases with postmenopausal hormone therapy use. The interplay of breast density, age, and hormone therapy use on the accuracy of mammography is uncertain.

## Contribution

For women with fatty breasts, the sensitivity of mammography was 87% and the specificity was 96.9%. For women with extremely dense breasts, the sensitivity of mammography was 62.9% and the specificity was 89.1%. Sensitivity increased with age. Hormone therapy use was not an independent predictor of accuracy.

## Implications

The accuracy of screening mammography is best in older women and in women with fatty breasts. Postmenopausal hormone therapy affects mammography accuracy only through its effects on breast density.

—The Editors

(craniocaudal and mediolateral) examination. To avoid including diagnostic examinations, we excluded any breast imaging study performed within the previous 9 months. Because our goal was to study routine screening, mammographic accuracy was calculated on the basis of the initial assessment of the screening views alone (only 6% required supplemental imaging). Interpretation codes included BI-RADS assessments of 0 (incomplete), 1 (negative), 2 (negative, benign), 3 (probably benign), 4 (suspicious abnormality), or 5 (highly suggestive of malignancy). In cases in which the initial screening visit included both a screening examination and additional imaging to determine an assessment, the initial screening assessment was assigned a 0 (incomplete assessment) for analysis. When a woman had different assessments by breast, we chose the highest-level assessment for the woman as a whole (woman-level assessment) on the basis of the following hierarchy of overall level of radiologic concern:  $1 < 2 < 3 < 0 < 4 < 5$ .

We defined a screening examination as positive if it was assigned a BI-RADS assessment code of 0, 4, or 5. An assessment code of 3 associated with a recommendation for immediate additional imaging, biopsy, or surgical evaluation was also classified as positive. Although the BI-RADS recommendation for a code 3 (probably benign) is short-interval follow-up, immediate work-up was recommended in 37% of code 3s in the pooled BCSC data; therefore, this assessment is more consistent with a BI-RADS code of 0 (incomplete assessment) (28). We defined a screening examination as negative if it received a BI-RADS assessment code of 1, 2, or 3 when associated with short-interval follow-up only or routine follow-up.

We classified breast pathology outcomes as cancer if



pathology or cancer registry data identified a diagnosis of invasive or ductal carcinoma in situ. Lobular carcinoma in situ (<0.01% of cancer cases in our pooled data) was not considered a diagnosis of cancer in our analyses because it cannot be detected by mammography and is not treated.

Examinations were classified as false-positive when the assessment was positive and breast cancer was not diagnosed within the follow-up period (365 days after the index screening examination or until the next examination, whichever occurred first). Examinations were classified as true-positive when the assessment was positive and cancer was diagnosed. A false-negative examination was a negative assessment with a diagnosis of cancer within the follow-up period. A true-negative examination was a negative assessment with no subsequent diagnosis of cancer within the follow-up period.

Radiographic breast density was defined according to BI-RADS as follows: 1) almost entirely fatty, 2) scattered fibroglandular tissue, 3) heterogeneously dense, and 4) extremely dense (25). We excluded one registry that collects two categories of breast density (dense or not dense) at some facilities.

### Statistical Analysis

For age, breast density, and HRT groups, we calculated rates of incident breast cancer, rates of breast cancer detected by mammography, and rates of missed cancer. To examine the nonlinear effects of age, we categorized age into 10-year groups, except for ages 40 to 59, which were divided into 5-year groups to explore changes around menopause. Accuracy indices included sensitivity and specificity. Sensitivity was calculated as true-positive/(true-positive + false-negative). Specificity was calculated as true-negative/(true-negative + false-positive). Accuracy indices were calculated separately by age groups, breast density, and HRT use.

We used logistic regression, adjusted for registry, to compare accuracy indices across age, breast density, and HRT use groups. For the analysis of sensitivity, we included women with a diagnosis of cancer in the follow-up period and modeled the probability of a true-positive mammogram (versus a false-negative mammogram) as the outcome in the logistic regression. For specificity, we included women without a diagnosis of cancer and modeled the probability of a true-negative mammogram (versus a false-positive mammogram) as the outcome.

We tested for two-way and three-way interactions by using likelihood ratios and eliminated nonsignificant interactions in the final models. We report adjusted sensitivities and specificities by age group, breast density category, and HRT use status, adjusted to the sample distribution of the other covariates in the model. To determine the adjusted sensitivity for women 40 to 44 years of age, we calculated the sensitivity for this age group and each possible combination of the remaining covariates on the basis of the estimated logistic regression model for the probability of

mammographic detection in women with cancer. We then took the average of these estimated sensitivities, weighted by the proportion of women in the analysis with the corresponding covariate combination (excluding age). The other rates were calculated in a similar manner. Simulation was used to estimate 95% CIs by sampling 100 000 values of the regression coefficients from their joint multivariate normal distribution and calculating the accuracy measures for each sample. We estimated upper and lower limits by the simulated 2.5th and 97.5th percentiles. We used SAS software, version 8.2 (SAS Institute, Inc., Cary, North Carolina), for all analyses.

### Role of the Funding Source

The National Cancer Institute supported this study as part of the BCSC. The National Cancer Institute supported the design, conduct, and reporting of the study and the decision to submit the manuscript for publication.

## RESULTS

### Characteristics of Women Eligible for the Study

Our analyses are based on information collected from 329 495 women 40 to 89 years of age who received 463 372 screening mammographic examinations. The percentages of data contributed by each registry ranged from 4% to 28%. The mean age of the women ( $\pm$ SD) was  $58 \pm 11.5$  years. Of the 80% of women who reported ethnicity, 88% were white, 6% were black, 3% were Asian, 2% were Native American or Alaskan natives, and 1% indicated "other" or "mixed" ethnicity. In a second question about Hispanic race, 8% of women reported being Hispanic. Of the 2223 cases of cancer that occurred within the follow-up period, 19% were ductal carcinoma in situ and 81% were invasive disease. Sixty-six percent of women had one screening examination, 28% had two, and 6% had three or more.

**Table 1** outlines characteristics of women included in the analysis. Current HRT use ranged from 7.6% to 48.3% and was highest around menopause. Most women had either scattered fibroglandular tissue or heterogeneously dense breasts, whereas fewer women had almost entirely fatty or extremely dense breasts.

Breast density decreased with age in HRT nonusers; a major decline was seen in the perimenopausal period (**Table 2**). When HRT users were compared with nonusers, radiographic breast density was higher in all age groups except the 40 to 44 and 45 to 49 age groups. This difference is probably due to the fact that women younger than 50 years of age who did not use HRT were less likely to be postmenopausal (only 20% to 22%) (**Table 1**). Use of HRT increases in the perimenopausal period, which indicates possible transitions in the study factors.

### Breast Cancer Rates

**Table 3** shows the increase in breast cancer per 1000 screening examinations by age. The rate of true-positive

**Table 1. Clinical Characteristics of Women and Mammograms Included in the Analysis\***

Variable	Value, n (%)
Characteristics of women (n = 329 495)	
Age 40–44 y (12.2%)	
Premenopausal	27 309 (72.5)
Peri- or postmenopausal; no HRT use	7523 (20.0)
Peri- or postmenopausal; HRT use	2845 (7.6)
Age 45–49 y (16.2%)	
Premenopausal	29 421 (58.2)
Peri- or postmenopausal; no HRT use	11 591 (22.9)
Peri- or postmenopausal; HRT use	9557 (18.9)
Age 50–54 y (17.5%)	
Premenopausal	17 001 (30.6)
Peri- or postmenopausal; no HRT use	15 943 (28.7)
Peri- or postmenopausal; HRT use	22 679 (40.8)
Age 55–59 y (13.8%)	
Peri- or postmenopausal; no HRT use	19 373 (42.5)
Peri- or postmenopausal; HRT use	26 236 (57.5)
Age 60–69 y (21.3%)	
Postmenopausal; no HRT use	38 621 (55.0)
Postmenopausal; HRT use	31 651 (45.0)
Age 70–79 y (21.3%)	
Postmenopausal; no HRT use	35 062 (70.1)
Postmenopausal; HRT use	14 474 (29.2)
Age 80–89 y (3.9%)	
Postmenopausal; no HRT use	10 671 (82.5)
Postmenopausal; HRT use	2256 (17.5)
Characteristics of mammograms (n = 463 372)	
Breast density	
Almost entirely fatty	42 237 (9.1)
Scattered fibroglandular tissue	218 129 (47.0)
Heterogeneously dense	167 003 (36.0)
Extremely dense	36 303 (7.8)

\* Menopausal status missing for 2% of women. HRT = hormone replacement therapy.

examinations increased with age. The rate of false-negative examinations was much lower than that of true-positive examinations and increased with age until age 70 years, when it leveled off. The rate of breast cancer and false-negative examinations also rose with increasing breast density. The rate of breast cancer was slightly higher in HRT users than in nonusers.

### Accuracy Indices of Screening Mammography by Age, Breast Density, and Current HRT Use

Accuracy measures of screening mammography by age, breast density, and current HRT use indicated that sensitivity and specificity increase with age (Table 3). Both indices decreased as breast density increased. Sensitivity and specificity decreased slightly among HRT users compared with nonusers.

The Figure shows the subgroup analysis of cancer rates, sensitivity, and specificity by age, breast density, and HRT use. The number of cancer cases per 1000 examinations was strongly associated with age but was also higher among women with dense breasts. In women 50 years of age and older with dense breasts, the number of cancer cases per 1000 examinations was higher for HRT users than for nonusers. For women who did not have dense breasts, current HRT use seemed to have little or no effect. Sensitivity increased with age for all categories of breast

density by HRT group; however, this age-related increase was less striking among HRT users than among nonusers. Use of HRT resulted in a more noticeable decline in sensitivity for women 70 years of age and older. Specificity increased with age in HRT nonusers but changed very little with age in HRT users. Sensitivity and specificity were highest in older women who did not use HRT and who did not have dense breasts and were lowest in younger women who had dense breasts, regardless of HRT use.

In the multivariable logistic regression analysis for sensitivity (Table 4), two-way or three-way interactions between age, breast density, and HRT use were not statistically significant after adjustment for registry ( $P > 0.2$ ). Main effects were statistically significant for age ( $P < 0.004$  for linear trend), breast density ( $P < 0.001$ ), and registry ( $P = 0.001$ ).

After adjustment for density, HRT use, and BCSC registry, adjusted sensitivity increased with age and decreasing breast density but was not significantly associated with HRT use ( $P = 0.18$ ). However, HRT use was significantly associated with lower sensitivity if breast density was excluded from the model, suggesting that breast density is a mediating factor in the effect of HRT on sensitivity.

In our analyses of specificity, we found significant two-way interactions between age and HRT use ( $P < 0.001$ ) and between age and density ( $P = 0.016$ ). Table 5 presents adjusted specificity values by these subgroups. Main effects were significant for age, breast density, HRT use, and registry ( $P < 0.001$ ). After adjustment for the other variables in the model, we found that adjusted specificity increased with age in HRT nonusers (test for trend,  $P < 0.001$ ) (Table 5). In contrast, age had no effect on specificity in women using HRT. For women 55 years of age and older, specificity was lower in HRT users compared with nonusers, even after adjustment for density. Specificity was similar in women with dense breasts but significantly increased as breast density decreased to fatty breasts in all age groups. Specificity increased with age in women with breasts showing any mammographic density ( $P < 0.001$  in all cases); however, there was no association be-

**Table 2. Mammograms in Women with Radiographically Dense Breasts, by Age and Hormone Replacement Therapy Use\***

Age Group	No HRT		HRT	
	Total	Dense	Total	Dense
y	n	n (%)	n	n (%)
40–44	46 947	28 628 (61.0)	4782	2537 (53.1)
45–49	55 362	32 112 (58.0)	16 023	8304 (51.8)
50–54	43 079	20 551 (47.7)	37 860	18 987 (50.2)
55–59	28 458	9905 (34.8)	39 105	17 832 (45.6)
60–69	57 542	16 511 (28.7)	47 379	20 336 (42.9)
70–79	49 854	13 389 (26.9)	20 551	8812 (42.9)
80–89	13 795	4107 (29.8)	2935	1295 (44.1)

\* "Dense" is defined as heterogeneously dense and extremely dense breasts combined. HRT = hormone replacement therapy.

**Table 3. Breast Cancer Occurrence, Detection, Missed Cancers, Sensitivity, and Specificity in the Screened Sample, by Age, Breast Density, and Hormone Replacement Therapy Use\***

Variable	Screening Mammograms	Cases of Cancer	Adjusted Cancer Rate per 1000 Screening Examination†	Adjusted True-Positives per 1000 Screening Examination†	Adjusted False-Negatives per 1000 Screening Examination†	Sensitivity	Specificity
	←————— <i>n</i> —————→					%	
All eligible women	463 672	2223	4.8	3.6	1.2	75.0	92.3
Age group							
40–44 y	51 729	125	2.0	1.4	0.6	65.6	90.9
45–49 y	71 385	218	2.7	2.0	0.8	69.7	90.7
50–54 y	80 939	328	3.9	2.9	1.0	72.9	91.6
55–59 y	67 563	325	4.8	3.6	1.2	73.8	92.3
60–69 y	104 921	611	5.9	4.2	1.7	73.3	93
70–79 y	70 405	501	7.1	5.7	1.4	81.4	94.1
80–89 y	16 730	115	7.3	5.9	1.5	86.1	94.3
Breast density group							
Almost entirely fatty	42 237	110	2.2	1.9	0.2	88.2	96.5
Scattered fibroglandular tissue	218 129	975	4.2	3.5	0.8	82.1	93
Heterogeneously dense	167 003	945	5.8	4.1	1.8	68.9	90.8
Extremely dense	36 303	193	6.1	3.9	2.2	62.2	89.9
Current use of HRT							
Yes	168 635	1319	4.7	3.5	1.3	76.6	92.6
No	295 037	904	4.6	3.5	1.1	72.7	91.7

\* HRT = hormone replacement therapy.

† Adjusted for mammography registry and the other covariates in the table.

tween age and specificity in women with fatty breasts ( $P > 0.2$ ).

## DISCUSSION

This study of pooled mammography registry data revealed several important findings. First, we observed that both breast density and age were important independent predictors of the sensitivity and specificity of screening mammography. Sensitivity and specificity were highest in older women, especially those older than 80 years of age, for whom sensitivity achieves the target of 85% set by the

**Table 4. Adjusted Sensitivity of Mammography Based on 2223 Women with Incident Breast Cancer, by Age, Breast Density, and Current Hormone Replacement Therapy Use\***

Variable	Adjusted Sensitivity (95% CI), %
Age group	
40–44 y	68.6 (60.2–75.9)
45–49 y	72.5 (66.4–77.8)
50–54 y	75.4 (70.7–79.5)
55–59 y	75.1 (70.3–79.4)
60–69 y	72.8 (69.1–76.1)
70–79 y	78.6 (74.6–82.1)
80–89 y	83.3 (74.9–89.4)
Breast density group	
Almost entirely fatty	87.0 (79.0–92.3)
Scattered fibroglandular tissue	81.5 (78.9–83.9)
Heterogeneously dense	69.4 (66.4–72.2)
Extremely dense	62.9 (55.8–69.4)
Current use of HRT	
No	76.0 (73.6–78.3)
Yes	73.3 (70.3–76.1)

\* Model includes age, HRT use, breast density, and mammography registry. HRT = hormone replacement therapy.

Agency for Healthcare Research and Quality (AHRQ) (formerly the Agency for Health Care Policy and Research) in 1994 (29). Mammography is most effective in detecting cancer in this group of women. The accuracy of mammography in women 50 to 79 years of age was lower than that recommended by AHRQ. Sensitivity and specificity were lowest in younger women with radiographically dense breasts who used HRT. In addition, we found that after adjustment for breast density, HRT use was not significantly associated with sensitivity. These findings suggest that the overall decrease in sensitivity associated with HRT use is due to increases in radiographic breast density that occur in some women who use HRT. Women with dense breasts, regardless of whether they use HRT, should be informed that mammography will be less likely to detect breast cancer.

Women taking HRT should be made aware that HRT may increase breast density, which could result in the need for additional imaging or breast ultrasonography (4, 5, 18). Studies show that HRT users have higher rates of additional imaging (30). Women and their health care providers may want to ensure that a routine mammographic examination is performed before starting HRT. In addition, mammography reports should include an assessment of breast density. This information allows women increased understanding of the level of detection mammography offers as well as how mammographic breast density may increase their risk for breast cancer.

Should women consider stopping HRT before a mammographic examination to allow for the best imaging of the breast? This question cannot be answered because we do not fully understand the effects of stopping and restart-

ing HRT on other health risks or the side effects that may be experienced. Randomized trials show that increases in myocardial infarction due to HRT use are highest during the first year of use (31). In addition, it is not clear how long HRT should be discontinued to allow for optimal imaging. These questions require additional study.

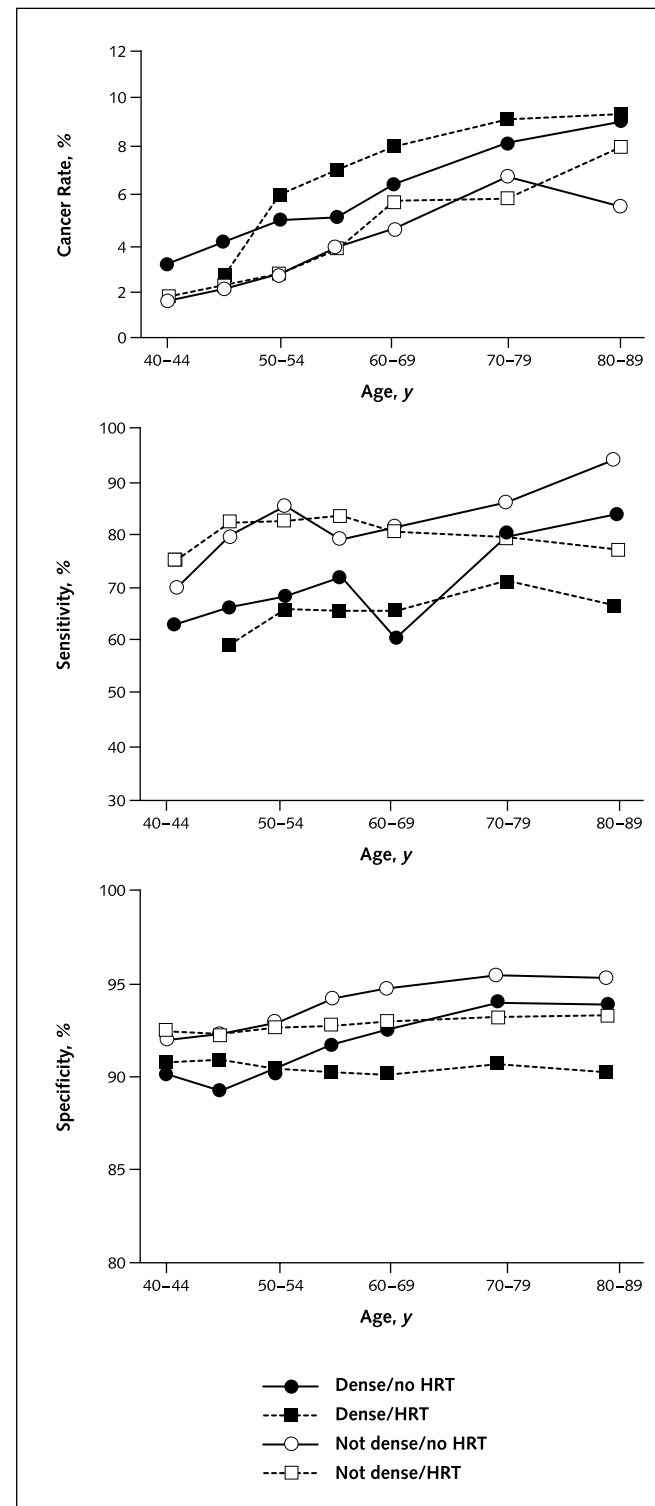
When examining how breast density, age, and HRT use independently affect specificity, we found that age and HRT use were independent predictors of accuracy but their effect is not as striking as that of breast density. In our multivariable model, we found that only age was associated with specificity for women not using HRT; specificity was higher in older women who do not use HRT compared with younger nonusers. In contrast, specificity is similar across all ages for all women using HRT. Not unexpectedly, we found that specificity significantly decreases as breast density increases and that additional imaging may be needed to evaluate a positive finding. This means that for every 1000 women with radiographically dense breasts who undergo screening, 100 will have false-positive examinations; for every 1000 women with fatty breasts who undergo screening, 35 will have false-positive examinations.

Because we could focus on screening in younger women, we believe that our study makes an important contribution to understanding the accuracy of screening mammography. In these women, we found some inconsistencies in the effect of HRT. In contrast to older women, women between 40 and 49 years of age who used HRT had lower sensitivities and higher specificities than women who did not use HRT. The higher proportion of postmenopausal women among younger women using HRT explains this apparent inconsistency. This subgroup of younger women with early menopause probably had less endogenous estrogen and, thus, decreased breast density. Additional strengths of our study are that data reflect current mammography techniques, which have changed as a result of the Mammography Quality Standards Act, and that ductal carcinoma in situ (with its recent increased incidence) was included in our analysis (32).

Our analyses may have several limitations. To evaluate screening mammography across diverse practice settings in the United States, the BCSC data are collected on a voluntary basis during routine delivery of care. Therefore, standardization of data collection is not as complete as it is in clinical trials, de novo observation studies, or other studies that administer surveys unconstrained by the need to fit within routine clinical care. This problem is reflected in the varying amounts of missing data across facilities and regions. However, given the similarity in overall accuracy between the subset of women excluded because data were missing and women included in our study, we do not believe that missing data substantially biased our results.

We also acknowledge that because we included women who had mammography within 2 years of the index examination, prevalent cancers were excluded from the analysis, which may affect the accuracy indices we studied.

**Figure.** Mean cancer rates, sensitivity, and specificity, by age, breast density, and current hormone replacement therapy (HRT) use.



Sensitivity and cancer rates were not plotted for women between 40 and 44 years of age with dense breasts who used HRT because there were only two women with cancer in this group.



**Table 5. Adjusted Specificity of Mammography in 461 449 Examinations with No Cancer Diagnosis within the Follow-up Period, by Age, Current Hormone Replacement Therapy Use, and Breast Density\***

Age Group	Adjusted Specificity by Breast Density (95% CI)				Adjusted Specificity by Current Use of HRT (95% CI)	
	Almost Entirely Fatty	Scattered Fibroglandular Tissue	Heterogeneously Dense	Extremely Dense	Yes	No
y	← % →					
40–44	95.8 (94.8–96.5)	91.8 (91.3–92.2)	90.2 (89.7–90.6)	90.3 (89.6–91.0)	91.6 (90.0–92.7)	91.4 (90.1–92.4)
45–49	96.6 (95.9–97.1)	91.9 (91.6–92.2)	90.0 (89.7–90.3)	89.1 (88.5–89.7)	91.9 (91.8–94.6)	91.2 (91.1–94.1)
50–54	96.1 (95.6–96.6)	92.6 (92.3–92.8)	90.5 (90.1–90.8)	90.1 (89.3–90.7)	91.9 (90.9–93.0)	92.0 (91.0–93.1)
55–59	96.8 (96.4–97.2)	93.0 (92.8–93.3)	90.8 (90.4–91.2)	91.6 (90.7–92.4)	91.5 (91.2–93.8)	93.1 (92.8–95.0)
60–69	96.6 (96.3–96.9)	93.5 (93.3–93.7)	91.2 (90.9–91.5)	91.4 (90.5–92.2)	91.4 (90.5–92.9)	93.6 (92.9–94.7)
70–79	96.9 (96.5–97.2)	93.9 (93.7–94.2)	92.4 (92.0–92.8)	93.2 (92.1–94.1)	91.9 (90.5–93.2)	94.6 (93.6–95.4)
80–89	96.4 (95.5–97.2)	93.8 (93.2–94.4)	92.4 (91.6–93.2)	92.4 (90.1–94.2)	91.8 (88.2–93.0)	94.4 (92.0–95.1)

\* Model includes age, HRT, breast density, and mammography registry. The specificities for age are shown by HRT status and breast density because there are significant interactions between age and HRT use and between age and breast density. HRT = hormone replacement therapy.

However, this reflects the reality of screening practices. In addition, we focused our analysis on recently screened women who represent most middle-aged to older women in the United States today. Finally, additional studies by the BCSC have revealed that radiologists' use of the BI-RADS categories for mammographic assessment is sometimes inconsistent with BI-RADS recommendation categories (28). These inconsistencies may influence the classification of mammography examinations as positive or negative and may ultimately influence the calculations of accuracy indices (33). We believe, however, that these inconsistencies had very little, if any, impact on our analyses.

In conclusion, we found that mammographic breast density, HRT use, and age were all important independent predictors of the accuracy of screening mammography. Age and breast density were the most important predictors. After adjustment for age and breast density, HRT use was not significantly associated with the sensitivity of screening mammography. However, because HRT use has been shown to increase breast density in many women, HRT can influence sensitivity through its effects on breast density.

Another clinical implication of our study is that all mammography reports should routinely include a statement about breast density. This information would prompt clinicians to inform women that breast density is influenced by HRT use and that it can alter mammographic interpretation.

It should be recognized that the BCSC provides an invaluable resource for the study of breast cancer surveillance. Our analyses could not have been conducted without the very large pooled samples provided by the BCSC, which allowed the study of the independent effect of factors influencing breast density.

From Dartmouth Medical School, Lebanon, New Hampshire; Group Health Cooperative, Seattle, Washington; University of North Carolina, Chapel Hill, North Carolina; Veterans Affairs Medical Center, San Francisco, California; University of New Mexico, Albuquerque, New Mexico; University of Vermont, Burlington, Vermont; AMC Cancer Research

Center, Denver, Colorado; and National Cancer Institute, Bethesda, Maryland.

**Grant Support:** By cooperative agreements UO1CA63731, UO1CA63736, UO1CA63740, UO1CA69976, UO1CA70013, UO1CA70040, UO1CA86076, UO1CA86082, and R01CA80888 from the National Cancer Institute as part of the National Cancer Institute's Breast Cancer Surveillance Consortium.

**Requests for Single Reprints:** Patricia A. Carney, PhD, Department of Community & Family Medicine, Dartmouth Medical School, 1 Medical Center Drive, HB 7925, Lebanon, NH 03756; e-mail, Patricia.A.Carney@dartmouth.edu.

Current author addresses and author contributions are available at [www.annals.org](http://www.annals.org).

## References

- Byrne C. Studying mammographic density: implications for understanding breast cancer [Editorial]. *J Natl Cancer Inst*. 1997;89:531-3. [PMID: 9106636]
- Boyd NF, Lockwood GA, Byng JW, Tritchler DL, Yaffe MJ. Mammographic densities and breast cancer risk. *Cancer Epidemiol Biomarkers Prev*. 1998;7:1133-44. [PMID: 9865433]
- Lam PB, Vacek PM, Geller BM, Muss HB. The association of increased weight, body mass index, and tissue density with the risk of breast carcinoma in Vermont. *Cancer*. 2000;89:369-75. [PMID: 10918168]
- Persson I, Thurfjell E, Holmberg L. Effect of estrogen and estrogen-progestin replacement regimens on mammographic breast parenchymal density. *J Clin Oncol*. 1997;15:3201-7. [PMID 9336356]
- Litherland JC, Evans AJ, Wilson AR. The effect of hormone replacement therapy on recall rate in the National Health Service Breast Screening Programme. *Clin Radiol*. 1997;52:276-9. [PMID: 9112944]
- U.S. Preventive Services Task Force. Guide to Clinical Preventive Services. 2nd ed. Washington, DC: U.S. Department of Health and Human Services; 1996.
- Coveney EC, Geraghty JG, O'Laoide R, Hourihane JB, O'Higgins NJ. Reasons underlying negative mammography in patients with palpable breast cancer. *Clin Radiol*. 1994;49:123-5. [PMID: 8124890]
- Kerlikowske K, Grady D, Barclay J, Sickles EA, Ernster V. Effect of age, breast density, and family history on the sensitivity of first screening mammography. *JAMA*. 1996;276:33-8. [PMID: 8667536]
- Harvey SC, DiPiro PJ, Meyer JE. Marked regression of a nonpalpable breast cancer after cessation of hormone replacement therapy. *AJR Am J Roentgenol*. 1996;167:394-5. [PMID: 8686614]

10. Laya MB, Larson EB, Taplin SH, White E. Effect of estrogen replacement therapy on the specificity and sensitivity of screening mammography. *J Natl Cancer Inst.* 1996;88:643-9. [PMID: 8627640]
11. Thurffell EL, Holmberg LH, Persson IR. Screening mammography: sensitivity and specificity in relation to hormone replacement therapy. *Radiology.* 1997;203:339-41. [PMID: 9114085]
12. Yankaskas BC, Cleveland RJ, Schell MJ, Kozar R. Association of recall rates with sensitivity and positive predictive values of screening mammography. *AJR Am J Roentgenol.* 2001;177:543-9. [PMID: 11517044]
13. Tabar L, Fagerberg G, Chen HH, Duffy SW, Smart CR, Gad A, et al. Efficacy of breast cancer screening by age. New results from the Swedish Two-County Trial. *Cancer.* 1995;75:2507-17. [PMID: 7736395]
14. Leung W, Goldberg F, Zee B, Sterns E. Mammographic density in women on postmenopausal hormone replacement therapy. *Surgery.* 1997;122:669-73; discussion 673-4. [PMID: 9347841]
15. Pankow JS, Vachon CM, Kuni CC, King RA, Arnett DK, Grabrick DM, et al. Genetic analysis of mammographic breast density in adult women: evidence of a gene effect. *J Natl Cancer Inst.* 1997;89:549-56. [PMID: 9106643]
16. Saftlas AF, Hoover RN, Brinton LA, Szklo M, Olson DR, Salane M, et al. Mammographic densities and risk of breast cancer. *Cancer.* 1991;67:2833-8. [PMID: 2025849]
17. Boyd NF, Byng JW, Jong RA, Fishell EK, Little LE, Miller AB, et al. Quantitative classification of mammographic densities and breast cancer risk: results from the Canadian National Breast Screening Study. *J Natl Cancer Inst.* 1995;87:670-5. [PMID: 7752271]
18. Steinberg KK, Thacker SB, Smith SJ, Stroup DF, Zack MM, Flanders WD, et al. A meta-analysis of the effect of estrogen replacement therapy on the risk of breast cancer. *JAMA.* 1991;265:1985-90. [PMID: 1826136]
19. White E, Velentgas P, Mandelson MT, Lehman CD, Elmore JG, Porter P, et al. Variation in mammographic breast density by time in menstrual cycle among women aged 40-49 years. *J Natl Cancer Inst.* 1998;90:906-10. [PMID: 9637139]
20. Baines CJ, Vidmar M, McKeown-Eyssen G, Tibshirani R. Impact of menstrual phase on false-negative mammograms in the Canadian National Breast Screening Study. *Cancer.* 1997;80:720-4. [PMID: 9264355]
21. Mandelson MT, Oestreicher N, Porter PL, White D, Finder CA, Taplin SH, et al. Breast density as a predictor of mammographic detection: comparison of interval- and screen-detected cancers. *J Natl Cancer Inst.* 2000;92:1081-7. [PMID: 10880551]
22. Morrison AS, Brisson J, Khalid N. Breast cancer incidence and mortality in the breast cancer detection demonstration project. *J Natl Cancer Inst.* 1988;80:1540-7. [PMID: 3193469]
23. Shapiro S. Periodic screening for breast cancer: the HIP Randomized Controlled Trial. *Health Insurance Plan. J Natl Cancer Inst Monogr.* 1997;27-30 [PMID: 9709271].
24. Houn F, Elliott ML, McCrohan JL. The Mammography Quality Standards Act of 1992. History and philosophy. *Radiol Clin North Am.* 1995;33:1059-65. [PMID: 7480655]
25. American College of Radiology (ACR) Illustrated Breast Imaging Reporting and Data System (BI-RADS™). 3rd ed. Reston, VA: American College of Radiology; 1998.
26. Ballard-Barbash R, Taplin SH, Yankaskas BC, Ernster VL, Rosenberg RD, Carney PA, et al. Breast Cancer Surveillance Consortium: a national mammography screening and outcomes database. *AJR Am J Roentgenol.* 1997;169:1001-8. [PMID: 9308451]
27. Carney PA, Geller BM, Moffett H, Ganger M, Sewell M, Barlow WE, et al. Current medicolegal and confidentiality issues in large, multicenter research programs. *Am J Epidemiol.* 2000;152:371-8. [PMID: 10968382]
28. Taplin SH, Ichikawa LE, Kerlikowske K, Ernster VL, Rosenberg RD, Yankaskas BC, et al. Concordance of breast imaging reporting and data system assessments and management recommendations in screening mammography. *Radiology.* 2002;222:529-35. [PMID: 11818624]
29. Quality Determinants of Mammography. Clinical Practice Guideline Number 13: Rockville, MD: U.S Department of Health and Human Services, Agency for Health Care Policy and Research; 1994. AHCPR publication 95-0632.
30. Rosenberg RD, Hunt WC, Williamson MR, Gilliland FD, Wiest PW, Kelsey CA, et al. Effects of age, breast density, ethnicity, and estrogen replacement therapy on screening mammographic sensitivity and cancer stage at diagnosis: review of 183,134 screening mammograms in Albuquerque, New Mexico. *Radiology.* 1998;209:511-8. [PMID: 9807581]
31. Hulley S, Grady D, Bush T, Furberg C, Herrington D, Riggs B, et al. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. Heart and Estrogen/progestin Replacement Study (HERS) Research Group. *JAMA.* 1998;280:605-13. [PMID: 9718051]
32. Ernster VL, Barclay J. Increases in ductal carcinoma in situ (DCIS) of the breast in relation to mammography: a dilemma. *J Natl Cancer Inst Monogr.* 1997;151-6. [PMID: 9709292]
33. Rosenberg RD, Yankaskas BC, Hunt WC, Ballard-Barbash R, Urban N, Ernster VL, et al. Effect of variations in operational definitions on performance estimates for screening mammography. *Acad Radiol.* 2000;7:1058-68. [PMID: 11131050]

**Current Author Addresses:** Dr. Carney: Department of Community & Family Medicine, Dartmouth Medical School, 1 Medical Center Drive, HB 7925, Lebanon, NH 03756.

Drs. Miglioretti, Rutter, Abraham, and Taplin: Center for Health Studies, Group Health Cooperative, Suite 1600, 1730 Minor Avenue, Seattle, WA 98101-1448.

Dr. Yankaskas: Department of Radiology, CB #7515, MRI, University of North Carolina, 106 Mason Farm Road, Chapel Hill, NC 27599-7515.

Dr. Kerlikowske: Department of Medicine and Epidemiology & Biostatistics, Suite 600, University of California, San Francisco, 74 New Montgomery, San Francisco, CA 94105.

Dr. Rosenberg: Department of Radiology—Health Sciences Center, University of New Mexico, Albuquerque, NM 87131.

Dr. Geller: University of Vermont, Health Promotions Research, Vermont Cancer Center, One South Prospect Street, Burlington, VT 05401-3498.

Dr. Dignan: Kentucky Prevention Research Center, 2365 Harrodsburg Road, Suite B100, Lexington, KY 40504-3381.

Dr. Cutter: University of Nevada at Reno Applied Research Facility, Mail Stop 199, Reno, NV 89557.

Dr. Ballard-Barbash, Applied Research Program—EPN 4005, National Cancer Institute, 6130 Executive Boulevard, Bethesda, MD 20892-7344.

**Author Contributions:** Conception and design: P.A. Carney, B.C. Yankaskas, K. Kerlikowske, S.H. Taplin, M. Dignan, R. Ballard-Barbash.

Analysis and interpretation of the data: P.A. Carney, D.L. Miglioretti, B.C. Yankaskas, K. Kerlikowske, L.A. Abraham, G. Cutter, R. Ballard-Barbash.

Drafting of the article: P.A. Carney, D.L. Miglioretti, K. Kerlikowske, L.A. Abraham, M. Dignan, R. Ballard-Barbash.

Critical revision of the article for important intellectual content: P.A. Carney, D.L. Miglioretti, B.C. Yankaskas, K. Kerlikowske, C.M. Rutter, B.M. Geller, S.H. Taplin, M. Dignan, G. Cutter, R. Ballard-Barbash.

Final approval of the article: P.A. Carney, D.L. Miglioretti, B.C. Yankaskas, K. Kerlikowske, C.M. Rutter, B.M. Geller, L.A. Abraham, S.H. Taplin, M. Dignan, G. Cutter.

Provision of study materials or patients: P.A. Carney, D.L. Miglioretti, B.C. Yankaskas, K. Kerlikowske, R. Rosenberg, S.H. Taplin, M. Dignan, Statistical expertise: D.L. Miglioretti, C.M. Rutter, L.A. Abraham, G. Cutter.

Obtaining of funding: P.A. Carney, B.C. Yankaskas, K. Kerlikowske, R. Rosenberg, G. Cutter, B.M. Geller, S.H. Taplin.

Administrative, technical, or logistic support: P.A. Carney.

Collection and assembly of data: P.A. Carney, B.C. Yankaskas, K. Kerlikowske, R. Rosenberg, B.M. Geller, and M. Dignan.

**Study Guide**  
**Session 3**  
**Determining the Value of a Diagnostic Test**  
**Allen F. Shaughnessy, PharmD, MMedEd**  
**Allen.Shaughnessy@tufts.edu**

**The aims of this session are to:**

1. Review and expand upon some ideas of disease screening and testing that have been covered in the epidemiology;
2. Introduce the concept of the difference between the technical precision and the clinical precision of a screening or diagnostic test;
3. Practice how to quickly and accurately evaluate studies evaluating a new diagnostic or screening test to determine their validity

**Specific Objectives:** By completing the initial reading and participating in class, students should be able to:

1. Explain the different ways the same testing procedure can be used;
2. Use the test characteristics of sensitivity, specificity, predictive values, and likelihood ratios to correctly interpret test results
3. Explain how pre-test probability can affect test accuracy
4. Determine whether the results of a study evaluating a test are relevant and valid and support that test's use in clinical practice.
5. Explain why the fact that simply using a test because it is highly sensitive or specific may not always be beneficial.

This study guide provides an outline of the concepts necessary to meet these objectives. It contains hyperlinks to short videos, web pages, and articles that explain the concepts in other ways or in greater detail. You can follow these hyperlinks if the explanations and examples I've given you are not sufficient to help you understand and to help you complete the readiness assessment test.

## **Purposes of testing**

A test, whether a physical exam maneuver, a laboratory test, an imaging study, or a performance measure (such as reading an eye chart), can play various roles, depending on our needs. The same test can:

- 1) **Screen** for an unapparent disease in an asymptomatic individual (*screening*)
- 2) **Identify** a disease in a patient suspected of having that disease (*case finding*)
- 3) **Confirm** or **refute** results of another test (*confirmation*)
- 4) **Evaluate** the effectiveness of treatment (*monitoring*)
- 5) Allow estimation of **prognosis** to help guide treatment decisions (*prognosis*)



# Example

## *Same test, different uses*

Measuring someone's blood pressure can be used for various purposes, depending on the setting and the need.

- 1) Screening: Checking blood pressure at a health fair or at a self-testing station – asymptomatic patients with low likelihood of hypertension;
- 2) Case finding: Checking patients' blood pressure at the start of an office visit – patients are still at low risk of hypertension (*unless already diagnosed*), but are at higher risk than when screened (as in example #1), since they are seeking health care
- 3) Confirmation: Checking someone's blood pressure who has had a previously high reading via screening or in the office;
- 4) Monitoring: Checking blood pressure in someone with diagnosed hypertension; and,
- 5) Prognosis: Checking blood pressure in patients with acute myocardial infarction or stroke can be used to estimate 30-day mortality.

## So What? Bayes' Theorem

Tests are not perfect – all tests have some risk of false positive and false negative results.\* These false results will vary based on the likelihood of the test being positive or negative *before we ever do the test*. This likelihood is called *prior probability*, *pre-test probability*, or *prevalence* of disease.

Simply put, [Bayes' theorem](#) is:

post-test probability = Pre-test probability, given the test result

The effect of pre-test probability on test results is intuitive: if someone is highly unlikely to have a disease but a test is positive, it makes sense to think that the test is wrong (i.e. still has a low post-test probability). Conversely, if another person has dramatic symptoms and signs of a disease but the test comes back negative, it makes sense that the test is a false negative (i.e. still has a high post-test probability).



For more information, see: [Bayes' Theorem and Breast Cancer](#) (9:56)

---

\* Except, perhaps, the “birth test” and the “death test”

For example,

- In a 70-year-old male with sudden onset chest pain that worsens with exercise, the likelihood of myocardial infarction, before additional testing, is 63%. A negative electrocardiogram would be **falsely negative** in 1-in-6 men like this patient.
- In a 40-year-old male with sudden onset chest pain that worsens with exercise, the likelihood of myocardial infarction, before additional testing, is around 0.6%.<sup>†</sup> A positive electrocardiogram would be **falsely positive** in 11 of 12 men like this patient.<sup>‡</sup>

Pretest probability is determined a number of ways—sometimes, we have done epidemiologic studies to find the population prevalence of certain diseases. Sometimes, the pretest probability is a best-guess estimate because we don't always have that kind of data on every condition.

So, when screening for disease, the pre-test probability of whatever we are screening for will be low, making positive test results suspect. Case-finding increases the probability somewhat, and the factors that prompt testing to confirm a diagnosis increase probability even more. *The key point is that the same test will perform differently given this background probability.*

We can quantify how likely a test is to be falsely positive using simple calculations. What is *not* intuitive for most people, though, is how likely a test is likely to be false in situations of low pre-test likelihood.

# Example

## *Some examples of high false positives*

1. In the U.S., lyme disease prevalence varies by geography, highest in Connecticut and lowest in Texas.

False positive results:

- Connecticut (20% prevalence): 17% false positive
- Texas (2% prevalence): 72% false positive

2. The prevalence of breast cancer in women increases with age. For a 30-year-old woman undergoing a screening mammography:

- Prevalence: 1 in 235 (0.43%)
- False positive rate: 94%

<sup>†</sup> [Ruling out coronary artery disease in primary care: development and validation of a simple prediction rule.](#)

<sup>‡</sup> [Acute chest pain in the emergency room. Identification and examination of low-risk patients.](#)

## Why is this important?

### *1. Hazards of testing*

Patients who have a positive test result, even if it is found later to be falsely positive, can have lasting psychological results:

[Three years after a false-positive mammogram result](#), women are more anxious about having breast cancer and have prolonged psychosocial effects such as, “my sense of well-being is less” and “my relationship with other people is worse”

[In a study of infants diagnosed with jaundice](#) (for which the evidence is inconclusive as to whether they benefit from diagnosis and treatment):

- Mothers are more likely to completely stop breast feeding
- Mothers were more likely to have never left their baby alone with anyone else, including the father
- The infant had more office visits and emergency department visits

This result has been called the “[vulnerable child syndrome](#)” that occurs, in this case, as a result of testing that may not produce benefit.

### *2. Lack of benefit of testing*

Physicians commonly express the belief that patients want diagnostic testing to check for serious but unlikely illnesses. This type of testing is often ordered with the aim of reassuring patients. However, “Diagnostic tests for symptoms with a low risk of serious illness do little to reassure patients, decrease their anxiety, or resolve their symptoms, although the tests may reduce further primary care visits.”

[JAMA Intern Med. 2013;173\(6\):407-416.](#)

## Technical vs. clinical precision of a test

*Sensitivity* and *specificity* of a test are characteristics used to judge the intrinsic technical precision of a test. They, for the most part, are insensitive to changes in prevalence.

*Positive predictive value* and *negative predictive value* are characteristics used to judge the clinical performance of a test, i.e., how well does the test represent the truth in clinical practice? Predictive values *are* sensitive to prevalence. As illustrated above, the same test, with the same sensitivity and specificity, will result in different rates of false positive and false negative results, depending on the pre-test likelihood of a positive or negative result.

In the above examples, the technical precision of the tests look pretty good:

### *Lyme disease detection*

Sensitivity: 95%

Specificity: 95%

### *Breast cancer detection (mammography)*

Sensitivity: 79%

Specificity: 89%

However, the effect of prevalence makes the tests perform much worse, in clinical practice, than is intuitively obvious.

### *Lyme disease positive predictive value*

High prevalence: 83%

Low prevalence: 28%

### *Mammography positive predictive value*

High prevalence: 34%

Low prevalence: 0.8%

In other words, in a place where Lyme disease is highly prevalent, 83% of the time, when the test is positive, the patient actually has the disease. Another way to interpret this is to say that in a high prevalence area, there is a 17% false-positive rate. In a low prevalence area, when the test is positive, 28% of those patients will truly have Lyme disease.

As a result, even though the sensitivity and specificity of these tests are high, the tests will be falsely positive a majority of the time when the prevalence of disease is low. In the case of Lyme disease, about 3 out of 4 people told they have Lyme disease will not; and 92 out of 100 low likelihood women with a finding on mammogram will not have breast cancer.



Watch [Sensitivity and Specificity – getting a feel \(8:14\)](#)

## Calculating Sensitivity, Specificity, and Predictive Values

There are several ways to calculate test characteristics, though test characteristics typically are calculated using a 2x2 table, listing the true status of the disease at the top and the test results along the side. The key to setting up the table is **to list the disease at the top of the square**.

	Disease truly present	Disease truly absent
Positive test	True Positives (TP) a	False Positives (FP) b
Negative test	c False Negatives (FN)	d True Negatives (TN)

### Sensitivity

- Is the percent of patients *with the disease* who have a *positive test*
- =  $TP / (TP + FN)$
- =  $a / (a + c)$

### Specificity

- Is the percent of patients *without the disease* who have a *negative test*
- =  $TN / (TN + FP)$
- =  $d / (d + b)$

Using the graph, the calculations proceed “down” and “up” the columns:

	Disease truly present	Disease truly absent
Positive test	True Positives (a)	False Positives (b)
Negative test	False Negatives (c)	True Negatives (d)

### Positive predictive value

- Is the percent of patients *with a positive test* who *have the disease*
- =  $TP / (TP + FP)$
- =  $a / (a + b)$

### Negative predictive value

- Is the percent of patients *with a negative test* who *do not have the disease*
- =  $TN / (TN + FN)$
- =  $d / (d + c)$

Using the graph, the calculations proceed “left” and “right” across the columns:

	Disease truly present	Disease truly absent
Positive test	True Positives (a)	False Positives (b)
Negative test	False Negatives (c)	True Negatives (d)



Watch “[Sensitivity, specificity, and predictive values](#)” (2:34)

## Practice:

Here are the results of a test for HIV.

	<b>HIV present</b>	<b>HIV absent</b>
<b>HIV test +</b>	475	4975
<b>HIV test -</b>	25	94525

Calculate the test characteristics:

Sensitivity:                      Positive Predictive Value:

Specificity:                      Negative Predictive Value:

Answers [here](#)

## Practice: Sample USMLE Step 1 question:

To protect blood supplies from contamination, screening for all donors for hepatitis C is required. The screening test has a sensitivity of 95% and a specificity of 90% and is used on a sample of donors in which 10% are known to have hepatitis C infection. Which of the following is the best estimate of the chance that a donor who tests negative is actually free of infection?

- A. 45%
- B. 50%
- C. 85%
- D. 90%
- E. 95%
- F. 99%



Calculations shown at: [Calculating NPV from sensitivity and specificity](#)  
(4:11)



## Using test characteristics: [SnNout and SpPin](#)

Tests that are highly sensitive are very good at identifying patients with disease. As a result of this quality, we can be sure that *negative* tests are truly negative. This relationship can be remembered using the mnemonic *SnNout*:

**SnNout:** If a test is highly **Sensitive**, and the test result is **Negative**, we can rule **out** the disease

Conversely, tests that are highly specific are negative unless patients truly have the disease. As a result, we can be sure that positive tests are truly positive. This relationship can be remembered using the mnemonic *SpPin*:

**SpPin:** If a test is highly **Specific**, and the test result is **Positive**, we can rule **in** the disease.

This rule holds when the likelihood is not too low (e.g., screening) or too high (e.g., confirmatory testing).



Watch [Using SpPin and SnNout](#) (2:23)

## Using test characteristics: [Predictive values](#)

The **good** thing about predictive values *is that they vary with prevalence of disease*. By knowing the prevalence, we can calculate the predictive value of a test for individual situations, helping us to make decisions regarding treatment or further testing.

The **bad** thing about predictive values is the same: *they vary with prevalence of disease*. It is often hard to calculate the pre-test likelihood of disease, especially in the moment, and we find it hard to find a resource that lists prevalence. Fortunately, there are calculators that can provide estimates of prevalence based on physical findings and calculate predictive values based on the test's sensitivity and specificity.

[Here](#) is an example of a clinical calculator that helps determine the pre-test probability of patients having sore throat based on symptoms.

## Combining all of the above: [Likelihood ratios](#)

The likelihood ratio gives us an understanding of how strongly a test result helps us rule in or rule out a disease. A **positive likelihood ratio** compares the likelihood that someone with the disease in question has a positive test as compared with someone who doesn't have the disease. A negative likelihood **ratio** compares the likelihood that someone without the disease in question will have a negative test as compared with someone who does have the disease.

Likelihood ratios are calculated based on sensitivity and specificity of the test:

Positive Likelihood Ratio = sensitivity / (1-specificity)

Negative Likelihood Ratio = (1-sensitivity) / specificity

In practice, likelihood ratios (LRs) are used in two ways.

- 1) To calculate post-test odds of a disease, given pre-test odds and the LR:

Pretest odds of disease X LR = Post-test odds of disease

Most of us, however, don't think in terms of odds, which makes the calculations difficult. However, many calculators are available that will take pre-test probability, convert it to pre-test odds, calculate the post-test odds from the LR, and convert this odds back into a probability. There are also [paper](#) or [computer-based](#) nomograms that will make the calculations easy.

- 2) To give a general interpretation of a test's quality. The size of a LR helps us to understand how valid a test result might be. General rules:

Likelihood ratio	Interpretation
>10	Good test to rule-in disease with a positive result
5 - 10	Moderately able to rule-in with a positive test
2 - 5	Small increase in probability with a positive test
1-2	No change in probability with a positive test
0.5 - 1	No change in probability with a negative test
0.2 - 0.5	Small increase in probability with a negative test
0.1 - 0.2	Moderately able to rule-out disease with a negative test
< 0.1	Good test to rule-out disease with a negative result

These are good, general rules, for gauging the quality of a diagnostic test. However, because prevalence still impacts a test's results, these rules don't apply if the pre-test probability is very high or very low.

For example, a test with a positive LR of 500 will only have a positive predictive value of 50% given a pre-test probability of 1 in 5,000.

## Beyond Test Characteristics: What Makes a Test Truly Helpful?

As we've discussed, tests can identify unknown disease, confirm presumed disease, help with estimates of prognosis, or monitor response to treatment. But can a test be held responsible to do more?

### Limits of Testing:

*Screening:* Early<sup>4</sup> identification of disease is only beneficial if treating before symptoms occur results in greater benefit than treating based on symptoms.



Watch [Overdiagnosed](#). (it's funny and informative, but long; over an hour)

*Diagnostic Testing:* A diagnosis, though the *sine qua non* of medicine, is, at its essence only a label placed on a patient that is useful when selecting the right treatment. Therefore, testing leading to a diagnosis is **only** helpful if it leads to a change in *treatment*.

### A Hierarchy of Evidence Regarding Tests<sup>5</sup>

As a result of these limitations, it's not enough to simply say that a test has a good sensitivity and specificity (or predictive values). We need our tests to do more:

*Basic Criteria:* Is the test sufficiently sensitive and specific?

We have many tests with low sensitivity and specificity.

*Minimally useful:* The test changes diagnosis

As a result of a positive test, we now have a label to put on a patient. That doesn't mean that we've done anything other than categorize their set of signs and symptoms. What's better is to . . .

*More useful:* The test changes treatment

A test that results in changes in treatment is a good start. However, tests don't always lead to changes. Sputum samples are often suggested in guidelines of the treatment of pneumonia. However, research has shown physicians frequently do not change treatment when the culture results are known a day or two later.

*Very useful:* The test changes outcomes

Routine monitoring (with A1c) did not affect outcomes in the United Kingdom Prospective Diabetes Study.<sup>6</sup>

---

<sup>4</sup> That is, before symptoms appear.

<sup>5</sup> Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Med Decis Making* 1991; 11:88-94

<sup>6</sup> Turner RC, et al. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). *Lancet* 1998;352:837-53.

On the other hand, testing in the emergency department to determine whether the patient does or doesn't have heart failure has been shown to decrease admissions, decrease length of stay, and speed the initiation of appropriate therapy.<sup>7</sup>

*Maximum benefit:* The test is worthwhile to patients and/or society.

Screening newborns for congenital hypothyroidism, phenylketonuria, and other diseases (but not all) results in early treatment that permanently alters the life of these children. On the other hand, screening for other diseases, and treating them (kernicterus, oxygenation) has not shown to be beneficial.

---

McCormack J, Greenhalgh T. Seeing what you want to see in randomised controlled trials: versions and perversions of UKPDS data. *BMJ* 2000;320:1720-1723)

<sup>7</sup> Mueller C, Scholer A, Laule-Kilian K, et al. Use of B-type natriuretic peptide in the evaluation and management of acute dyspnea. *N Engl J Med* 2004; 350: 647-54).

[\(jump to worksheet\)](#)

## **Evaluating a Study about a New Diagnostic Test**

### **Is the new test reasonable? What are its limitations?**

The first step in evaluating a new screening test, diagnostic test, or other diagnostic maneuver is to determine whether the test could be used in clinical practice. A test that is too expensive, takes too long to perform, requires too much blood, or other limitation may not be worth considering.

#### *Example*

Now there is a rapid test for influenza, but in years past an influenza test had to be sent to a central laboratory. Results were returned in 5 to 7 days and were not useful for making decisions regarding treatment. Influenza testing was only useful after the fact for tracking patterns.

### **Is the reference (gold) standard appropriate?**

A new test has to be compared to an existing standard. This reference standard, or “gold standard,” should be the best currently available way of identifying disease. Lesser standards, with their own limitations, can confound the results.

#### *Example*

Some years ago, I conducted a [study](#) to determine whether two methods of skin caliper measurement and bioimpedence measurement were accurate measures of body fat; we wanted to determine the best way to identify the minimum weight for high school wrestlers. The gold standard for body fat analysis is total body immersion in water; we didn’t have a facility to do this, which dramatically limited our ability to draw conclusions regarding the tests’ accuracy.

### **Did all participants receive both the new test and the reference test?**

All study subjects should receive both the new test and the gold standard test. Sometimes testing is done so that only patients with a positive result on the new test get the gold standard test, or vice versa. This approach biases the study.

### **Were the results of the test interpreted without knowledge (blinded) of the reference test result and vice versa?**

We want to assure that both tests are interpreted independently, that is, without knowledge of the results of the other test. This approach prevents interpretation bias.

#### *Example*

Imagine checking the blood pressure in a patient just after someone else has. If you knew the first reading, you would expect your reading to be similar, and you might try a little harder to get a very similar. If, on the other hand, you don’t know the results of the first blood pressure measurement, you will not have an inherent bias toward that result.

## **Were the patients enrolled randomly or consecutively?**

Ideally, all patients eligible for testing would be enrolled; if not, we would like to see a random selection of patient. This criterion helps to assure a broad spectrum of patients, some with relatively mild disease and others with more severe disease, which gives us a more accurate understanding of the test's characteristics.

### *Example*

Investigators conducted an early study of a test to determine whether patients with shortness of breath presenting to an emergency department have heart failure. The study was conducted during daylight hours (because that is when the research associate was available). However, patients with heart failure often don't develop symptoms until the evening when they are lying flat in bed. This study, therefore, likely enrolled patients whose disease was more severe, since they had symptoms during the day, rather than enrolling patients with a wide spectrum of heart failure.

## **Does the study population generalize to your practice?**

Most studies will present the prevalence of disease for their population. The source of patients will also produce different spectra of disease states. Patients presenting to a primary care physician, for example, are likely to have less severe disease than patients who are subsequently referred to subspecialists. The study population should be similar to the population of patients you treat.

[Jump to instructions](#)

## A Worksheet for Articles about Diagnostic Tests

*Description of the tests:*

1. Is the new test **reasonable**? What are its **limitations**? (stop)

2. Is the **reference (gold) standard** appropriate? (stop)

YES (if yes, describe)

NO

EXPLAIN:

3. Did all participants receive **both** the new test and the reference test? (stop)

YES

NO

4. Were the results of the test interpreted without knowledge (**blinded**) of the reference test result and vice versa? YES NO

*Study Population:*

1. Were the patients enrolled randomly or consecutively?

YES

NO

2. Does the study population **generalize** to your practice?

YES

NO

(Consider the spectrum of patient characteristics, co-morbidities, and clinical presentation)

EXPLAIN:

D. *Test Characteristics:*

1. What are the **sensitivity**, **specificity** and **predictive values** of the test?

a. Sensitivity=  $\frac{a}{a+c}$  \_\_\_\_\_

c. P.P.V.=  $\frac{a}{a+b}$  \_\_\_\_\_

b. Specificity=  $\frac{d}{b+d}$  \_\_\_\_\_

d. N.P.V.=  $\frac{d}{d+c}$  \_\_\_\_\_

		DISEASE	
		+	-
TEST	+	a b	
	-	c d	

2. Calculate the **prevelance** of disease in the study

$\frac{a+c}{a+b+c+d}$

3. How does this compare to your practice?



Answer to the HIV calculation

Sensitivity: 95%

Positive Predictive Value: 8.7%

Specificity: 95%

Negative Predictive Value: 99.9%

[Back to the example](#)

# Effectiveness and safety of nicotine replacement therapy assisted reduction to stop smoking: systematic review and meta-analysis

David Moore, senior reviewer Paul Aveyard, NIHR career scientist Martin Connock, systematic reviewer Dechao Wang, systematic reviewer Anne Fry-Smith, information specialist Pelham Barton, senior lecturer

School of Health and Population Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT

Correspondence to: D Moore  
d.j.moore@bham.ac.uk

Cite this as: *BMJ* 2009;338:b1024  
doi:10.1136/bmj.b1024

## ABSTRACT

**Objective** To determine the effectiveness and safety of nicotine replacement therapy assisted reduction to stop smoking.

**Design** Systematic review of randomised controlled trials.

**Data sources** Cochrane Library, Medline, Embase, CINAHL, PsychINFO, Science Citation Index, registries of ongoing trials, reference lists, the drug company that sponsored most of the trials, and clinical experts.

**Review methods** Eligible studies were published or unpublished randomised controlled trials that enrolled smokers who declared no intention to quit smoking in the short term, and compared nicotine replacement therapy (with or without motivational support) with placebo, no treatment, other pharmacological therapy, or motivational support, and reported quit rates. Two reviewers independently applied eligibility criteria. One reviewer assessed study quality and extracted data and these processes were checked by a second reviewer. The primary outcome, six months sustained abstinence from smoking beginning during treatment, was assessed by individual patient data analysis. Other outcomes were cessation and reduction at end of follow-up, and adverse events.

**Data synthesis** Seven placebo controlled randomised controlled trials were included (four used nicotine replacement therapy gum, two nicotine replacement therapy inhaler, and one free choice of therapy). They were reduction studies that reported smoking cessation as a secondary outcome. The trials enrolled a total of 2767 smokers, gave nicotine replacement therapy for 6-18 months, and lasted 12-26 months. 6.75% of smokers receiving nicotine replacement therapy attained sustained abstinence for six months, twice the rate of those receiving placebo (relative risk (fixed effects) 2.06, 95% confidence interval 1.34 to 3.15; (random effects) 1.99, 1.01 to 3.91; five trials). The number needed to treat was 29. All other cessation and reduction outcomes were significantly more likely in smokers given nicotine replacement therapy than those given placebo. There were no statistically significant differences in adverse events (death, odds ratio 1.00, 95% confidence interval 0.25 to 4.02; serious adverse events, 1.16, 0.79 to 1.50;

and discontinuation because of adverse events, 1.25, 0.64 to 2.51) except nausea, which was more common with nicotine replacement therapy (8.7% v 5.3%; odds ratio 1.69, 95% confidence interval 1.21 to 2.36).

**Conclusions** Available trials indicate that nicotine replacement therapy is an effective intervention in achieving sustained smoking abstinence for smokers who have no intention or are unable to attempt an abrupt quit. Most of the evidence, however, comes from trials with regular behavioural support and monitoring and it is unclear whether using nicotine replacement therapy without regular contact would be as effective.

## INTRODUCTION

Smoking is one of the greatest causes of illness and premature death in developed and developing countries, but giving up smoking can prevent most of the harm. Although nearly half of all smokers in the United Kingdom try to stop every year, only 2-3% succeed.<sup>1</sup> One reason for the low success is that many quit attempts are unplanned<sup>2</sup> so that the most effective cessation aids may not be used.<sup>1</sup> The most widely used cessation aid is nicotine replacement therapy.<sup>1</sup> Standard instructions for using such therapy and guidance from the National Institute for Health and Clinical Excellence require smokers to set a day when they will abruptly stop smoking and use nicotine replacement therapy or other pharmacotherapy as a substitute for smoking. Despite 70% of smokers wanting and intending to stop at some time,<sup>3</sup> only 12% are ready to stop smoking in the next month<sup>4</sup> and thus only this small proportion are suitable for abrupt quit interventions.

In the UK the licence for some nicotine replacement therapies (gum, inhaler, and, most recently, lozenge) has been extended to allow longer term use in those who are not willing or able to quit abruptly, thereby aiding them to cut down smoking and to facilitate quitting. This is termed nicotine assisted reduction to stop; also called cut down then stop,<sup>5</sup> cut down to stop, and cut down to quit. We carried out a systematic review of randomised controlled trials to determine the effectiveness of nicotine assisted reduction to stop and whether there are associated harms. Unlike previous

reviews,<sup>6,7</sup> which reported only point prevalence of cessation at end of follow-up, we focused on sustained cessation from smoking, widely considered the superior outcome measure for effectiveness.<sup>8,9</sup> This was possible because of access to unpublished trial reports. This review is an updated extension and summary of our Health Technology Assessment on this topic.<sup>10</sup>

An ancillary paper will report on an economic analysis to determine whether nicotine assisted reduction to stop provides good value for money from the perspective of the UK National Health Service.

## METHODS

We electronically searched the Cochrane library, Medline, Embase, CINAHL, PsychINFO, and Science Citation Index from at least 1992 to November 2007 for relevant trials, using a combination of free text and MeSH terms (see web extra appendix 1). We contacted authors, experts, and the pharmaceutical company that sponsored most trials, and checked reference lists of retrieved documents for further trials. All titles and abstracts were screened for relevance and we obtained the full paper if appropriate.

Studies were included in the review if they were randomised controlled trials meeting the following criteria:

- The population comprised smokers who were unable or unwilling to stop abruptly
- The intervention was gum or inhaler nicotine replacement therapy alone or as part of combination therapy, such as motivational support. Some studies considered nicotine replacement therapy as a generic intervention and allowed a choice, and such studies were considered to meet the inclusion criteria irrespective of whether data could be disaggregated for different forms of therapy (the licensing of lozenges for gradual smoking cessation coincided with the latter stages of this review and is not dealt with specifically here)
- The comparator was placebo, no treatment, non-nicotine replacement therapy drugs for smoking cessation, or psychological interventions, such as motivational support. If the intervention arm included an adjunct therapy the comparator had to include one too

- The outcome was abstinence from smoking.

The criteria were applied independently by two reviewers and discrepancies resolved by discussion and with the involvement of a third reviewer if required.

## Data extraction and quality assessment

The quality of included studies was assessed according to standard guidelines<sup>11</sup> and data extracted using a data extraction form. Both tasks were undertaken by one reviewer and checked for accuracy by a second. Disagreements were resolved by discussion, and with a third reviewer if necessary. When information was missing it was sought from the authors or sponsors of trials.

## Data synthesis

Studies were grouped according to outcome and comparison groups. The primary outcome for the review was six months' sustained abstinence starting any time before the end of treatment. We regard this as definitive evidence of the effectiveness of treatment.<sup>8,9</sup> Secondary outcomes were point prevalence abstinence at end of follow-up; sustained abstinence from early in treatment to end of follow-up; sustained reduction from week 6 to end of follow-up; point prevalence reduction at end of follow-up; and adverse events throughout follow-up—death, serious adverse events (death, admission to hospital, or permanent disability), discontinuation owing to side effects, and nausea (as an index symptom of possible nicotine overdose).

Meta-analysis was carried out using Stata (version 10). For smoking outcomes we summarised data with relative risks; the preferred statistic of the Cochrane Tobacco Addiction Review Group. For adverse events we summarised data using Peto odds ratio, which is the preferred statistic for rare occurrences.<sup>12</sup>

## Developing a measure of sustained abstinence

In most studies on smoking cessation all individuals set a quit day near the beginning of the study and once they relapse they are counted forever as a sustained abstinence failure, even if they subsequently make a quit attempt and succeed. In studies of nicotine assisted reduction to stop, participants have the opportunity to use nicotine replacement therapy for a prolonged period (up to 18 months) during which time they make several quit attempts. Unlike normal studies on cessation, where the index quit attempt is the first, in studies on nicotine assisted reduction to stop, treatment continues whether or not someone attempted to stop and failed. Thus, previous failures do not nullify later success. We counted the number who had started to abstain during treatment and had maintained abstinence for at least six months. Some smokers started quit attempts late in the treatment and because follow-up did not continue for six months beyond the end of the treatment, follow-up ceased with these people having been abstinent continuously for several months, but fewer than six months. To count them as

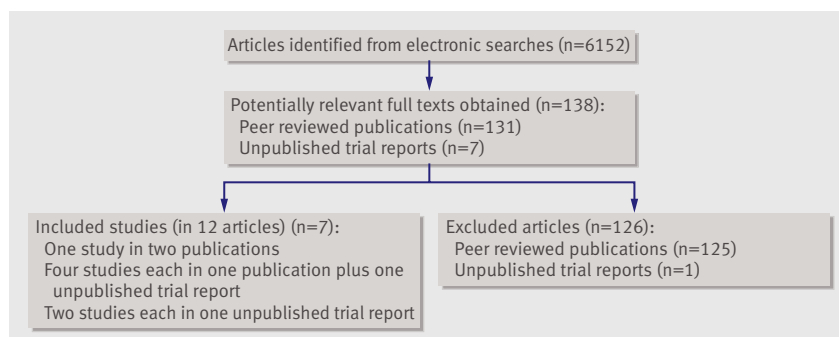


Fig 1 | Flow of papers through study

Table 1 | Main characteristics of included studies

Reference, country, trial dates	Treatment duration; follow-up (months)	Indication	No in group, mean age in years (% female)		Baseline cigarettes smoked/day, exhaled carbon monoxide level (ppm), Fagerström score*		NRT intervention† (nicotine content)	Comparator	Other treatment components	Main outcomes measured	Funding (trial code)
			NRT	Control	NRT	Control					
Batra, <sup>w1</sup> Germany and Switzerland (NR)	12; 13	Not intending to quit in next month; willing to change behaviour	184, 42.6 (45.9)	180, 43.5 (35.2)	27.9, 29.1, 5.7	29.6, 28.2, 5.9	Gum (4 mg) for 12 months	Placebo gum for 12 months	Clinic visits (n=9), telephone support, additional clinic visits as necessary	Smoking reduction; abstinence (exhaled carbon monoxide level <10 ppm)‡; NRT use (self report and records); serum cotinine and thiocyanate levels (ppm); adverse events; haematological risk factors§	Industry (980-CHC 1013-028)
Bolliger, <sup>w2</sup> Sweden and Switzerland (Feb 1997 to May 1999)	18; 24	Unwilling or unable to quit; wanted to reduce cigarette consumption	200, 46.4 (57)	200, 45.8 (48)	28.2, 27.1, 5.5	30.3, 27.1, 5.6	Inhaler (10 mg)¶ for 18 months	Placebo inhaler as required	Clinic visits (n=9)	Smoking reduction; abstinence (exhaled carbon monoxide level <10 ppm)‡; NRT use (self report), acceptability; plasma cotinine and thiocyanate levels, (ppm); quality of life** and adverse events; haematological risk factors§	Industry (96-NNIN 016)
Haustein, <sup>w3</sup> unpublished††, Germany (Mar 2000 to Nov 2001)	9; 12	Not intending to quit in next month want to reduce cigarette consumption	97, 42.3 (50)	96, 41.7 (50)	24.3, 27.5, 5.4	24.4, 28.9, 5.5	Gum (4 mg) as required for 9 months	Placebo gum as required for 9 months	Clinic visits (n=8)	Smoking reduction; abstinence (exhaled carbon monoxide level <10 ppm); product use; change in Fagerström score; adverse events	Industry (980 CHC-9021-0013)
Rennard, <sup>w4</sup> USA (Feb 2000 to Apr 2001)	12; 15	Not intending to quit within next month, wanted to reduce cigarette consumption	215, 45.9 (59)	214, 44.8 (54)	29.3, 29.7, 6.5	30.4, 29.5, 6.6	Inhaler (10 mg) for 12 months	Placebo inhaler for 12 months	Clinic visits (n=9)	Smoking reduction; abstinence (exhaled carbon monoxide level <10 ppm)‡; NRT use (self report), acceptability; plasma cotinine and thiocyanate levels (ppm); quality of life††† and adverse events; haematological risk factors§	Industry (98-NNIN-027)
Wennike, <sup>w5</sup> Denmark (Feb 1999 to May 2000)	12; 24	Not intending to quit within next month, wanted to reduce cigarette consumption	205, 45 (65)	206, 44 (59)	24, 29, 6.4	24, 27, 6.4	Gum (2 or 4 mg; depending on Fagerström score) for 12 months	Placebo gum for 12 months	Clinic visits (n=9)	Smoking reduction; abstinence (exhaled carbon monoxide level <10 ppm)‡; NRT use (self report) and compliance; plasma cotinine and thiocyanate levels; quality of life** and adverse events; haematological risk factors§	Industry (98 NNCG-014)
Wood-Baker, <sup>w6</sup> unpublished, Australia (Jun 1999 to Mar 2001)	12; 15	Not intending to quit within next month, wanted to reduce cigarette consumption	218, 42.9 (54)	218, 45.3 (55)	29.0, 25.8, 6.6	27.4, 25.9, 6.4	Gum (2 or 4 mg; depending on Fagerström score) for 12 months	Placebo gum for 12 months	Clinic visits (n=9)	Smoking reduction; abstinence (exhaled carbon monoxide level <10 ppm)‡; NRT use and compliance; plasma cotinine and thiocyanate levels; quality of life** and adverse events; haematological risk factors§	Industry (98 NNCG-017)
Etter, <sup>w7</sup> Switzerland (1999 to 2002)§§	6¶¶; 26	Not intending to quit within next 6 months, wanted to reduce cigarette consumption	265, 269, 389; 43.2, 41.7, 42.9; (46, 51, 56)		29.8, 29.4, 30.2; NR, NR, NR; 6.0, 5.9, 6.2		Free choice***: inhaler (10 mg), gum (4 mg), or patch (25 mg) for 6 months	Placebo NRT for 6 months and no intervention	Literature only	Smoking reduction; abstinence†††; product use; change in Fagerström score; adverse events	Government and industry (no trial code)

NR=not reported; NRT=nicotine replacement therapy.

\*Test for nicotine dependence.

†Gum and inhaler were Nicorette products (Pharmacia).

‡Seven day point prevalence.

§Examples include C reactive protein, fibrinogen, white blood cell count.

¶Total available nicotine 4-5 mg.

\*\*Short form 36.

††This study had two further arms that compared short term quit intervention using gum with placebo.

†††Revised RAND 36 item health survey 1.0.

§§This study had a third arm in which participants received no treatment.

¶¶Quitters continued to receive NRT after six months.

\*\*\*Switching between products was allowed.

†††Point prevalence for past seven days and one month.

failed quitters or as successes would be inappropriate, so we developed a method to determine what proportion of these would sustain abstinence of six months if follow-up had been long enough. We applied the probability that a smoker who abstained for *x* months would go on to abstain for six months to those smokers who were abstinent for *x* months at the end of study. This calculation was based on probabilities derived from analyses using individual person data of all quit attempts made in each of the studies for which individual person data were available (see web extra appendix 2).

## RESULTS

Figure 1 shows the flow of papers through the systematic review. Seven randomised controlled trials<sup>w1-w7</sup> (12 articles) met the inclusion criteria (see web extra appendix 3 for excluded articles).

Table 1 shows the characteristics of the included studies. Six of the randomised controlled trials were sponsored by industry,<sup>w1-w6</sup> two of which were unpublished.<sup>w3 w6</sup> Full, unpublished trial reports were obtained for all six trials, of which five reports<sup>w1 w3-w6</sup> contained individual patient data allowing calculation of at least six months' sustained abstinence. The seventh trial<sup>w7</sup> was independent, and unpublished data were obtained from the authors.

All the studies recruited smokers who were unwilling or unable to quit abruptly, and none emphasised reduction then stop on recruitment. Consequently the primary outcome was reduction and not cessation.

### Trial design

All the studies were randomised parallel group trials with nicotine replacement therapy and placebo arms. One trial<sup>w3</sup> randomised people to four arms; two of the arms were not included in this review because participants were randomised to reduction over only one month (with active nicotine replacement therapy or

placebo). Another trial<sup>w7</sup> had three arms, comprising no pharmacotherapy, placebo, and nicotine replacement therapy. For consistency we analysed differences between nicotine replacement therapy and placebo.

### Population

The populations had similar personal and smoking characteristics typical of heavy smokers attending smoking cessation clinics. Potential participants with heart disease, those receiving psychiatric drugs, pregnant or lactating women, or people with other drug problems were excluded. Recruitment was by advertisement.

### Intervention

Four trials used gum,<sup>w1 w3 w5 w6</sup> two used inhalers,<sup>w2 w4</sup> and one used free choice of gum, inhaler, or patch.<sup>w7</sup> Prior to randomisation in two trials,<sup>w5 w6</sup> smokers were stratified by nicotine dependence (Fagerström score); the less nicotine dependent were given 2 mg gum whereas the more dependent received 4 mg gum. The other gum trials used 4 mg gum. The trial with three arms<sup>w7</sup> used a 15 mg/16 hour patch, 4 mg gum, or inhaler.

Nicotine replacement therapy was available for six months in one trial<sup>w7</sup> (although people who remained abstinent could have extended use). The other trials provided nicotine replacement therapy for nine months,<sup>w3</sup> 12 months,<sup>w1 w4-w6</sup> and 18 months.<sup>w2</sup>

### Behavioural support

The trial with three arms<sup>w7</sup> had no clinic visits and no behavioural support, but participants received a 20 page booklet covering reasons for reducing cigarette consumption and the methods for achieving reduction.

In the other publications behavioural support was described as moderate (visits lasting 15-30 minutes<sup>w5</sup>), or participants were "instructed to reduce their smoking . . . and provided with ways to do so,"<sup>w4</sup> or

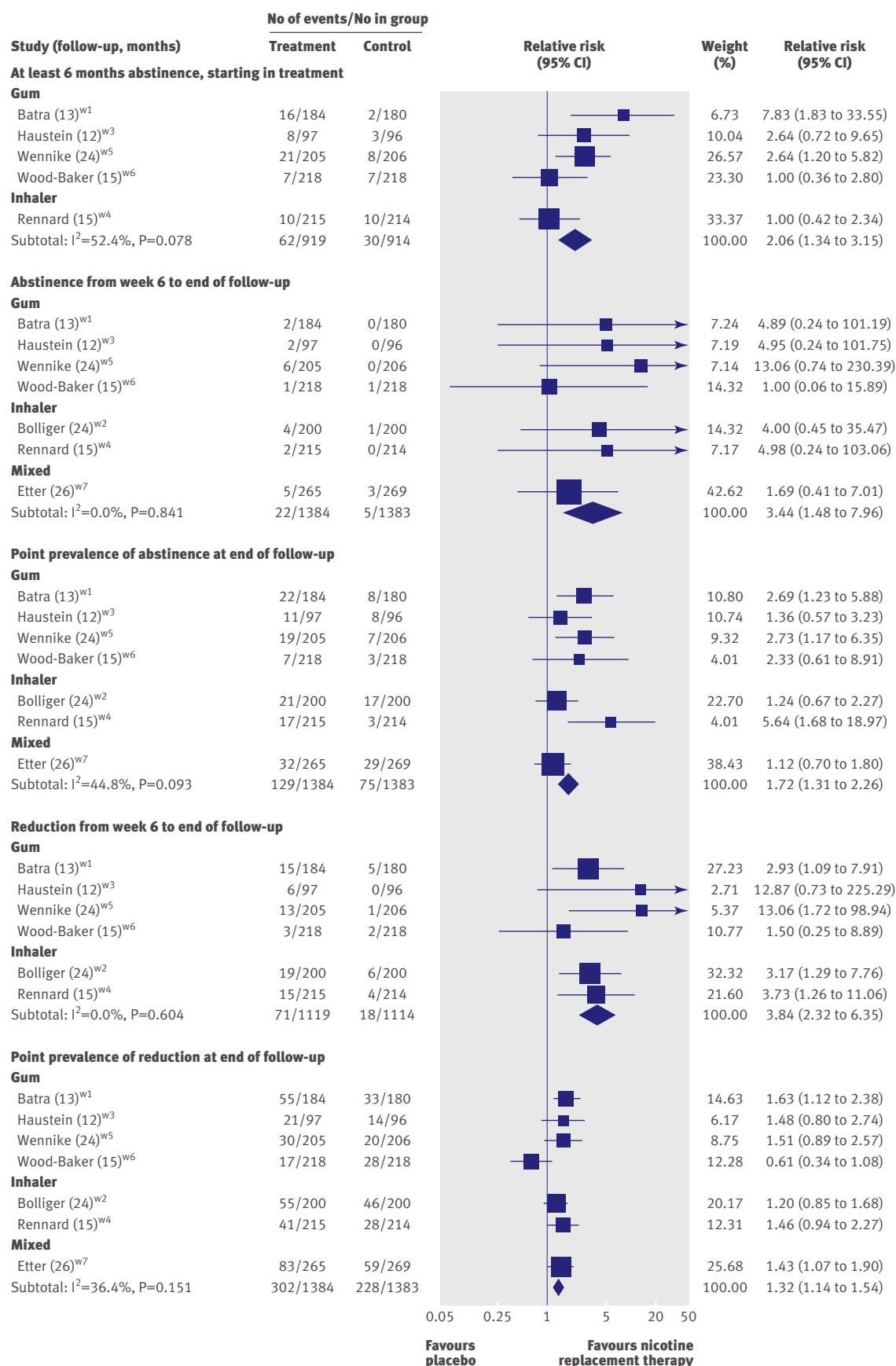
**Table 2** | Summary of quality assessment of included randomised controlled trials

Study	Was assignment of treatment really random?	Was allocation concealed and concealment method described?	Were groups similar at baseline?	Were eligibility criteria specified?	Who was blinded to treatment allocation?	Was intention to treat analysis used and were drop outs accounted for?
Batra <sup>w1</sup>	Yes; computer generated list	Yes; sealed envelopes	Yes*	Yes	Participants, therapists, and outcome assessors	Yes, yes
Bolliger <sup>w2</sup>	Yes; computer generated list	Yes; sealed envelopes	Yes*	Yes	Participants, therapists, and outcome assessors	Yes, yes
Haustein <sup>w3</sup>	Yes; computer generated list	Yes; sealed envelopes	Yes	Yes	Participants, therapists, and outcome assessors	Yes, yes
Rennard <sup>w4</sup>	Likely, but method not described	Likely, but method not reported	Yes	Yes	Participants, therapists, and outcome assessors	Yes, yes
Wennike <sup>w5</sup>	Yes (stratified by Fagerström score); computer generated list	Yes; sealed code list	Yes*	Yes	Participants, therapists, and outcome assessors	Yes, yes
Wood-Baker <sup>w6</sup>	Yes (stratified by Fagerström score); computer generated list	Yes; sealed envelopes	Yes	Yes	Participants, therapists, and outcome assessors	Yes, yes
Etter <sup>w7</sup>	Yes; computer generated list	Unclear	Yes	Yes	Participants and outcome assessors	Most outcomes†, yes

When extensive unpublished study reports were available, they were used for quality analysis.

\*Except for small imbalance in sex distribution.

†Not intention to treat for product usage and for completeness of blinding of participants (determined at six months).



**Fig 2** | Meta-analysis of smoking outcomes. Pooled estimates are Mantel Haenszel relative risks (fixed effects). Heterogeneity statistic Q for at least six months' abstinence was 8.4 (P=0.078), for abstinence from week 6 to end of follow-up was 2.74 (P=0.840), for point prevalence of abstinence at end of follow-up was 10.86 (P=0.093), for reduction from week 6 to end of follow-up was 3.63 (P=0.604), and for point prevalence of reduction at end of follow-up was 9.43 (P=0.151)



behavioural support was not really described.<sup>w1 w2</sup> The unpublished trial reports,<sup>w1-w6</sup> however, indicated that the behavioural support programme was similar in all these studies. Participants were given a sheet of paper with written advice on how to use gum or inhaler to reduce or stop smoking. Clinic staff followed a written behavioural support protocol giving information on how much nicotine replacement therapy to use and how to use it to substitute for cigarettes. In addition, at each visit the therapist elicited problems from the participants, helped them find solutions, and related their progress back to their goals negotiated at the start of the programme. Smokers were encouraged to quit during the study. At six and nine months, participants were instructed to stop smoking completely, regardless of reduction achieved to that point. At all visits smoking status was monitored, exhaled carbon monoxide recorded, and feedback given on progress towards agreed goals. Typically, behavioural support and clinic visits were repeated on five or more occasions up to at least a year and in some trials beyond, to 18 or 24 months.

### Outcomes

The primary outcome in the trials was sustained reduction. In the industry sponsored trials<sup>w1-w6</sup> sustained reduction was defined as reported cigarette consumption of less than 50% of baseline from week 6 to week 16, although in some trials this was also to later visits. Sustained reduction was measured by self reported cigarettes smoked a day and validated by the carbon monoxide level that was at least 1 ppm less than at baseline on each occasion it was checked. The secondary outcomes were prolonged abstinence from the week 6 visit to end of follow-up and 7 day point prevalence abstinence and point prevalence of reduction at various follow-up times.

In the trial with three arms,<sup>w7</sup> point prevalence abstinence and point prevalence reduction for the past seven days and four weeks were the main outcomes at six and 26 months.

### Quality of included studies

Table 2 summarises the quality of the included studies. All were of high quality.

Although trials blinded participants to allocation, it is difficult to blind people to psychoactive drugs. At six months, participants in the three arm trial<sup>w7</sup> guessed more accurately than would be expected by chance whether they had received active drug or placebo.

### Sustained six months' abstinence

Individual person data were available from one trial using inhaler<sup>w4</sup> and four using gum<sup>w1 w3 w5 w6</sup> and allowed the calculation and meta-analysis of sustained abstinence of at least six months.<sup>w1 w3-w6</sup> The proportion of smokers achieving sustained abstinence at six months with nicotine replacement therapy was double that with placebo (relative risk 2.06, 95% confidence interval 1.34 to 3.15; fig 2), but the rates were low (6.75% v 3.28%, respectively). Moderate heterogeneity

was suggested ( $\chi^2=8.4$ ,  $df=4$ ,  $P=0.08$ ,  $I^2=53\%$ ). There was no evidence to indicate that this was due to the type of nicotine replacement therapy used, and the inclusion criteria and protocols of the trials were similar. By a random effects model the relative risk was 1.99 (1.01 to 3.91).

### Other smoking outcomes

Sustained abstinence was measured from six weeks (two weeks in one study<sup>w1</sup>) to the end of follow-up. Point prevalence abstinence was also measured at last follow-up, which was one month,<sup>w1</sup> three months,<sup>w3 w4 w7</sup> six months,<sup>w2</sup> 12 months,<sup>w5</sup> and 20 months<sup>w7</sup> after the end of treatment. Sustained reduction and point prevalence reduction was measured at these time points during treatment and at follow-up. Figure 2 summarises these results.

As might be expected of smokers unwilling or unable to quit in the short term, sustained abstinence rates starting from six weeks were low; across all studies 1.6% in the nicotine replacement therapy group and 0.4% in the placebo group. Point prevalence rates of abstinence at the end of follow-up were 9.3% and 5.4%, respectively.

Successful reduction was more common. In those receiving active nicotine replacement therapy, 21.8% had reduced consumption by more than 50% at final follow-up compared with 16.5% receiving placebo. Sustained reduction from early in treatment to final follow-up occurred in 6.3% of those receiving active treatment and 1.6% receiving placebo.

### Adverse events

Overall, 1384 predominantly middle aged smokers were treated with nicotine replacement therapy for six to 18 months and 1383 were treated with placebo. Four deaths occurred in those randomised to nicotine replacement therapy and four in those randomised to placebo: odds ratio 1.00 (95% confidence interval 0.25 to 4.02; fig 3). Serious adverse events occurred in fewer than 8% of participants in both arms: 1.09 (0.79 to 1.50; fig 3). In no cases were these judged likely to have been due to treatment. Discontinuation of treatment because of adverse events was rare, with 1.7% in the nicotine replacement therapy group and 1.3% in the placebo group: odds ratio 1.27 (0.64 to 2.51; fig 3). Nausea was selected as an index symptom to indicate possible nicotine overdose. It was slightly and significantly more common in the nicotine replacement therapy group, with 8.6% experiencing nausea compared with 5.3% in the placebo group: 1.69 (1.21 to 2.36; fig 3).

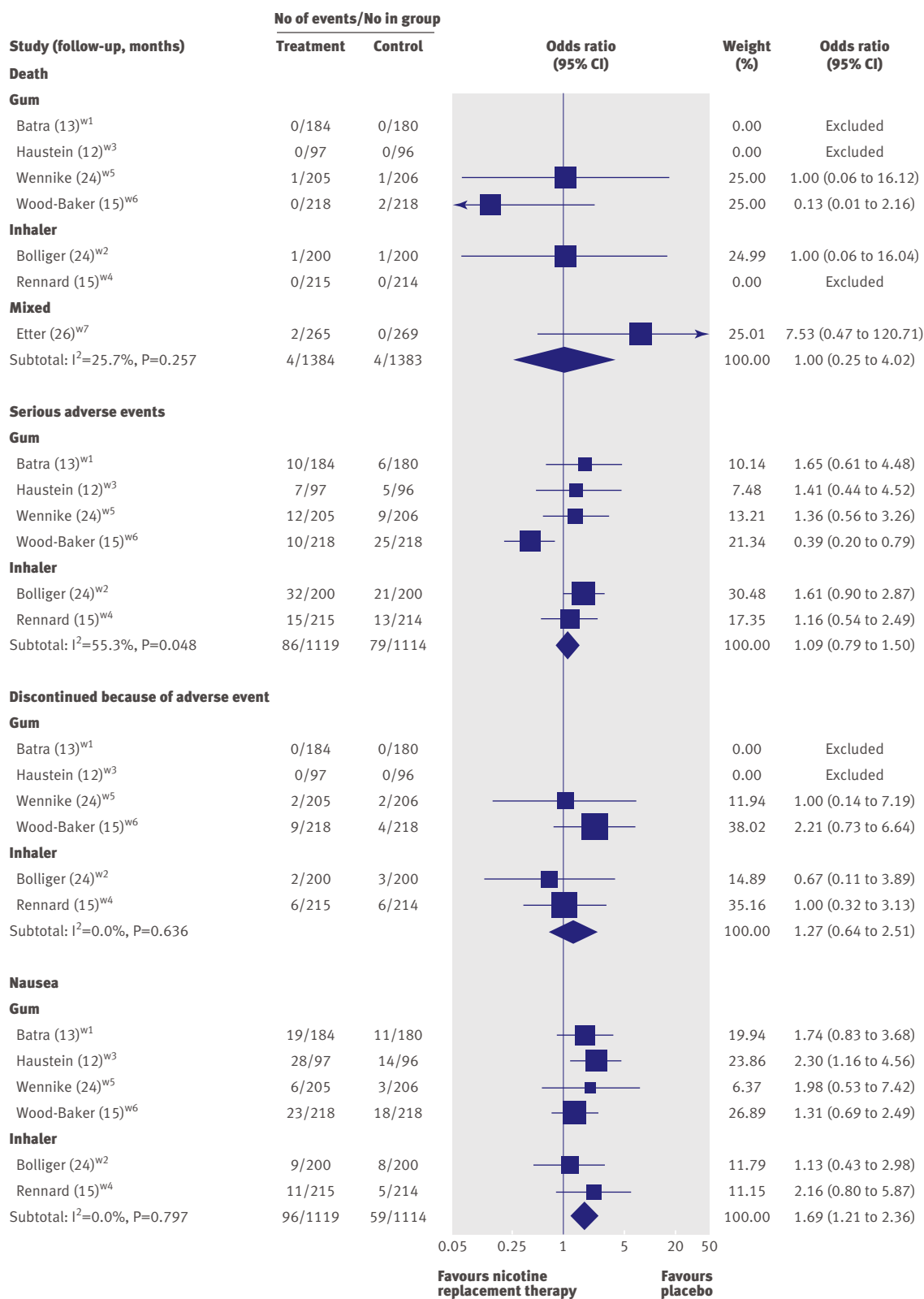
### DISCUSSION

This review found evidence that nicotine assisted reduction to stop programmes can be effective in achieving sustained abstinence from smoking of six months. There was no evidence of an increase in life threatening problems, and nicotine replacement therapy was well tolerated, with almost no difference in discontinuation because of side effects in those



receiving nicotine replacement therapy compared with those receiving placebo. Nausea was significantly

more common with nicotine replacement therapy than with placebo, but only one in 30 users became



**Fig 3** | Meta-analysis of safety outcomes; pooled estimates are Peto's odds ratio (fixed effects). Heterogeneity statistic Q for death was 4.04 (P=0.257), for serious adverse events was 11.19 (P=0.048), for discontinuation of treatment because of adverse events was 1.70 (P=0.636), and for nausea was 2.36 (P=0.797). I<sup>2</sup> was 0 (negative value [100×[(Q-DF)/Q]] except for serious adverse events, where I<sup>2</sup> was 55%

nauseous because of treatment. The results imply that compared with placebo twice the number of smokers sustained six months' abstinence as a result of nicotine replacement therapy. This equates to about an additional 3% of all smokers quitting who would otherwise not have done so. This is a similar effect size to treating smokers who are motivated to quit, where 4-5% might be expected to abstain for six months owing to use of nicotine replacement therapy.<sup>13</sup> Previous data suggest that half of those who sustain six months of abstinence will maintain it for the rest of their lives.<sup>14 15</sup>

Three reviews, comprising a Health Technology Assessment,<sup>10</sup> a Cochrane review,<sup>7</sup> and a qualitative review<sup>6</sup> have examined smoking cessation achieved by smokers recruited to randomised controlled trials of smoking reduction interventions. The present review is an extension and update of the Health Technology Assessment report<sup>10</sup> and differs from the Cochrane review<sup>7</sup> and the qualitative review.<sup>6</sup> We report sustained abstinence rates derived from analysis of individual patient data, whereas the Cochrane and qualitative reviews were restricted to point prevalence of smoking cessation at the end of follow-up, a measure that cannot inform about the duration of cessation, which is the outcome most relevant to health. Sustained abstinence is the preferred outcome of the Society for Research in Nicotine and Tobacco<sup>8</sup> and advised by other experts.<sup>9</sup> We included an additional trial<sup>w6</sup> not included in the qualitative review. Our review focused on nicotine replacement therapy whereas both the Cochrane and the qualitative reviews encompassed multiple interventions and did not meta-analyse data on safety outcomes. The qualitative review concluded that smoking reduction increased the probability of future cessation, whereas the Cochrane review concluded that people unwilling to quit were helped by nicotine replacement therapy to cut down on number of cigarettes smoked a day. Our use of sustained abstinence and measurement of safety has allowed us to draw stronger conclusions on the public health benefit of smoking reduction with nicotine replacement therapy for unwilling quitters.

The licence for nicotine replacement therapy is for reduction then stopping, whereas the trials in our review recruited smokers motivated only to reduce

their consumption. We excluded one study in which participants wanted to quit by reduction,<sup>16</sup> which was included in both the Cochrane and the qualitative reviews. The odds ratio for point prevalence of abstinence at the end of follow-up from this study was similar to our pooled effect estimate (2.34, 95% confidence interval 1.16 to 4.74); this suggests that whether smokers are motivated to reduce then quit or simply motivated to reduce may make little difference to the efficacy of nicotine replacement therapy in supporting cessation. There is further evidence that this difference between trial populations (reducers) and the smokers for whom the products are marketed (reducing to quit) is probably not important. Nearly half of surveyed American smokers planning to quit would choose reduction over abrupt cessation, and two thirds of these were interested in the assistance of drugs.<sup>17</sup> In these smokers there was little interest in reduction as an end in itself, only as a means to stop. Even among those not planning to stop soon, cessation was the goal for half. Intentions to stop smoking are volatile<sup>18</sup> so a stated intention to stop at a specified future time may have little long term meaning for many smokers. Instruction to stop was delivered in the trials, although the importance of this instruction has not been tested. We therefore believe that encouraging smokers prepared to reduce consumption to use nicotine replacement therapy regardless of their subsequent intention to quit is appropriate because this is the population that was included in the trials.

For health services an important issue is whether nicotine replacement therapy should be reimbursed in nicotine assisted reduction to stop programmes and whether and how such programmes should be implemented. All the industry sponsored trials took place in specialist smoking cessation clinics with extensive monitoring and moderate behavioural support. The remaining trial<sup>w7</sup> was rather analogous to use of nicotine replacement therapy purchased directly from retail outlets, but even here a 20 page booklet was given to participants to motivate and instruct on reduction. This trial showed lower relative efficacy than the overall effect estimate and a lower absolute benefit, but whether this was due to the setting or chance is unclear.

Currently, nicotine assisted reduction to stop is licensed in the UK but recent guidance from the National Institute for Health and Clinical Excellence and a recent US Clinical Practice Guideline recommend its use only in the context of further research.<sup>19 20</sup> Survey data show that large numbers are using nicotine replacement therapy to reduce consumption,<sup>1</sup> but whether they are truly reading the packet inserts and seeking to follow a nicotine assisted reduction to stop programme is uncertain. Furthermore, most people who are reducing with nicotine replacement therapy are using a patch,<sup>1</sup> which is not licensed for this use and does not come with such instructions. It is therefore unclear whether the outcomes observed in the trials are being achieved through such use.<sup>21</sup>

#### WHAT IS ALREADY KNOWN ON THIS TOPIC

Most smokers are not ready to quit and might not respond to interventions of abrupt cessation

Nicotine replacement therapy (NRT) is licensed for smoking reduction in smokers not ready to stop but there is no evidence that it leads to sustained abstinence

No review has assessed the safety of concurrent smoking and use of long term NRT

#### WHAT THIS STUDY ADDS

This systematic review of randomised clinical trials in smokers not ready to stop found that with NRT support twice as many quitters achieve six months of sustained abstinence

This equates to an additional 3% of sustained quitters compared with placebo

Using NRT while smoking did not lead to serious health problems

In summary, these trials have shown proof of concept. People who would answer “no” to “do you want to stop smoking now?” may be helped to stop over a longer period by applying drugs formerly reserved only for abrupt cessation. The contribution of the behavioural support programme is unknown, and the optimum advice to give people in reduction programmes is also unknown as these have not been manipulated in comparative trials. The importance of these trials is that they show that treating a population of smokers not ready to stop means more of them stop. Therefore it is important to examine how nicotine assisted reduction to stop can be incorporated into tobacco control programmes.

**Contributors:** AFS designed and implemented the searches. DW, MC, and PA extracted the data. DW and MC selected studies and did the meta-analyses. DM supervised the project and is the guarantor. All authors wrote the manuscript.

**Funding:** This work was funded by the UK Health Technology Assessment Programme (National Institute for Health Research).

**Competing interests:** PA has accepted hospitality and money from McNeil (Helsinborg, Sweden), which sponsored the trials in the report; he has not received hospitality or money in relation to any nicotine assisted reduction research.

**Ethical approval:** Not required.

- 1 West R. Smoking and smoking cessation in England, 2006. Reference paper 4. 2008. <http://aspsilverbackwebsites.co.uk/smokinginengland/>.
- 2 West R, Sohal T. “Catastrophic” pathways to smoking cessation: findings from national survey. *BMJ* 2006;332:458-60.
- 3 Office for National Statistics. Smoking-related behaviour and attitudes. 2008. [www.statistics.gov.uk/downloads/theme\\_health/Smoking2005.pdf](http://www.statistics.gov.uk/downloads/theme_health/Smoking2005.pdf).
- 4 Taylor T, Lader D, Bryant A, Keyse L, Joloza MT. *Smoking-related behaviour and attitudes, 2005*. London: Office for National Statistics, 2006.
- 5 Medicines and Healthcare Products Regulatory Agency, Committee on Safety of Medicines: Report of the committee on safety of medicines working group on nicotine replacement therapy. 2008. [www.mhra.gov.uk/home/idcplg?IdcService=GET\\_FILE&dDocName=CON2023239&RevisionSelectionMethod=LatestReleased](http://www.mhra.gov.uk/home/idcplg?IdcService=GET_FILE&dDocName=CON2023239&RevisionSelectionMethod=LatestReleased).
- 6 Hughes JR, Carpenter MJ. Does smoking reduction increase future cessation and decrease disease risk? A qualitative review. *Nicotine Tob Res* 2006;8:739-49.
- 7 Stead LF, Lancaster T. Interventions to reduce harm from continued tobacco use. *Cochrane Database Syst Rev* 2007;(3):CD005231.
- 8 Hughes JR, Keely JP, Niaura RS, Ossip-Klein DJ, Richmond RL, Swan GE. Measures of abstinence in clinical trials: issues and recommendations. *Nicotine Tob Res* 2003;5:13-25.
- 9 West R, Hajek P, Stead L, Stapleton J. Outcome criteria in smoking cessation trials: proposal for a common standard. *Addiction* 2005;100:299-303.
- 10 Wang D, Connock M, Barton P, Fry-Smith A, Aveyard P, Moore D. “Cut down to quit” with nicotine replacement therapies in smoking cessation: a systematic review of effectiveness and economic analysis. *Health Technol Assess* 2008;12:1-135.
- 11 NHS Centre for Reviews and Dissemination. *Undertaking systematic reviews of research on effectiveness: CRD’s guidance for those carrying out or commissioning reviews*. 2nd ed. University of York, Centre for Reviews and Dissemination, 2001.
- 12 Deeks JJ, Altman DG. Effect measures for meta-analysis of trials with binary outcomes. In: Egger M, Davey Smith G, Altman DG, ed. *Systematic reviews in health care; meta-analysis in context*. 2nd ed. London: BMJ Books, 2001:313-35.
- 13 Silagy C, Lancaster T, Stead L, Mant D, Fowler G. Nicotine replacement therapy for smoking cessation. [update in *Cochrane Database Syst Rev* 2004;(3):CD000146; PMID: 15266423]. *Cochrane Database Syst Rev* 2002;(4).
- 14 Stapleton J. Cigarette smoking prevalence, cessation and relapse. *Stat Methods Med Res* 1998;7:187-203.
- 15 Etter JF, Stapleton JA. Nicotine replacement therapy for long-term smoking cessation: a meta-analysis. *Tob Control* 2006;15:280-5.
- 16 Kralikova E, Kozak J, Rasmussen T, Cort N. The clinical benefits of NRT-supported smoking reduction. *Nicotine Tob Res* 2002;4:243.
- 17 Shiffman S, Hughes JR, Ferguson SG, Pillitteri JL, Gitchell JG, Burton SL. Smokers’ interest in using nicotine replacement to aid smoking reduction. *Nicotine Tob Res* 2007;9:1177-82.
- 18 Hughes JR, Keely JP, Fagerstrom KO, Callas PW. Intentions to quit smoking change over short periods of time. *Addict Behav* 2005;30:653-62.
- 19 National Institute for Health and Clinical Excellence. NICE public health guidance 10. Smoking cessation services in primary care, pharmacies, local authorities and workplaces, particularly for manual working groups, pregnant women and hard to reach communities. 2008. [www.nice.org.uk/nicemedia/pdf/PH010guidance.pdf](http://www.nice.org.uk/nicemedia/pdf/PH010guidance.pdf).
- 20 Fiore MC, Jaen CR, Baker TB, Bailey WC, Benowitz NL, Curry SJ, et al. *Treating tobacco use and dependence: 2008 update*. Clinical Practice Guideline. Rockville, MD: US Department of Health and Human Services Public Health Service, 2008.
- 21 Levy DE, Thorndike AN, Biener L, Rigotti NA. Use of nicotine replacement therapy to reduce or delay smoking but not to quit: prevalence and association with subsequent cessation efforts. *Tob Control* 2007;16:384-9.

**Accepted:** 14 January 2009

**Study Guide**  
**Session 4**  
**Reading and Determining the Validity of Review Articles**  
**Jessica Early, MD**  
**[JEarly@challiance.org](mailto:JEarly@challiance.org)**

**The aims of this session are to:**

1. Help you identify review articles that are more likely to be valid; and,
2. Explain how to read the results of meta-analyses

**Specific Objectives:** By completing the initial reading and participating in class, students should be able to:

1. Differentiate synthesis and summary review articles
2. List the components of a systematic review
3. Use a worksheet to evaluate a systematic review for:
  - a. The quality of the search and selection of evidence
  - b. The quality of the included evidence
  - c. Homogeneity of the results
  - d. Evidence of publication bias
4. Interpret a forest plot of a meta-analysis

This study guide provides an outline of the concepts necessary to meet these objectives. It contains hyperlinks to short videos, web pages, or articles that explain the concepts in other ways or in greater detail. You can follow these hyperlinks if the explanations and examples I've given you are not sufficient to help you understand and to help you complete the readiness assessment test.

## What are review articles?

Review articles summarize or analyze research previously published by others, rather than reporting new experimental results (although, as we will see, they also can report new data). They are often called “secondary literature” since they build on research literature, which is called “primary literature.”

There are two main types of review articles. **Summary reviews** are the traditional type of review. They cover the full breadth of a particular topic, typically providing an overview of the disease etiology, diagnosis, prognosis, or management, and will usually address a number of questions. Experts usually write them. Book chapters are summary reviews. This type of review is useful and often fine for [background](#) questions.



Read [Example of s summary review](#)

**Synthesis reviews** are systematic reviews, either with or without meta-analysis. Defining one or two specific questions, writers of systematic reviews carefully find all available evidence, evaluate its validity, and report their answer to the question. This type of review is useful when answering [foreground](#) questions.



Read [Synthesis review example](#)

**Meta-analysis is a statistical technique** for combining the findings from independent studies. It can be performed following a systematic review, to treat the data from different studies as if they were from one large study, rather than simply counting the studies (“4 studies say it works, 2 studies say it doesn’t, so I guess it works.”)



Read [Meta-analysis example](#)



Watch [Meta-analysis](#) (4:46)

## The Cochrane Collaboration



The [Cochrane Collaboration](#) is an international network of more than 28,000 people from [over 100 countries](#) (The New England Cochrane Center is housed at Tufts in the [Center for Clinical Evidence Synthesis](#). The group has produced over 5,000 systematic reviews using a process that is considered to be the gold standard for systematic review and meta-analysis.

The logo for the Cochrane Collaboration is a forest plot (explained below) of the results of using corticosteroid treatment of pregnant women at risk of premature delivery. The use of this simple and inexpensive treatment decreases mortality in the newborns by 30% - 50%. It was not until publication of this meta-analysis in 1991 that maternity care physicians started using this treatment regularly, saving thousands of lives (more about the [history of the Cochrane logo](#))

## What are the issues with review articles?

Typically, authors of summary reviews are experts in the area of the review. As such, the review writer usually makes little or no attempt to be systematic in the formulation of the question they are addressing, searching for evidence, or summarizing the evidence they consider. As a result, the information in summary reviews has to be taken at face value.\*

There are other issues with summary reviews:

- 1) *Misreferenced statements.* The citation at the end of the sentence doesn't support the statement. In some studies this has been as high as 40% of all citations.

- 2) *Information imposters:* The article seems to convey information but uses wish-washy phrase such as "may be effective" or "should be useful, leaving the reader unsure

- 3) *Missing information* due to a lack of a literature search.

- 4) *Lagging recommendations.* Recommendations in review articles may not be based on the best current evidence. In an analysis of review articles

of [treatment of acute myocardial infarction](#), an average of 13 years passed between the time good evidence was available to support a treatment and the recommendation of that treatment for routine use in review articles. In review articles on the [treatment of type 2 diabetes](#), most reviews did not accurately convey the results of a landmark study.

- 5) *Reliance on the expert's knowledge* rather than a systematic approach to evidence. The methodologic rigor of the review, in one study,<sup>†</sup> was inversely related to the self-rated clinical expertise of the review writer.

are available, but their long-term efficacy is unknown. A major study of the natural history of macular degeneration is under way.

### Treatment for more than the lucky few?

It used to be thought that only about 25% of patients affected with neovascular macular degeneration were candidates for laser treatment because the subretinal neovascularization was in an untreatable area of the fovea. It now seems possible to treat new vessels in the fovea with less damage to central vision than was previously feared. Eyesight cannot be improved by this treatment, but visual decline can be arrested. In addition, this treatment may prevent a more devastating loss of vision caused by subretinal neovascularization and hemorrhage. For example, if the patient's vision is reduced from 20/100 to 20/400 by treatment of the subfoveal mem-

### Stopping it before it starts

The DCCT results proved what many diabetologists and ophthalmologists have assumed for some time: Tight control of blood glucose levels has the potential to decrease the incidence of nonproliferative diabetic retinopathy and progression to proliferative diabetic retinopathy by 50-75%.

The drawbacks to this approach include the intense effort and large time expenditure required by patients and health care providers. Hypoglycemia is more likely when patients adopt this approach to diabetes self-management, which usually requires multiple daily insulin injections and maintenance of glycosylated hemoglobin levels within the normal range. Nonetheless, an increasing number of patients will probably embark on a program of tight glycemic con-

## Wishy-washy terms in a single article

\* I call these reviews, "trust me, I'm the expert" reviews.

<sup>†</sup> Oxman AD, Guyatt GH. The science of reviewing research. Authority, superstition, and science. *Ann NY Acad Sciences* 1993;703:125-33.



## Evaluating Review Articles for Relevance and Validity

The goal of using this worksheet ([jump to worksheet](#)) is to quickly determine whether the review article is likely to present relevant and valid information. It allows determination, based on answers to the questions, whether the information is relevant to you, and whether the study design has sufficient rigor to apply the results to your patients.

The questions focus on the study design issues described above. The first 6 questions address study “musts” – they ask about issues of relevance or validity that must be present if the study results are to be applied to clinical practice. The answers to these questions must be yes regardless of the answers to the rest of the questions.

The goal of this worksheet is **not** to determine whether a study is “good” or “bad.” Instead, we will use it to determine whether the results reported by the study are likely to occur if we use the same approach in our patients. As a result, the information is either useful to us, or not.

### Step 1. How was relevant research identified?

#### *Were the methods used to locate relevant studies comprehensive and clearly stated?*

This question quickly separates summary reviews from synthesis reviews. The latter type of review will start by explaining how the authors assembled evidence for review.

*“A Medline search was performed,” is not an adequate explanation of a literature search. A Medline search will miss [30% - 50%](#) of applicable controlled studies.*

Instead, the methods should include:

- 1) A [detailed explanation](#) of the method used to search Medline, including search terms and strategies.
- 2) Searching of at least two databases. If the review involves a treatment, the [Cochrane Central Register of Controlled Trials](#) must be one of the searched databases. Other databases include
  - a. The [Health Technology Assessment Database](#)
  - b. [EMBASE](#)
  - c. [LILACS](#) (Latin American and Caribbean Health Sciences Literature)
  - d. [DARE](#), The Database of Abstracts of Reviews of Effectiveness
  - e. [Web of Science](#)
  - f. [Scopus](#)
  - g. [HerbMed](#) for botanicals
  - h. [Clinical trial registries](#). Most journals require that a study protocol be registered before the study is started, and a registry can identify studies that may have been completed but not published.
- 3) Unpublished and [gray](#) literature.

- a. Results of studies may be published in meeting abstracts and, especially if a negative study, may not be published in a journal. The authors should look at the appropriate meeting abstracts for relevant research.
  - b. Not all research is published in journals but may be in government or other database.
- 4) Reference lists of identified articles. Since researchers will reference previous research in their own research descriptions, bibliographies are useful to find additional research.

***Did they clearly outline study inclusion criteria that generalize to my practice?***

The researchers should explain how they decided on which articles to include and exclude from their analysis. The included studies should include patients similar to patients in your care. For example, a meta-analysis of the effect of blood glucose control in intensive care surgical patients may not apply to patients on a medical ward.

***Was the study selection independently performed by at least two investigators?***

The literature search and article selection project should be performed by two researchers and their results compared.

***Step 2. How valid was the identified research?***

Garbage in = garbage out. There are several steps researchers must take to evaluate the research they have found:

***a. Did the authors perform a validity assessment of the studies using appropriate criteria?***

The assessment should be similar to those discussed for evaluating research regarding a therapy or diagnostic test.

***b. Was the assessment independently performed by at least two investigators?***

As mentioned above, the results should be compared and differences resolved, usually through discussion or by adjudication by a third researcher.

***c. Were the included studies reasonably valid?***

If not, how did the authors handle it? One option, which should be decided before the analysis is undertaken, is to include only studies of a certain quality level. A second approach is to separately analyze studies of high quality and low quality.

### Step 3: Analyzing the data. Is it reasonable to combine these studies?

#### *a. Were the included studies statistically homogenous?*

Studies, conducted at different times, on different populations, with slight differences in design, will not produce the same results. The variability among studies' results is termed heterogeneity. Statistical heterogeneity occurs when the difference among study results is greater than chance alone.

Researchers will report evidence of heterogeneity (or lack thereof) in a couple ways:

1. Chi-squared test. Heterogeneity will produce a p-value  $< .05$ . Therefore, a higher p-value (e.g.,  $> .05$ ) is evidence of homogeneity
2. Degree of inconsistency ( $I^2$ ):
  - a. 0% to 40%: might not be important;
  - b. 30% to 60%: may represent moderate heterogeneity;
  - c. 50% to 75%: may represent substantial heterogeneity;
  - d. 75% to 100%: considerable heterogeneity.
3. Below we will discuss a visual method of identifying heterogeneity.

#### *b. If the results were heterogenous, is there a reasonable explanation?*

Authors should try to identify a reason for the heterogeneity. It may be different study populations, study quality, etc.

## **E x a m p l e**

A meta-analysis compared administering an asthma drug by two different methods. The analysis found significant heterogeneity among the studies. The authors reasoned that the two methods might work differently in adults than in children. When they separated the results by age (children vs. adults), the heterogeneity was removed. The two administration methods were found to be equivalent in adults but not in children.

#### *c. Were the populations, interventions, outcomes, and outcome measurements combined in a way that makes intuitive sense?*

We often call this the “apples and oranges” issue. Meta-analysis only makes sense when combining results in a way that makes sense. A recent meta-analysis combined all “alternative medicine” approaches to treatment of a specific illness into a single analysis, which does not make intuitive sense.

#### ***d. Could publication bias have occurred?***

Publication bias is the likelihood that negative results, i.e., studies that do not show a difference or benefit of a treatment, are less likely to be published.

*Why?*

1. Journal editors and journal reviewers are less interested in studies that “don’t show anything.”
2. A study is small, or smaller than previously published studies, and the results will not be interesting to readers.
3. Pharmaceutical companies suppress publication of research they paid for that isn’t flattering (see TED talk, below, and a fascinating report of an example [here](#)).
4. Researchers do not submit research for publication because they know about #1.

*So what?*

Since publication bias favors publication of research showing a benefit, a meta-analysis combining on published studies could inflate the real benefit of an intervention.


*How to detect?*

Researchers conducting meta-analysis can analyze the data to determine whether the risk of publication bias is high.

Statistically, different results from studies of the same topic should form a normal distribution (Gaussian curve) around the average calculated from those studies. That is, some results from individual studies should be below the mean, and some should be above. Also, the smaller the study, the greater the inherent variability of the data and the more likely the study is farther away from the mean.

A funnel plot compares the effect size in different studies with some measure of the variability of the data from each study. The example below compares the effect size as measured by odds ratio with the standard error (SE). Sample size is often used. Studies with small standard error cluster near the mean and studies with a larger SE are farther away. A “funnel” formed by the data that is balanced on both sides of the mean shows there was no publication bias. In the example that follows, the funnel is missing data at the bottom, left side, which is indicative of publication bias. Statistics can also be used to determine whether publication bias is likely.



 Watch [TED Talk on the Effect of Publication Bias and Evidence Suppression](#) (13:29)

 [Animation explaining publication bias](#) (3:02)

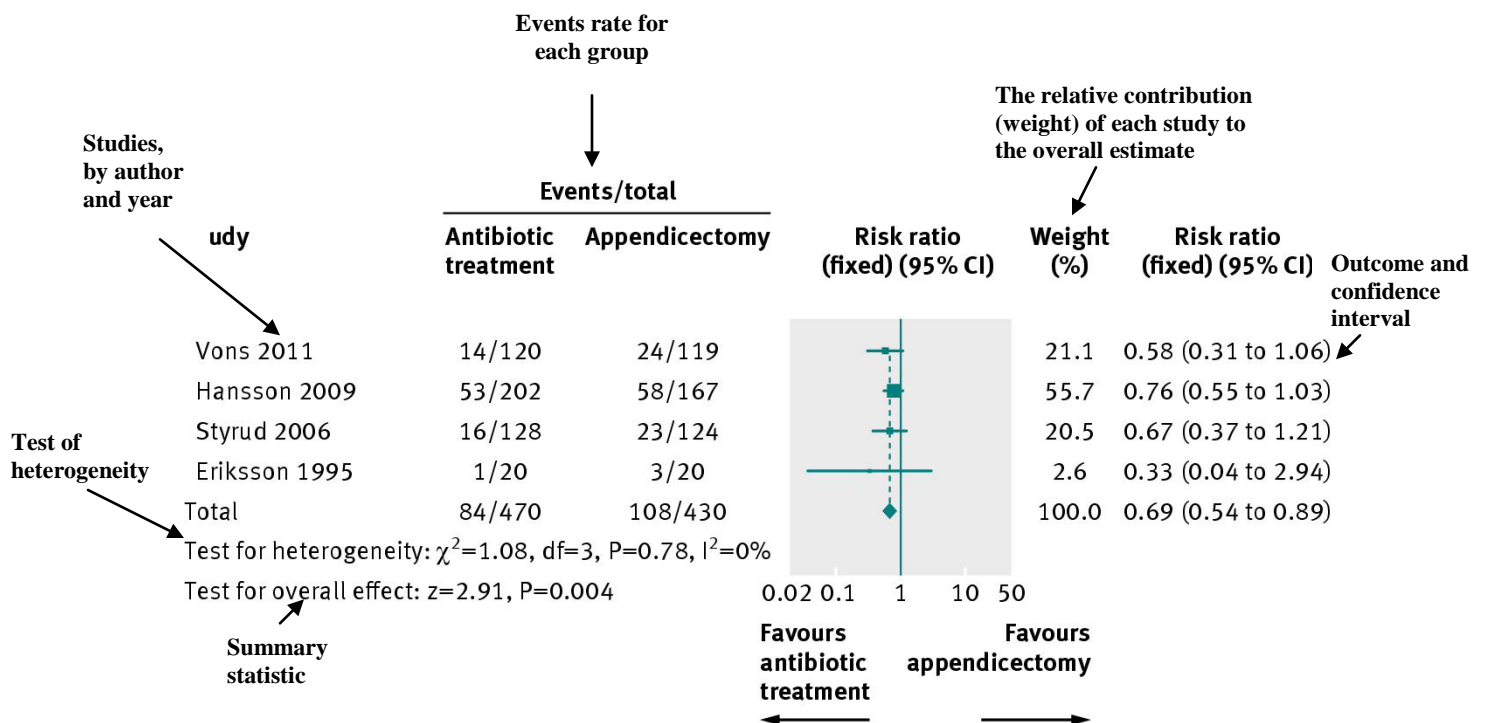
## Interpreting the Box – Understanding the forest plot

The results of a meta-analysis are reported in a complex figure called a forest plot. It graphically illustrates the data from individual studies, their relative strength, and how the studies combine into a single result. The results of studies are reported in rows of numbers and graphically using a combination of boxes representing the result and horizontal lines representing the confidence intervals around the result. The resulting “forest of lines” is where the graph gets its name.



Read “[Interpreting and Understanding Meta-Analysis Graphs](#)”

A typical presentation from a meta-analysis in a forest plot:





Note that this is a log scale and differences get rapidly larger as the result moves from 1.

Each study's relative risk is presented as a box and horizontal line.

- The **box** represents the relative risk and its size conveys the relative weight of that study.
- The **horizontal bars** represent the confidence interval for that study.
- The **diamond** at the bottom is the result for the combined study results, with the horizontal points representing the confidence interval.

Interpretation:

- Study results with **horizontal bars** (confidence intervals) crossing the solid vertical line are not statistically significant.
- The **vertical dashed line** shows the relationship of combined result to the results for each study (heterogeneity). For this plot, there is little difference between the combined result and any of the individual studies.

[Jump to instructions](#)

# Evaluating the Usefulness of Review Articles

## Determine *Validity*

### A. **Finding** the studies?

- Were the methods used to **locate** relevant studies comprehensive and clearly stated? ..Yes No
- Did they clearly **outline** study inclusion criteria that generalize to my practice?.....Yes No
- Was the study selection independently performed by at least **two** investigators? .....Yes No

### E. **Validity**: Was the validity of the original studies appropriately assessed?

- Were the validity criteria **appropriate**? .....Yes No
- Was the assessment **independently** performed by at least two investigators?.....Yes No
- How were the **validity determinations** used?
  - If studies were *excluded*, were the criteria reasonable? .....Yes No
  - If all studies were *included*, did the authors perform a subanalysis based on study quality or sufficiently explain the influence? .....Yes No

### F. **Analyzing** the Data: Is it reasonable to combine these studies?

- Were the studies **reasonably** valid? .....Yes No
- Were the included studies statistically **homogenous**? If not, did they provide an adequate explanation to account for the heterogeneity? .....Yes No
- Were the populations, interventions, outcomes, and outcome measurements **combined** in a way that makes intuitive sense? .....Yes No
- Could **publication bias** have occurred? .....Yes No





**Morbidity and Mortality Weekly Report**

[www.cdc.gov/mmwr](http://www.cdc.gov/mmwr)

---

Recommendations and Reports

June 25, 2010 / Vol. 59 / No. RR-5

---

**Updated Guidelines for Using  
Interferon Gamma Release Assays  
to Detect *Mycobacterium tuberculosis*  
Infection – United States, 2010**

The *MMWR* series of publications is published by the Office of Surveillance, Epidemiology, and Laboratory Services, Centers for Disease Control and Prevention (CDC), U.S. Department of Health and Human Services, Atlanta, GA 30333.

**Suggested Citation:** Centers for Disease Control and Prevention. [Title]. MMWR 2010;59(No. RR-#):[inclusive page numbers].

### Centers for Disease Control and Prevention

Thomas R. Frieden, MD, MPH  
*Director*

Harold W. Jaffe, MD, MA  
*Associate Director for Science*

James W. Stephens, PhD  
*Office of the Associate Director for Science*

Stephen B. Thacker, MD, MSc  
*Deputy Director for  
Surveillance, Epidemiology, and Laboratory Services*

### Editorial and Production Staff

Frederic E. Shaw, MD, JD  
*Editor, MMWR Series*

Christine G. Casey, MD  
*Deputy Editor, MMWR Series*

Teresa F. Rutledge  
*Managing Editor, MMWR Series*

David C. Johnson  
*Lead Technical Writer-Editor*

Jeffrey D. Sokolow, MA  
*Project Editor*

Martha F. Boyd  
*Lead Visual Information Specialist*

Malbea A. LaPete  
Stephen R. Spriggs  
Terraye M. Starr  
*Visual Information Specialists*

Quang M. Doan, MBA  
Phyllis H. King  
*Information Technology Specialists*

### Editorial Board

William L. Roper, MD, MPH, Chapel Hill, NC, Chairman

Virginia A. Caine, MD, Indianapolis, IN

Jonathan E. Fielding, MD, MPH, MBA, Los Angeles, CA

David W. Fleming, MD, Seattle, WA

William E. Halperin, MD, DrPH, MPH, Newark, NJ

King K. Holmes, MD, PhD, Seattle, WA

Deborah Holtzman, PhD, Atlanta, GA

John K. Iglehart, Bethesda, MD

Dennis G. Maki, MD, Madison, WI

Patricia Quinlisk, MD, MPH, Des Moines, IA

Patrick L. Remington, MD, MPH, Madison, WI

Barbara K. Rimer, DrPH, Chapel Hill, NC

John V. Rullan, MD, MPH, San Juan, PR

William Schaffner, MD, Nashville, TN

Anne Schuchat, MD, Atlanta, GA

Dixie E. Snider, MD, MPH, Atlanta, GA

John W. Ward, MD, Atlanta, GA

### CONTENTS

Introduction .....	1
Methods for Updating IGRA Guidelines .....	2
Background .....	2
The Epidemiology of Tuberculosis and <i>M. tuberculosis</i> Infection .	2
Development of Interferon Gamma Release Assays (IGRAs) and Interpretation Criteria .....	2
FDA-Approved Intended Use for IGRAs .....	5
Assessment of QFT-GIT and T-Spot Accuracy, Specificity, and Sensitivity .....	5
Use of QFT-GIT and T-Spot in Contact Investigations .....	7
Value of QFT-GIT and T-Spot in Predicting Subsequent Active Tuberculosis .....	7
Use of QFT-GIT and T-Spot for Testing Children .....	8
Use of QFT-GIT and T-Spot for Testing Immunocompromised Persons .....	9
Considerations for Programs.....	9
Recommendations .....	10
General Recommendations for Use of IGRAs .....	10
Test Selection.....	10
Medical Management After Testing .....	12
Areas for Additional Research.....	12
References .....	13

# Updated Guidelines for Using Interferon Gamma Release Assays to Detect *Mycobacterium tuberculosis* Infection – United States, 2010

Prepared by

Gerald H. Mazurek, MD, John Jereb, MD, Andrew Vernon, MD, Phillip LoBue, MD, Stefan Goldberg, MD, Kenneth Castro, MD  
Division of Tuberculosis Elimination, National Center for HIV, STD, and TB Prevention, CDC

## Summary

In 2005, CDC published guidelines for using the QuantiFERON-TB Gold test (QFT-G) (Cellestis Limited, Carnegie, Victoria, Australia) (CDC. Guidelines for using the QuantiFERON-TB Gold test for detecting *Mycobacterium tuberculosis* infection, United States. MMWR;54[No. RR-15]:49–55). Subsequently, two new interferon gamma (IFN- $\gamma$ ) release assays (IGRAs) were approved by the Food and Drug Administration (FDA) as aids in diagnosing *M. tuberculosis* infection, both latent infection and infection manifesting as active tuberculosis. These tests are the QuantiFERON-TB Gold In-Tube test (QFT-GIT) (Cellestis Limited, Carnegie, Victoria, Australia) and the T-SPOT.TB test (T-Spot) (Oxford Immunotec Limited, Abingdon, United Kingdom). The antigens, methods, and interpretation criteria for these assays differ from those for IGRAs approved previously by FDA.

For assistance in developing recommendations related to IGRA use, CDC convened a group of experts to review the scientific evidence and provide opinions regarding use of IGRAs. Data submitted to FDA, published reports, and expert opinion related to IGRAs were used in preparing these guidelines. Results of studies examining sensitivity, specificity, and agreement for IGRAs and TST vary with respect to which test is better. Although data on the accuracy of IGRAs and their ability to predict subsequent active tuberculosis are limited, to date, no major deficiencies have been reported in studies involving various populations.

This report provides guidance to U.S. public health officials, health-care providers, and laboratory workers for use of FDA-approved IGRAs in the diagnosis of *M. tuberculosis* infection in adults and children. In brief, TSTs and IGRAs (QFT-G, QFT-GIT, and T-Spot) may be used as aids in diagnosing *M. tuberculosis* infection. They may be used for surveillance purposes and to identify persons likely to benefit from treatment. Multiple additional recommendations are provided that address quality control, test selection, and medical management after testing.

Although substantial progress has been made in documenting the utility of IGRAs, additional research is needed that focuses on the value and limitations of IGRAs in situations of importance to medical care or tuberculosis control. Specific areas needing additional research are listed.

## Introduction

Before 2001, the tuberculin skin test (TST) was the only practical and commercially available immunologic test for *Mycobacterium tuberculosis* infection approved in the United States (1). Recognition that interferon gamma (IFN- $\gamma$ ) plays a critical role in regulating cell-mediated immune responses to *M. tuberculosis* infection led to development of interferon gamma release assays (IGRAs) for the detection of *M. tuberculosis* infection (2–4). IGRAs detect sensitization to *M. tuberculosis*

by measuring IFN- $\gamma$  release in response to antigens representing *M. tuberculosis*. In 2001, the QuantiFERON-TB test (QFT) (Cellestis Limited, Carnegie, Victoria, Australia) became the first IGRA approved by the Food and Drug Administration (FDA) as an aid for diagnosing *M. tuberculosis* infection (5,6). In 2005, the QuantiFERON-TB Gold test (QFT-G) (Cellestis Limited, Carnegie, Victoria, Australia) became the second IGRA approved by FDA as an aid for diagnosing *M. tuberculosis* infection (7,8). CDC published guidelines for using QFT in 2003 and for using QFT-G in 2005 (6,8).

Updated IGRA guidelines are needed because since 2005, two new IGRAs have been approved by FDA, and several hundred peer-reviewed articles describing clinical studies of IGRAs have been published. This report provides updated guidance to U.S. public health officials, health-care providers, and laboratory workers for use of FDA-approved IGRAs in the diagnosis of *M. tuberculosis* infection in adults and children.

The material in this report originated in the National Center for HIV, STD, and TB Prevention, Kevin Fenton, MD, PhD, Director; and the Division of Tuberculosis Elimination, Kenneth G. Castro, MD, Director.

**Corresponding preparer:** Gerald H. Mazurek, MD, Division of Tuberculosis Elimination, National Center for HIV, STD, and TB Prevention, CDC, 1600 Clifton Rd., N.E., MS E-10, Atlanta, GA 30333. Telephone: 404-639-8174; Fax: 404-639-8961; E-mail: gym6@cdc.gov.

## Methods for Updating IGRA Guidelines

CDC identified relevant reports published through August 2008 by searching PubMed for articles written in English that listed “tuberculosis” as the major MeSH topic and that included either “QuantiFERON” or “T-Spot” in the title or abstract. CDC identified additional published reports by contacting test manufacturers and examining references listed in retrieved articles. These search methods identified 152 potentially relevant articles. CDC reviewed the methods used in each study to select 96 primary reports that provided data related to 1) sensitivity or specificity of QFT-GIT or T-Spot; 2) agreement of QFT-GIT and T-Spot results with each other or with TST results; 3) association of QFT-GIT or T-Spot results with risk for *M. tuberculosis* infection or subsequent active tuberculosis; or 4) evaluation of QFT-GIT or T-Spot use in contact investigations, immunocompromised persons, or children.

During August 4–5, 2008, CDC convened a meeting in Atlanta, Georgia, to consider the use of QFT-GIT and T-Spot in U.S. tuberculosis-control activities. At this meeting, tabulated study results, descriptive summaries, explanations by study authors, and commentaries from test manufacturers were presented to an Expert Committee\* comprising tuberculosis-control officials, clinicians, laboratorians, and leading researchers with IGRA expertise, together with representatives of the American Academy of Pediatrics, the American Thoracic Society, the Advisory Council for the Elimination of Tuberculosis, the Association of Public Health Laboratories, CDC, FDA, the Infectious Disease Society of America, the National Tuberculosis Controllers Association, Stop TB USA, the U.S. Army, the U.S. Air Force, and the Veterans Health Administration. Data from most of the 96 primary reports used by CDC as the evidence on which these guidelines are based were available for review by the expert committee either as published articles or articles accepted for publication. CDC asked members of the Expert Committee to provide written opinions regarding how FDA-approved IGRAs should be used.

CDC used the published reports, data submitted to FDA, the product package inserts, and expert opinion related to QFT-GIT and T-Spot to prepare these guidelines. CDC coordinated development of these guidelines with the American Academy of Pediatrics, the American Thoracic Society, and the Infectious Disease Society of America.

## Background

### The Epidemiology of Tuberculosis and *M. tuberculosis* Infection

Globally, nine million persons develop active disease attributable to *M. tuberculosis* infection annually, and one third of the world's population, approximately 2 billion persons, are thought to be latently infected with *M. tuberculosis* (9). Although persons with latent *M. tuberculosis* infection (LTBI) do not manifest overt symptoms of active tuberculosis and are not infectious, they are at increased risk for developing active disease and becoming infectious. Approximately two million persons die each year from active tuberculosis despite the existence of effective treatments for both latent infection and active disease.

The prevalence of active tuberculosis in the United States has declined from 6.2 cases per 100,000 persons in 1998 to 4.2 cases per 100,000 persons in 2008 (10). During 1998–2007, of the 153,555 persons in the United States who had received a diagnosis of active tuberculosis, 3,708 (2.4%) died before treatment for active tuberculosis was started, and 10,777 (7.0%) died after starting treatment but before treatment was completed (CDC, unpublished data, 2008). A TST survey in 2000 indicated that an estimated 11,213,000 U.S. residents (4.2% of the civilian, noninstitutionalized U.S. population aged >1 year) had LTBI, representing a 60% decline from 1972 (11). However, the declines were not uniform among all segments of the U.S. population, and rates of *M. tuberculosis* infection and active tuberculosis vary considerably. Categorization of the risk for infection (Box 1) and for progression to active disease (Box 2) facilitates targeted testing and selection of those persons likely to benefit from treatment for latent infection (12). Identification of persons who are at increased risk for a poor clinical outcome (e.g., meningitis, disseminated disease, or death) if active tuberculosis occurs (Box 2) is an important component of targeted testing and treatment. U.S. residents with none of the recognized risk characteristics are considered to be at low risk for both infection and disease from *M. tuberculosis*. The prevalence of *M. tuberculosis* infection among such persons is estimated to be ≤1% (11).

### Development of Interferon Gamma Release Assays (IGRAs) and Interpretation Criteria

TSTs have been used worldwide for more than a century as an aid in diagnosing both LTBI and active tuberculosis. A positive TST result is associated with an increased risk for current or future active tuberculosis (13–16). However, certain limitations

\*The names of the members of the IGRA Expert Committee and the IGRA Expert Committee presenters appear on page 25 of this report.

**BOX 1. Risk factors for *Mycobacterium tuberculosis* infection**

Persons at increased risk\* for *M. tuberculosis* infection

- close contacts of persons known or suspected to have active tuberculosis;
- foreign-born persons from areas that have a high incidence of active tuberculosis (e.g., Africa, Asia, Eastern Europe, Latin America, and Russia);
- persons who visit areas with a high prevalence of active tuberculosis, especially if visits are frequent or prolonged;
- residents and employees of congregate settings whose clients are at increased risk for active tuberculosis (e.g., correctional facilities, long-term care facilities, and homeless shelters);
- health-care workers who serve clients who are at increased risk for active tuberculosis;
- populations defined locally as having an increased incidence of latent *M. tuberculosis* infection or active tuberculosis, possibly including medically underserved, low-income populations, or persons who abuse drugs or alcohol; and
- infants, children, and adolescents exposed to adults who are at increased risk for latent *M. tuberculosis* infection or active tuberculosis.

**Source:** Based on CDC. Targeted tuberculin testing and treatment of latent tuberculosis infection. MMWR 2000;49(No. RR-6).

\* Persons with these characteristics have an increased risk for *M. tuberculosis* infection compared with persons without these characteristics.

**BOX 2. Risk factors for progression of infection to active tuberculosis**

Persons at increased risk\* for progression of infection to active tuberculosis include

- persons with human immunodeficiency virus (HIV) infection;<sup>†</sup>
- infants and children aged <5 years;<sup>†</sup>
- persons who are receiving immunosuppressive therapy such as tumor necrosis factor- $\alpha$  (TNF- $\alpha$ ) antagonists, systemic corticosteroids equivalent to  $\geq 15$  mg of prednisone per day, or immune suppressive drug therapy following organ transplantation;<sup>†</sup>
- persons who were recently infected with *M. tuberculosis* (within the past 2 years);
- persons with a history of untreated or inadequately treated active tuberculosis, including persons with fibrotic changes on chest radiograph consistent with prior active tuberculosis;
- persons with silicosis, diabetes mellitus, chronic renal failure, leukemia, lymphoma, or cancer of the head, neck, or lung;
- persons who have had a gastrectomy or jejunioileal bypass;
- persons who weigh <90% of their ideal body weight;
- cigarette smokers and persons who abuse drugs or alcohol; and
- populations defined locally as having an increased incidence of active tuberculosis, possibly including medically underserved or low-income populations

**Source:** Based on CDC. Targeted tuberculin testing and treatment of latent tuberculosis infection. MMWR 2000;49(No. RR-6).

\* Persons with these characteristics have an increased risk for progression of infection to active tuberculosis compared with persons without these characteristics.

<sup>†</sup> Indicates persons at increased risk for a poor outcome (e.g., meningitis, disseminated disease, or death) if active tuberculosis occurs.

are associated with the use of TSTs. A valid TST requires proper administration by the Mantoux method with intradermal injection of 0.1 mL of tuberculin-purified protein derivative (PPD) into the volar surface of the forearm. In addition, patients must return to a health-care provider for test reading, and inaccuracies and bias exist in reading the test. Also, false-positive TSTs can result from contact with nontuberculous mycobacteria or vaccination with Bacille Calmette-Guerin (BCG), because the TST test material (PPD) contains antigens that are also in BCG and certain nontuberculous mycobacteria (13,17,18).

In 2001, QFT became the first IGRA approved by FDA as an aid for diagnosing *M. tuberculosis* infection (5,6). This test used an enzyme-linked immunosorbent assay (ELISA) to measure the amount of IFN- $\gamma$  released in response to PPD compared with controls. CDC issued guidelines on the use of QFT in 2003 (6). However, QFT specificity was less than that of TST despite the use of *M. avium* antigen as a control for nontuberculous mycobacterial sensitization and saline as

a negative control (19). QFT has not been available commercially since 2005.

To improve specificity, new IGRAs were developed. These IGRAs assess response to synthetic overlapping peptides that represent specific *M. tuberculosis* proteins, such as early secretory antigenic target-6 (ESAT-6) and culture filtrate protein 10 (CFP-10). These proteins are present in all *M. tuberculosis* and they stimulate measurable release of IFN- $\gamma$  in most infected persons, but they are absent from BCG vaccine strains and from most nontuberculous mycobacteria (20). Thus, as test antigens, these proteins offer improved test specificity com-



pared with PPD. However, ESAT-6 and CFP-10 are present in *M. kansasii*, *M. szulgai*, and *M. marinum*, and sensitization to these organisms might contribute to the release of IFN- $\gamma$  in response to these antigens and cause false-positive IGRA results. Because ESAT-6 and CFP-10 are recognized by fewer T lymphocytes and stimulate less IFN- $\gamma$  release compared with PPD, a more sensitive ELISA than was used for QFT is required to measure IFN- $\gamma$  concentrations and responses to ESAT 6 and CFP-10.

In 2005, the QuantiFERON-TB Gold test (QFT-G) (Cellestis Limited, Carnegie, Victoria, Australia) became the second IGRA approved by FDA as an aid for diagnosing *M. tuberculosis* infection (7,8). It assesses the immunologic responsiveness of tested patients to ESAT-6 and CFP-10. For QFT-G, separate aliquots of fresh whole blood are incubated with controls and with two separate mixtures of peptides, one representing ESAT-6 and the other representing CFP-10. The amount of IFN- $\gamma$  released in response to ESAT-6 or CFP-10 (i.e., the ESAT-6 Response or the CFP-10 Response) is calculated as the difference in IFN- $\gamma$  concentration in plasma from blood stimulated with antigen minus the IFN- $\gamma$  concentration in plasma from blood incubated with saline (i.e., Nil). For QFT-G, the TB Response is the higher of the ESAT-6 Response or the CFP-10 Response. A stipulation for FDA approval was inclusion of interpretation criteria that addressed the potential for false-positive results accompanying high Nil values (i.e., >0.7 IU/ml).

In 2005, CDC issued guidelines for using QFT-G (8), but the criteria that addressed interpretation when Nil values are high were subsequently revised (Table 1) (21). The 2005 QFT-G guidelines indicated that QFT-G may be used in all circumstances in which a TST was recommended, including contact investigations, evaluation of recent immigrants, and serial-testing surveillance programs for infection control (e.g., those for health-care workers) (8). The guidelines provided cautions for testing persons from selected populations, including persons at increased risk for progression to active disease if infected.

For IGRAs to measure IFN- $\gamma$  response accurately, a fresh blood specimen that contains viable white blood cells is needed. This requirement limited the use of early IGRAs to facilities in which trained laboratorians could begin testing blood within a few hours of its collection. The QuantiFERON-TB Gold In-Tube test (QFT-GIT) (Cellestis Limited, Carnegie, Victoria, Australia) was developed to address this limitation. In October 2007, QFT-GIT became the third IGRA approved by FDA as an aid for diagnosing *M. tuberculosis* infection (22). Control materials and antigens for QFT-GIT are contained in special tubes used to collect blood for the test, thus allowing more direct testing of fresh blood. One tube contains test antigens that consist of a single mixture of 14 peptides representing

the entire amino acid sequences of ESAT-6 and CFP-10 and part of the sequence of TB7.7. The two accompanying tubes serve as negative and positive controls: the negative-control tube contains heparin alone, and the positive-control tube contains heparin, dextrose, and phytohemagglutinin. Blood (1 ml) is collected into each of the three tubes, mixed with the reagents already in the tubes, and incubated for 16–24 hours. Plasma is separated, and the IFN- $\gamma$  concentration in the plasma is determined using the same sensitive ELISA used for QFT-G. To interpret QFT-GIT as approved by the FDA (Table 2), the TB Response is calculated as the difference in IFN- $\gamma$  concentration in plasma from blood stimulated with antigen (i.e., the single cocktail of peptides representing ESAT-6, CFP-10, and TB7.7) minus the IFN- $\gamma$  concentration in plasma from blood incubated without antigen (i.e., Nil).

QFT-GIT was evaluated in the United States and used in other countries prior to FDA approval in 2007, and users of the test promulgated a variety of interpretation criteria. Some published reports used criteria for QFT-GIT that were similar to those being used for QFT-G. As compared with FDA-approved QFT-G interpretation criteria (Table 1), the FDA criteria approved for QFT-GIT in 2007 (Table 2) interpret tests with a Nil of 0.7–8.0 and a TB Response of 25%–50% of Nil as positive rather than as indeterminate. Also, tests with a Nil of 0.7–8.0 and a TB Response that is <25% of Nil are interpreted as negative, whereas for QFT-G they are interpreted as indeterminate.

In July 2008, T-Spot became the fourth IGRA to be approved by FDA (23). For this test, peripheral blood mononuclear cells (PBMCs) are incubated with control materials and two mixtures of peptides, one representing the entire amino acid sequence of ESAT-6 and the other representing the entire amino acid sequence of CFP-10. The test uses an enzyme-linked immunospot assay (ELISpot) to detect increases in the number of cells that secrete IFN- $\gamma$  (represented as spots in each test well) after stimulation with antigen as compared to the media control (Nil). The T-Spot interpretation criteria approved by FDA for use in the United States (24) differ from those used in other countries (25). Also, the majority of published studies evaluating T-Spot have used criteria that differ from those approved by FDA. The 2008 FDA-approved interpretation criteria for T-Spot (Table 3) included a borderline interpretation for a TB Response equal to five, six, or seven spots. Use of a borderline category might address test variation and uncertainty for results near a dichotomous cut point. This might increase the assay's apparent specificity and sensitivity by minimizing false-positive and false-negative results near a dichotomous cut point. In addition, through the use of a borderline category, test conversions from negative to positive are more likely to represent a newly acquired infection.

## FDA-Approved Intended Use for IGRAs

FDA has approved both QFT-GIT and T-Spot as in vitro diagnostic aids for detection of *M. tuberculosis* infection (22,23). Both tests are approved as indirect tests for *M. tuberculosis* infection (including infection resulting in active disease) when used in conjunction with risk assessment, radiography, and other medical and diagnostic evaluations. The FDA-approved indications for QFT-GIT and T-Spot are similar to indications for QFT-G and TST using either Tubersol PPD (Sanofi Pasteur Ltd., Toronto, Ontario, Canada) or Aplisol PPD (JHP Pharmaceuticals, LLC, Rochester, Michigan). Because QFT-G, QFT-GIT, T-Spot, and TST each measure different aspects of the immune response and use different antigens and interpretation criteria, test results might not be interchangeable. Different tests can yield different results.

## Assessment of QFT-GIT and T-Spot Accuracy, Specificity, and Sensitivity

### Limitations in Assessing Accuracy

Assessments of accuracy of tests for *M. tuberculosis* infection are hampered by the lack of confirmatory tests to diagnose LTBI and culture-negative active tuberculosis. Accuracy is a measure of the proportion of test results that are correct and encompasses assessment of specificity (the proportion of true negatives that have negative test results) and sensitivity (the proportion of true positives that have positive test results). Assessments of accuracy of tests for *M. tuberculosis* infection are difficult because there is no “gold standard” to confirm a diagnosis of LTBI or culture-negative active tuberculosis. However, approximations of accuracy, sensitivity, and specificity can be made by testing populations with known characteristics. For example, to assess the sensitivity of IGRAs, researchers can observe the proportion of positive IGRA results among persons with culture-confirmed active tuberculosis, a group for whom the IGRA should be positive (i.e., true positives). Likewise, to assess the specificity of IGRAs, researchers can observe the proportion of negative IGRA tests among persons who are very unlikely to have *M. tuberculosis* infection (i.e., assumed negatives). Researchers also can characterize factors associated with discordance between different tests or conduct follow-up studies to determine the subsequent rate of active tuberculosis for persons with positive or negative IGRA results. However, although sensitivity and specificity are inherent characteristics of the tests, with no “gold standard,” estimates of test performance might fluctuate as a result of differences in the study population and the rate of diagnostic misclassification (e.g., as a result of differences in prevalence *M. tuberculosis*

and nontuberculous mycobacterial infection, malnutrition, and immune suppression). In addition, because TSTs and IGRAs are indirect tests that measure immunologic responses and are not direct tests that detect the causative organism or components of the organism, assessments of sensitivity among persons with culture-confirmed active tuberculosis might not provide reliable estimates of sensitivity for LTBI. Immunologic differences that allow progression of infection to disease might affect immunologic test results. In addition, treatment can alter immunologic responses and might alter test results. Estimates of specificity among low-risk populations might underestimate specificity because some persons might have infection resulting from unrecognized exposure.

Assessment of test accuracy is complicated further by the use of different test methods and interpretation criteria for TST, QFT-GIT, and T-Spot in published reports. Most published reports evaluating QFT-GIT or T-Spot accuracy (26–53) (Tables 4–7) have used interpretation criteria different from those approved by FDA. Also, in published studies in which IGRA results have been compared with TST results (27,28,31–41,43–45,49,50), the TST antigens and cut points in indurations used to separate negative and positive results differed. In addition, for evaluations of QFT-GIT, some investigators used methods that did not include a positive control for QFT-GIT (28,30), in contrast to the methods approved by FDA. Inclusion of a positive control increases estimates of sensitivity by excluding indeterminate results with low Mitogen Responses, which otherwise might be interpreted as negative. For example, if blood samples are processed improperly to the point that they lose the ability to produce IFN- $\gamma$  and a positive control is not used, the IGRA results for these samples will be interpreted as negative. With a positive control, they will be interpreted as indeterminate and not be included in the calculations of sensitivity (i.e., they will be removed from the denominator). This is similar to excluding persons who do not return to have their TST read from estimates of TST sensitivity.

Incorporation of a borderline category for the T-Spot as approved by FDA (Table 3) increases test accuracy by classifying results near the cut point (at which small variations might affect the interpretation) as neither positive or negative. Although not included in FDA-approved interpretation criteria for QFT-GIT (Table 2), an appropriate borderline category for QFT-GIT might increase its accuracy for the same reasons. Another tactic for improving detection sensitivity is to use any positive result from multiple tests, as is done with culture or nucleic acid amplification tests. Interpreting any positive result from multiple tests as evidence of infection typically increases detection sensitivity and decreases specificity. On the other hand, requiring positive results from two or more tests

typically has the opposite effect (i.e., decreasing sensitivity and increasing specificity).

### Estimates of Sensitivity

Estimates of QFT-GIT and T-Spot sensitivity have varied widely in published studies (Tables 4 and 5), which have involved predominantly adults with culture-confirmed active tuberculosis. In general, QFT-GIT and T-Spot sensitivities are considered similar to those for TST. However, caution is required when comparing test sensitivity from these studies because 1) some cohorts were not limited to subjects with microbiologically confirmed active tuberculosis (and in reality might not have had active tuberculosis); 2) in the majority of studies, head-to-head comparisons of IGRAs were not performed in the same subjects; and 3) test methods and interpretation criteria used in reported studies often differed from those approved by FDA.

When data from published studies related to QFT-GIT sensitivity in patients with culture-confirmed active tuberculosis (26–30,32,33,35,37–39) were pooled (Table 4) and sensitivity was determined as the number of subjects with positive QFT-GIT results divided by the number with positive or negative results, pooled QFT-GIT sensitivity was 81%, compared with 70% reported by a study that estimated sensitivity on the basis of a meta-analysis (54). In studies that compared the sensitivity of QFT-GIT to that of TST in patients with culture-confirmed active tuberculosis (27,28,32,33,35,37–39), pooled QFT-GIT sensitivity was 83% and pooled TST sensitivity was 89%. In the 11 studies that compared QFT-GIT and TST in patients in whom active tuberculosis (not necessarily culture-confirmed) was diagnosed, six studies (28,32,34,37–39) demonstrated no statistically significant difference between the two tests, three studies (27,31,33) demonstrated greater sensitivity for TST, and two studies (35,36) demonstrated greater sensitivity for QFT-GIT.

When data from published studies related to T-Spot sensitivity in patients with culture-confirmed active tuberculosis (28,33,38,42,46,48,50–52) were pooled (Table 5), and sensitivity was determined as the number of subjects with positive T-Spot results divided by the number with positive or negative results, pooled T-Spot sensitivity was 91%. In studies that compared the sensitivity of T-Spot to that of TST in patients with culture-confirmed tuberculosis (28,33,38,39,50), pooled T-Spot sensitivity was 90% and TST sensitivity was 89%. In the 12 studies that compared T-Spot and TST sensitivity in patients diagnosed with active tuberculosis (not necessarily culture-confirmed), nine demonstrated no statistically significant difference in the two tests (28,31,33,38,39,43,44,49,50), and three demonstrated greater sensitivity for T-Spot (39,40,44).

In three published studies that evaluated TST, QFT-GIT, and T-Spot (28,33,39), pooled sensitivity for TST, T-Spot, and QFT-

GIT were 95%, 91%, and 84%, respectively. The largest of these studies was conducted in Singapore and involved more than 270 persons with culture-confirmed active tuberculosis (33). In that study, the estimates of sensitivity of T-Spot and of TST (using a 10-mm cutoff) were similar (94% and 95% respectively;  $p=0.84$ ), and significantly greater than QFT-GIT (83%;  $p<0.01$ ).

### Estimates of Specificity

QFT-GIT and T-Spot are expected to be more specific than a TST because the antigens used in these tests are relatively specific to *M. tuberculosis* and should produce fewer false-positive tests (i.e., they should not produce cross-reactions after sensitization by BCG and most nontuberculous mycobacteria, such as *M. avium* complex). Estimates of QFT-GIT and T-Spot specificity in tested populations considered to be at low risk for *M. tuberculosis* infection generally are high (Tables 6 and 7). Caution is required when estimating and comparing test specificity from these studies because 1) the background risk for infection varied among studies, 2) the test methods and interpretation criteria used in the studies often differed from those approved by FDA, and 3) some persons classified as false positives might have infection resulting from unrecognized risk. Most studies comparing the specificity of QFT-GIT or T-Spot with TST have been conducted outside the United States.

In tested populations of persons unlikely to have *M. tuberculosis* infection, pooled QFT-GIT specificity was 99% (Table 6) (26,28,32,34), and pooled TST specificity from these cohorts, when available, was 85% (28,34). Pooled T-Spot specificity was 88% (Table 7) (28,40,53), and pooled TST specificity from these cohorts, when available, was 86% (28,40). Because of the small sample sizes in studies examining T-Spot specificity, additional independent studies are needed to increase the certainty of the T-Spot specificity estimate. The lower estimates of TST specificity compared with QFT-GIT and T-Spot might be attributable to false-positive TST results following BCG vaccination or exposure to nontuberculous mycobacteria. Lower estimates of TST specificity have been demonstrated for BCG-vaccinated cohorts, and in those with nontuberculous lymphadenitis (28,55,56). However, in a study in which cohorts with similar risks for infection were compared, the specificity of IGRA using ESAT-6 or CFP-10 did not differ significantly between those vaccinated with BCG and those not vaccinated (57). The effect of BCG on specificity is difficult to assess because BCG is used predominately in populations already at increased risk for *M. tuberculosis* infection.

### Agreement Among Tests

Agreement among tests for *M. tuberculosis* infection varies widely in reported studies (33,58–60). Agreement in these studies has been affected by test interpretation criteria, prevalence of



infection and the proportion of infections that are confirmed microbiologically, estimates of recent and remote exposure, age, race, prior BCG vaccination, recent TST, and coexisting diseases, including nontuberculous mycobacterial infection and conditions with immunosuppression (e.g., human immunodeficiency virus [HIV] infection). Increasing age is a risk for *M. tuberculosis* infection because of longer time for potential exposure and because older persons might have been alive when tuberculosis was more prevalent. The association of older age with positive TST and IGRA results generally is attributed to *M. tuberculosis* infections that accumulate over time. The observation in some studies that increasing age is associated more strongly with TST results than with IGRA results suggests that a TST might be more sensitive than IGRAs in detecting remote infections that occurred years earlier (58,61).

Investigations examining the effect of PPD injection on subsequent IGRAs have produced conflicting results (59,62–66); outcome differences probably are attributable to differences in the study population (infected versus noninfected subjects, recent versus temporally remote infection, and risk for ongoing exposure), timing of IGRA testing after PPD injection, the IGRA format, and the definition of boosting used. PPD injection should be expected to boost anamnestic immune responses measured by IGRA originating from *M. tuberculosis* infection, but not from BCG vaccination or in nonsensitized persons. Additional studies examining the effect of PPD injection on IFN- $\gamma$  responses are needed to define the frequency, magnitude, induction time, and longevity of IGRA boosting following a TST.

Uncertainty exists regarding the reproducibility of IGRA results in individual patients and the clinical significance of fluctuations in measured IFN- $\gamma$  responses. Longitudinal studies have revealed considerable fluctuation in IFN- $\gamma$  responses with serial testing in individual patients (59,62,63,65,67–71). These fluctuations might be attributed to limitations in the precision of IGRAs or to actual fluctuations in IFN- $\gamma$  responses in the patient. Some increases in IFN- $\gamma$  response might be attributed to new infection or boosting following a TST. Some decreases in IFN- $\gamma$  response in individual persons might be attributed to antimycobacterial treatment. However, for the most part, fluctuations in IFN- $\gamma$  responses among serially tested individual patients reported in longitudinal studies remain unexplained and nonspecific. The magnitude of these fluctuations can be of sufficient size to cause test interpretations to change from negative to positive (conversion) or from positive to negative (reversion), especially when the IFN- $\gamma$  responses are near cut points separating positive and negative results. Well-controlled studies are needed to further define the causes of individual variations in IFN- $\gamma$  response and to develop criteria to differentiate nonspecific variation from that associated with new or resolving infection.

## Use of QFT-GIT and T-Spot in Contact Investigations

Several reports of contact investigations have included results from QFT-GIT and T-Spot (Table 8) (30,31,58,61,72–74). In two of these investigations (58,73), greater recent exposure (as measured by duration of exposure or infectiousness of the source based on a higher number of acid-fast bacilli in their sputa) was more strongly associated with positive IGRA results than with positive TST results, suggesting that IGRAs might be better than the TST at detecting recent infection. In these studies, persons with lower amounts of recent exposure were more likely to be positive by TST than IGRA, suggesting that the TST might have been better than the IGRAs at detecting remote infection that was present prior to (and therefore did not occur as a result of) the recent exposure (58). In two other investigations (72,74), neither TST nor IGRA results were associated with measures of recent exposure. In another investigation (30), the proximity of recent exposure (i.e., same room, different room, or different house) was more strongly associated with TST results than QFT-GIT results.

## Value of QFT-GIT and T-Spot in Predicting Subsequent Active Tuberculosis

Of critical importance, is a test's ability to predict risk for subsequent active tuberculosis. For a person with a positive TST, the lifetime risk for active tuberculosis is estimated to be 5%–10% (16,75). However, very few longitudinal data exist on the ability of IGRAs to predict risk for subsequent active tuberculosis.

In one study in Germany involving 601 close contacts of persons with smear-positive, culture-confirmed active tuberculosis, QFT-GIT was reported to perform better than a TST using a 5 mm cut point in predicting subsequent active tuberculosis (76). Whereas five (2.3%) of 219 contacts with TST induration  $\geq 5$  mm developed tuberculosis, six (14.6%) of 41 contacts with positive QFT-GIT results developed the disease ( $p=0.003$ ). However, an unusually large proportion (59%) of the contacts had TST induration that ranged from 5 mm to 9 mm. The proportion of those considered positive by TST using a 10 mm cutoff who developed active tuberculosis (five of 90 [5.6%]) was similar to the proportion positive by QFT-GIT (six of 41 [14.6%];  $p=0.1$ ). In addition, only two of the six contacts with positive QFT-GIT results who developed active tuberculosis had the diagnosis confirmed by culture. As noted in a published comment on the article, the sensitivity for predicting subsequent active tuberculosis did not differ significantly for the two tests (77). The QFT-GIT sensitivity was 100% (95% confidence interval [CI] = 54%–100%) and

the TST sensitivity was 83% (CI = 36%–100%) ( $p=0.50$ ) using either a 5 mm or a 10 mm TST cut point.

Results from another study indicated that active tuberculosis developed in three of 36 (8.3%) HIV-infected persons who had positive QFT-GIT results at baseline and in none of 705 HIV-infected persons with negative QFT-GIT results at baseline during a median of 19 months of active follow up ( $p<0.001$ ) (37). TST was performed for a subset of subjects who had positive QFT-GIT results. TST was positive for all of the tested subjects who developed active tuberculosis.

In a study of 339 immigrants to the Netherlands, TST and QFT-GIT were reported to perform similarly in predicting subsequent active tuberculosis (78). Contacts whose TST was  $\geq 5$  mm at 0 or 3 months after diagnosis of the index patient were followed for up to 2 years. Nine (3.1%) of 288 contacts with TST  $\geq 10$  mm developed active tuberculosis whereas seven (3.8%) of 184 with TST  $\geq 15$  mm, five (2.8%) of 178 with a positive QFT-GIT, and six (3.3%) of 181 with a positive T-Spot developed active tuberculosis. The proportions of contacts with positive results by the different tests who developed active tuberculosis were not statistically different. The sensitivity for subsequent active tuberculosis during the period of follow-up was 100% for a TST using a 10 mm cutoff, 88% for a TST using a 15 mm cutoff, 63% for QFT-GIT, and 75% for a T-Spot. While TST using a 10 mm cutoff identified the greatest number of contacts who developed active tuberculosis (nine of nine [100%]), and QFT-GIT identified the lowest number of contacts who developed active tuberculosis (five of nine [63%]), the sensitivity of the two tests were not statistically different ( $p=0.08$ ).

In another large study, an ELISpot assay that was developed by the investigators to detect responses to ESAT-6 and CFP-10 was used to study tuberculosis household contacts in The Gambia. The ELISpot assay was positive for 11 (52%) of 21 secondary cases of active tuberculosis, compared with 14 (56%) of 25 secondary cases who were positive by TST (79). Of the 21 persons with secondary cases tested with both tests, 15 (71%) were positive by at least one of the tests. Although this proportion was not significantly greater than the proportion positive by TST alone (56%;  $p=0.2$ ), the study indicated that positivity by either test might be the best indication for preventive treatment in this setting. Additional, larger studies are needed to estimate more accurately the performance of IGRA tests compared with TSTs.

## Use of QFT-GIT and T-Spot for Testing Children

Assessment of the accuracy of IGRAs has been more difficult in children than in adults because study enrollment is more

complicated, phlebotomy is more difficult in younger children, microbiologic confirmation of infection is less frequent, and BCG might have been administered more recently. This is especially true for children aged  $<5$  years. Few performance data exist for QFT-GIT and T-Spot testing in children (especially for those aged  $<5$  years). For this reason, and because rates of progression from latent infection to active disease (including severe forms of the disease, such as meningitis, disseminated disease, or death as a result of *M. tuberculosis*) are higher in infants and young children, caution is warranted when using IGRAs in children aged  $<5$  years (80).

The higher rate of active tuberculosis and severe forms of the disease in infants and children aged  $<5$  years compared with older children suggests that the immune response to *M. tuberculosis* infection differs in these groups. Age-related immunologic differences might explain reported variations in IGRA test performance, including poorer test sensitivity, and lower production of IFN- $\gamma$  in response to mycobacterial antigens and mitogen (used as a positive control) among children aged  $<4$  years compared with children aged 4–15 years (81), an increase in response to mitogen with increasing age (82), and a higher proportion of indeterminate QFT-GIT results among children aged  $<5$  years (43). In contrast, one large study in a tuberculosis-endemic setting found that infants and young children had robust IFN- $\gamma$  responses to *M. tuberculosis* antigens, and that their responses were comparable to responses in adults and older children (83).

Older children (i.e., those aged  $\geq 5$  years) are less likely than children aged  $<5$  years to develop active tuberculosis or to have severe forms of the disease; in this way, older children resemble adults. In addition, for older children, IGRA testing might be logistically easier (e.g., in the ability to draw sufficient quantities of blood). Therefore, less caution might be required when implementing IGRA testing in children aged  $\geq 5$  years than in children aged  $<5$  years.

Use of IGRAs in children is subject to several limitations. First, studies evaluating IGRAs performance in children are scant. In only a few studies are separate results provided for children, and even fewer studies divide results by narrow age categories. This means that IGRA performance in children is less well understood than IGRA performance in adults. Second, indeterminate results for children are a potential limitation to implementing IGRAs into clinical practice. The frequencies of indeterminate IGRA results in children vary greatly among studies (range: 0–17%) and between different IGRA formats (31,39,43,84–89). Although the majority of indeterminate results are attributable to a low Mitogen Response, the reasons for low Mitogen Responses in young children are unclear. The mitogen might not work well in young children as a result of a lack of immunologic maturity. Differences in the mitogen

concentration used for stimulation and differences in interpretation criteria can affect the number of indeterminate results, especially when different IGRA formats are compared. Third, concerns relate to difficulties in collecting blood for these tests and the need for a relatively large volume of blood from small children (especially for infants). Finally, certain pediatricians have expressed concern that IGRAs might have lower sensitivity than TSTs in children (81,90,91).

In general, sensitivity of IGRAs in children is expected to be comparable to TST. In one study of 28 children with culture-confirmed active tuberculosis who were aged 4 months–7 years, estimates of sensitivity for TST, QFT-GIT, and T-Spot were comparable at 100%, 93%, and 93% respectively ( $p=0.15$ ) (28). Sensitivities of these tests were also similar in another study of nine children who had active tuberculosis; six (67%) were positive by T-Spot, six (67%) were positive by QFT-GIT, and nine (100%) were positive by TST (31). In another study involving 25 children with culture-confirmed active tuberculosis, estimates of sensitivity were 88% for TST at 10 mm and 83% for TST at 15 mm, 80% for QFT-GIT, and 58% for T-Spot (39). In the same study, when children with probable active tuberculosis were included (defined on the basis of epidemiologic, clinical, and radiographic findings in the absence of a positive culture), sensitivity for TST at 10 mm fell to 71%, sensitivity for TST at 15 mm fell to 60%, and sensitivity for QFT-GIT and T-Spot fell to 64% and 50%, respectively. However, the methods used for diagnosing active tuberculosis in this study were not stated specifically and might have included use of TST results. In another study that evaluated 154 children aged 5–15 years with culture-confirmed active tuberculosis, results indicated that TST was more sensitive than QFT-GIT (90% and 76%, respectively;  $p<0.01$ ) (27).

In general, specificity of IGRAs in children is expected to be high. For example, QFT-GIT and T-Spot demonstrated high specificity for *M. tuberculosis* infection even among children whose TST specificity was reduced to 22% because of nontuberculous mycobacterial infections (28). Additional larger studies are needed to evaluate the performance of IGRAs in children.

## Use of QFT-GIT and T-Spot for Testing Immunocompromised Persons

Limited data are available regarding the use of QFT-GIT for testing immunocompromised persons (Table 9) (27,36,37,92–100). In two studies with a total of 34 HIV-infected subjects with culture-confirmed active tuberculosis, the sensitivities of QFT-GIT were 81% and 88% (27,37). In one study, the sensitivities of QFT-GIT and TST were similar (81% and 85% respectively,  $p>0.99$ ) (27). QFT-GIT sensitivity was not

significantly different among persons with HIV infection than among those without infection (81% and 73%, respectively;  $p=0.59$ ). In another study in Zambia involving 112 persons (59 were infected with HIV, 37 were not infected with HIV, and 16 were not tested) in whom active tuberculosis was diagnosed on the basis of sputum smear (36), QFT-GIT and TST were significantly less sensitive in persons infected with HIV than in persons not infected with HIV (76% compared with 97% for QFT-GIT;  $p=0.02$  and 55% compared with 81% for TST,  $p=0.04$ ). Among persons with HIV infection, QFT-GIT sensitivity tended to be higher than TST sensitivity (76% and 55%, respectively;  $p=0.06$ ). However, in this study, reduced TST sensitivity might have resulted from delayed reading of TSTs, which were read 48–164 hours after PPD injection. Low CD4 counts were associated with increases in false-negative TST results and indeterminate and false-negative QFT-GIT results.

Published comparisons have not demonstrated significant differences in the proportion of positive QFT-GIT results as compared with the proportion of positive TST results among HIV-infected persons screened for *M. tuberculosis* infection (93–96). QFT-GIT results from two studies suggest that the proportion of indeterminate QFT-GIT results among HIV-infected persons (17% and 19%, respectively) is similar to the proportion among uninfected persons (14% and 0, respectively;  $p=0.88$  and  $p=0.18$ , respectively) (27,36). However, in another study among HIV-infected persons, CD4 counts were lower in those with indeterminate QFT-GIT results as compared with those with positive or negative results ( $p<0.01$ ) (37). Among persons with other immunosuppressive conditions, published comparisons do not show consistent agreement between results of QFT-GIT and those of TST (97–100). Without a diagnostic “gold standard” for LTBI, the accuracies of both the QFT-GIT and the TST are uncertain.

Information related to T-Spot in immunocompromised persons has been provided in relatively few published reports (Table 10) (60,96,97,101–108) with very little information related to test sensitivity in such persons (Table 5). Among persons with various immunosuppressive conditions being screened for *M. tuberculosis* infection, published comparisons of T-Spot with TST generally demonstrate either similar proportions of positive results (60,96,97,101,104,108) or that T-Spot is more often positive (103,105–107). Without a diagnostic “gold standard” for LTBI, the accuracies of both the TST and the T-Spot are suspect.

## Considerations for Programs

Because of administrative and logistic difficulties associated with the TST, IGRAs are attractive diagnostic aids for detect-



ing *M. tuberculosis* infection. Unlike TSTs, IGRA results can be available within 24 hours without the need for a second visit. As laboratory-based assays, IGRAs are not subject to the biases and errors associated with TST placement and reading. However, errors in collecting, labeling, or transporting blood specimens, or while performing and interpreting these assays can decrease IGRA accuracy. Also, availability of IGRAs is limited by the need for a fresh blood sample and the potential for delays as a result of the long distances to laboratories that offer these tests.

The cost for an IGRA is substantially greater than that for a TST (109). However, this additional cost might be offset by decreases in the number of persons testing positive and the associated costs of evaluating and treating persons with positive test results (110). Use of an IGRA might increase acceptance of treatment for LTBI (111). However, cost-effectiveness studies are limited by the lack of critical data on the relative ability of these tests to predict subsequent disease.

## Recommendations

### General Recommendations for Use of IGRAs

- TSTs and IGRAs (QFT-G, QFT-GIT, and T-Spot) should be used as aids in diagnosing infection with *M. tuberculosis*. These tests may be used for surveillance purposes or to identify persons likely to benefit from treatment, including persons who are or will be at increased risk for *M. tuberculosis* infection (Box 1) or for progression to active tuberculosis if infected (Box 2).
- IGRAs should be performed and interpreted according to established protocols using FDA-approved test formats. They should be performed in compliance with Clinical Laboratory Improvement Amendment (CLIA) standards.
- Both the standard qualitative test interpretation and the quantitative assay measurements should be reported together with the criteria used for test interpretation. This will permit more refined assessment of results and promote understanding of the tests.
- Arrangement for IGRA testing should be made prior to blood collection to ensure that the blood specimen is collected in the proper tubes, and that testing can be performed within the required timeframe.
- Prior to implementing IGRAs, each institution and tuberculosis-control program should evaluate the availability, overall cost, and benefits of IGRAs for their own setting. In addition, programs should consider the characteristics of the population to be tested.
- As with the TST, IGRAs generally should not be used for testing persons who have a low risk for both infection and

progression to active tuberculosis if infected (except for those likely to be at increased risk in the future). Screening such persons diverts resources from higher priority activities and increases the number of false-positive results. Even with a test specificity approaching 99%, when the prevalence of *M. tuberculosis* infection is  $\leq 1\%$ , the majority of positive results will be false positives. If persons at low risk for both infection and progression are to be tested, selection of the test with the greatest specificity will minimize false-positive results, reduce unnecessary evaluation and treatment, and minimize the potential for adverse events from unnecessary treatment.

### Test Selection

- Selection of the most suitable test or combination of tests for detection of *M. tuberculosis* infection should be made on the basis of the reasons and the context for testing, test availability, and overall cost effectiveness of testing. Results of studies examining sensitivity, specificity, and agreement for IGRAs and TST vary with respect to which test is better. Although data on the accuracy of IGRAs and their ability to predict subsequent active tuberculosis are limited, to date, no major deficiencies have been reported in studies involving various populations. As use of these tests increases, greater understanding of their value and limitations will be gained.
- An IGRA may be used in place of (but not in addition to) a TST in all situations in which CDC recommends tuberculin skin testing as an aid in diagnosing *M. tuberculosis* infection, with preferences and special considerations noted below. Despite the indication of a preference in these instances, use of the alternative test (FDA-approved IGRA or TST) is acceptable medical and public health practice.

### Situations in Which an IGRA Is Preferred But a TST Is Acceptable

- An IGRA is preferred for testing persons from groups that historically have low rates of returning to have TSTs read. For example, use of an IGRA might increase test completion rates for homeless persons and drug-users. The use of IGRAs for such persons can increase test completion rates, so control efforts can focus on those most likely to benefit from further evaluation and treatment.
- An IGRA is preferred for testing persons who have received BCG (as a vaccine or for cancer therapy). Use of IGRAs in this population is expected to increase diagnostic specificity and improve acceptance of treatment for LTBI.

### Situations in Which a TST Is Preferred But an IGRA Is Acceptable

- A TST is preferred for testing children aged <5 years. Use of an IGRA in conjunction with TST has been advocated by some experts to increase diagnostic sensitivity in this age group. Recommendations regarding use of IGRAs in children have also been published by the American Academy of Pediatrics (112).

### Situations in Which Either a TST or an IGRA May Be Used Without Preference

- An IGRA or a TST may be used without preference to test recent contacts of persons known or suspected to have active tuberculosis with special considerations for follow-up testing. IGRAs offer the possibility of detecting *M. tuberculosis* infection with greater specificity than with a TST. Also, unlike TSTs, IGRAs do not boost subsequent test results and can be completed following a single patient visit. However, data on the ability of IGRAs to predict subsequent active tuberculosis are limited. If IGRAs are to be used in contact investigations, negative results obtained prior to 8 weeks after the end of exposure typically should be confirmed by repeat testing 8–10 weeks after the end of exposure. This recommendation is similar to one used for TST, because data on the timing of IGRA conversion after a new infection are not currently available. Use of the same test format for repeat testing will minimize the number of conversions that occur as a result of test differences.
- An IGRA or a TST may be used without preference for periodic screening of persons who might have occupational exposure to *M. tuberculosis* (e.g., surveillance programs for health-care workers) with special considerations regarding conversions and reversions. For serial and periodic screening, IGRAs offer technical, logistic, and possible economic advantages compared with TSTs but also have potential disadvantages. Advantages include the ability to get results following a single visit. Two-step testing is not required for IGRAs, because IGRA testing does not boost subsequent test results. Disadvantages of IGRAs in this setting include a greater risk of test conversion due to false-positive IGRA results with follow-up testing of low-risk health-care workers who have tested negative at prior screening. CDC has published criteria for identifying conversions for TSTs and IGRAs (113). TST conversion is defined as a change from negative to positive with an increase of  $\geq 10$  mm in induration within 2 years. TST conversion is associated with an increased risk for active tuberculosis. An IGRA conversion is defined as a change from negative to positive within 2 years without any consideration of the magnitude of the change in TB Response. Using this lenient criterion

to define IGRA conversion might produce more conversions than are observed with the more stringent criteria applied to TSTs. Furthermore, an association between an IGRA conversion and subsequent disease risk has not been demonstrated. The criteria for interpreting changes in an IGRA that identify new infections remain uncertain. CDC encourages institutions and programs in which IGRAs are used to publish their experiences, particularly in regard to rates of conversion, reversion, and progression to active tuberculosis over time.

### Situations in Which Testing with Both an IGRA and a TST May Be Considered

- Although routine testing with both a TST and an IGRA is not generally recommended, results from both tests might be useful when the initial test (TST or IGRA) is negative in the following situations: 1) when the risk for infection, the risk for progression, and the risk for a poor outcome are increased (e.g., when persons with HIV infection or children aged <5 years are at increased risk for *M. tuberculosis* infection) or 2) when clinical suspicion exists for active tuberculosis (such as in persons with symptoms, signs, and/or radiographic evidence suggestive of active tuberculosis) and confirmation of *M. tuberculosis* infection is desired. In such patients with an initial test that is negative, taking a positive result from a second test as evidence of infection increases detection sensitivity. However, multiple negative results from any combination of these tests cannot exclude *M. tuberculosis* infection.
- Using both a TST and an IGRA also might be useful when the initial test is positive in the following situations: 1) when additional evidence of infection is required to encourage compliance (e.g., in foreign-born health-care workers who believe their positive TST result is attributable to BCG) or 2) in healthy persons who have a low risk for both infection and progression. In the first situation, a positive IGRA might prompt greater acceptance of treatment for LTBI as compared with a positive TST alone. In the latter situation, requiring a positive result from the second test as evidence of infection increases the likelihood that the test result reflects infection. For the second situation, an alternative is to assume, without additional testing, that the initial result is a false positive or that the risk for disease does not warrant additional evaluation or treatment, regardless of test results. Steps should be taken to minimize unnecessary and misleading testing of persons at low risk.
- Repeating an IGRA or performing a TST might be useful when the initial IGRA result is indeterminate, borderline, or invalid and a reason for testing persists. A second test also might be useful when assay measurements from the

initial test are unusual, such as when the Nil value is higher than typical for the population being tested (e.g., IFN- $\gamma$  concentration for Nil by QFT-G or QFT-GIT  $>0.7$  IU/ml for most of the U.S. populations), the Nil value is appreciably greater than the value obtained with *M. tuberculosis* antigen stimulation (e.g. when IFN- $\gamma$  concentration for Nil by QFT-G is 0.35 IU/ml greater than the concentration obtained with either ESAT-6 or CFP-10 stimulation, or when the number of spots for Nil by T-Spot is four spots greater than the number with either ESAT-6 or CFP-10 stimulation), or the Mitogen value is lower than is expected for the population being tested (e.g., the Mitogen Response by QFT-G or QFT-GIT is  $<0.5$  IU/ml, or the number of spots in the mitogen well by T-Spot is  $<20$ ). If an IGRA is to be repeated, a new blood sample should be used. In such situations, repeat testing with another blood sample usually provides interpretable results.

## Medical Management After Testing

- Diagnoses of *M. tuberculosis* infection and decisions about medical or public health management should not be based on IGRA or TST results alone, but should include consideration of epidemiologic and medical history as well as other clinical information.
- Persons with a positive TST or IGRA result should be evaluated for the likelihood of *M. tuberculosis* infection, for risks for progression to active tuberculosis if infected, and for symptoms and signs of active tuberculosis. If risks, symptoms, or signs are present, additional evaluation is indicated to determine if the person has LTBI or active tuberculosis.
- A diagnosis of LTBI requires that active tuberculosis be excluded by medical evaluation, which should include taking a medical history and a physical examination to check for suggestive symptoms and signs, a chest radiograph, and, when indicated, testing of sputum or other clinical samples for the presence of *M. tuberculosis*. Neither an IGRA nor TST can distinguish LTBI from active tuberculosis.
- In persons who have symptoms, signs, or radiographic evidence of active tuberculosis or who are at increased risk for progression to active tuberculosis if infected, a positive result with either an IGRA or TST should be taken as evidence of *M. tuberculosis* infection. However, negative IGRA or TST results are not sufficient to exclude infection in these persons, especially in those at increased risk for a poor outcome if disease develops, and clinical judgment dictates when and if further diagnostic evaluation and treatment are indicated.
- In healthy persons who have a low likelihood both of *M. tuberculosis* infection and of progression to active tuber-

culosis if infected, a single positive IGRA or TST result should not be taken as reliable evidence of *M. tuberculosis* infection. Because of the low probability of infection, a false-positive result is more likely. In such situations, the likelihood of *M. tuberculosis* infection and of disease progression should be reassessed, and the initial test results should be confirmed. Repeat testing, with either the initial test or a different test, may be considered on a case-by-case basis. For such persons, an alternative is to assume, without additional testing, that the initial result is a false positive.

- In persons with discordant test results (i.e., one positive and the other negative), decisions about medical or public health management require individualized judgment in assessing the quality and magnitude of each test result (e.g., size of induration and presence of blistering for a TST; and the TB Response, Nil, and Mitogen values for an IGRA), the probability of infection, the risk for disease if infected, and the risk for a poor outcome if disease occurs.
  - Taking a positive result from either of two tests as evidence of infection is reasonable when 1) clinical suspicion exists for active tuberculosis (e.g., in persons with symptoms, signs, and/or radiographic evidence of active tuberculosis) or 2) the risks for infection, progression, and a poor outcome are increased (e.g., when persons with HIV infection or children aged  $<5$  years are at increased risk for *M. tuberculosis* infection).
  - For healthy persons who have a low risk for both infection and progression, discounting an isolated positive result as a false positive is reasonable. This will increase detection specificity and decrease unnecessary treatment.
  - For persons who have received BCG and who are not at increased risk for a poor outcome if infected (Box 2), TST reactions of  $<15$  mm in size may reasonably be discounted as false positives when an IGRA is clearly negative.
  - In other situations, inadequate evidence exists on which to base recommendations for dealing with discordant results. However, in the absence of convincing evidence of infection, diagnostic decisions may reasonably be deferred unless an increased risk exists for progression if infected and/or a high risk exists for a poor outcome if disease develops.

## Areas for Additional Research

Although substantial progress has been made in documenting the utility of IGRAs, further studies and research are needed. Future studies should focus on determining the value



and limitations of IGRAs in situations of importance to medical care or tuberculosis control. Questions to address include the following (not listed in any order of priority):

- Are IGRAs better at predicting subsequent active tuberculosis than TST?
- Are persons with discordant TST and IGRA results at increased risk for active tuberculosis compared with persons with concordant negative results?
- Are higher IFN- $\gamma$  responses associated with a greater risk for developing active tuberculosis?
- Do IGRAs perform differently in children than in adults, in those with extrapulmonary versus pulmonary tuberculosis, in those with HIV infection versus those without HIV infection, in those recently infected as compared with those infected years earlier, and in those with latent infection as compared with those with active tuberculosis?
- Why do simultaneously performed TST, QFT-GIT, QFT-G, and T-Spot results differ?
- Can sensitivity and specificity of IGRAs be improved by modification in testing methods, application of different interpretation criteria, or inclusion of additional antigens?
- What is the best approach for determining cut points for IGRA interpretation, including situations where Nil values are high or Mitogen values are low?
- To what extent does inclusion of a “borderline” interpretation improve IGRA accuracy?
- What causes variation in IGRA results and to what extent?
- What magnitude of change in IFN- $\gamma$  response indicates new infection?
- After exposure, how long does it take for an IGRA to become positive?
- What is the clinical significance of IGRA reversion?
- What methods should be used to monitor IGRA quality?
- Is there an association between lymphocyte count and IFN- $\gamma$  response (with or without HIV infection)?
- What effect does treatment of *M. tuberculosis* infection have on IGRA results?
- How do host and bacterial genetic factors affect IGRA results?

## References

1. American Thoracic Society, CDC. Diagnostic standards and classification of tuberculosis in adults and children. *Am J Respir Crit Care Med* 2000;161:1376–95.
2. Rothel JS, Jones SL, Corner LA, Cox JC, Wood PR. A sandwich enzyme immunoassay for bovine interferon-gamma and its use for the detection of tuberculosis in cattle. *Aust Vet J* 1990;67:134–7.
3. Converse PJ, Jones SL, Astemborski J, Vlahov D, Graham NM. Comparison of a tuberculin interferon-gamma assay with the tuberculin skin test in high-risk adults: effect of human immunodeficiency virus infection. *J Infect Dis* 1997;176:144–150.
4. Streeton JA, Desem N, Jones SL. Sensitivity and specificity of a gamma interferon blood test for tuberculosis infection. *Int J Tuberc Lung Dis* 1998;2:443–50.
5. Food and Drug Administration. QuantiFERON-TB - P010033. Available at <http://www.fda.gov/MedicalDevices/ProductsandMedicalProcedures/DeviceApprovalsandClearances/Recently-ApprovedDevices/ucm084025.htm>. Accessed June 16, 2010.
6. Mazurek GH, Villarino ME. Guidelines for using the QuantiFERON-TB test for diagnosing latent *Mycobacterium tuberculosis* infection. *MMWR* 2003;52(No. RR-2):15–8.
7. Food and Drug Administration. QuantiFERON-TB Gold - P010033/S006. Available at <http://www.fda.gov/MedicalDevices/ProductsandMedicalProcedures/DeviceApprovalsandClearances/PMAApprovals/ucm110838.htm>. Accessed June 16, 2010.
8. CDC. Guidelines for using the QuantiFERON-TB Gold test for detecting *Mycobacterium tuberculosis* infection, United States. *MMWR* 2005;54(No. RR-15):49–55.
9. World Health Organization. Global tuberculosis control: epidemiology, strategy, financing: WHO report 2009. Geneva, Switzerland: World Health Organization; 2009. Available at [http://www.who.int/tb/publications/global\\_report/2009/pdf/report\\_without\\_annexes.pdf](http://www.who.int/tb/publications/global_report/2009/pdf/report_without_annexes.pdf). Accessed June 16, 2010.
10. CDC. Reported tuberculosis in the United States, 2008. Atlanta, GA: U.S. Department of Health and Human Services, CDC; 2009.
11. Bennett DE, Courval JM, Onorato I, et al. Prevalence of tuberculosis infection in the United States population: the national health and nutrition examination survey, 1999–2000. *Am J Respir Crit Care Med* 2008;177:348–55.
12. CDC. Targeted tuberculin testing and treatment of latent tuberculosis infection. *MMWR* 2000;49(No. RR-6).
13. Edwards PQ, Edwards LB. Story of the tuberculin skin test from an epidemiologic viewpoint. *Am Rev Respir Dis* 1960;81:1–47.
14. Antonucci G, Girardi E, Raviglione MC, Ippolito G. Risk factors for tuberculosis in HIV-infected persons. *JAMA* 1995;274:143–8.
15. Selwyn PA, Hartel D, Lewis VA, et al. A prospective study of the risk of tuberculosis among intravenous drug users with human immunodeficiency virus infection. *N Engl J Med* 1989;320:545–50.
16. Horsburgh CR, Jr. Priorities for the treatment of latent tuberculosis infection in the United States. *N Engl J Med* 2004;350:2060–7.
17. Judson FN, Feldman RA. Mycobacterial skin tests in humans 12 years after infection with *Mycobacterium marinum*. *Am Rev Respir Dis* 1974;109:544–7.
18. Snider DE Jr. Bacille Calmette-Guerin vaccinations and tuberculin skin tests. *JAMA* 1985;253:3438–39.
19. Mazurek GH, Weis SE, Moonan PK, et al. Prospective comparison of tuberculin skin test and two whole blood interferon-gamma release assays in tuberculosis suspects. *Clin Infect Dis* 2007;45:837–45.
20. Andersen P, Munk ME, Pollock JM, Doherty TM. Specific immune-based diagnosis of tuberculosis. *Lancet* 2000;356:1099–104.
21. Cellestis Limited. QuantiFERON-TB Gold [Package insert]. Available at <http://www.cellestis.com/IRM/Company/ShowPage.aspx?CPID=1247>. Accessed June 16, 2010.
22. Food and Drug Administration. QuantiFERON-TB Gold In-Tube - P010033/S011. Available at <http://www.fda.gov/MedicalDevices/ProductsandMedicalProcedures/DeviceApprovalsandClearances/PMAApprovals/ucm106548.htm>. Accessed June 16, 2010.
23. Food and Drug Administration. T-SPOT-TB - P070006. Available at <http://www.fda.gov/MedicalDevices/ProductsandMedicalProcedures/DeviceApprovalsandClearances/PMAApprovals/ucm102794.htm>. Accessed June 16, 2010.
24. Oxford Immunotec Limited. T-Spot.TB [U.S. package insert]. Available at <http://www.oxfordimmunotec.com/USPageInsert>. Accessed June 16, 2010.
25. Oxford Immunotec Limited. T-Spot.TB [U.K. package insert]. Available at <http://www.oxfordimmunotec.com/96-UK>. Accessed June 16, 2010.
26. Harada N, Higuchi K, Yoshiyama T, et al. Comparison of the sensitivity and specificity of two whole blood interferon-gamma assays for *M. tuberculosis* infection. *J Infect* 2008;56:348–53.



27. Tsiouris SJ, Coetzee D, Toro PL, et al. Sensitivity analysis and potential uses of a novel gamma interferon release assay for diagnosis of tuberculosis. *J Clin Microbiol* 2006;44:2844–50.
28. Detjen AK, Keil T, Roll S, et al. Interferon-gamma release assays improve the diagnosis of tuberculosis and nontuberculous mycobacterial disease in children in a country with a low incidence of tuberculosis. *Clin Infect Dis* 2007;45:322–8.
29. Pai M, Joshi R, Bandyopadhyay M, et al. Sensitivity of a whole-blood interferon-gamma assay among patients with pulmonary tuberculosis and variations in T-cell responses during anti-tuberculosis treatment. *Infection* 2007;35:98–103.
30. Adetifa IM, Lugos MD, Hammond A, et al. Comparison of two interferon gamma release assays in the diagnosis of *Mycobacterium tuberculosis* infection and disease in The Gambia. *BMC Infect Dis* 2007;7:122.
31. Dominguez J, Ruiz-Manzano J, De Souza-Galvao M, et al. Comparison of two commercially available gamma interferon blood tests for immunodiagnosis of tuberculosis. *Clin Vaccine Immunol* 2008;15:168–71.
32. Palazzo R, Spensieri F, Massari M, et al. Use of whole-blood samples in in-house bulk and single-cell antigen-specific gamma interferon assays for surveillance of *Mycobacterium tuberculosis* infections. *Clin Vaccine Immunol* 2008;15:327–37.
33. Chee CB, Gan SH, KhinMar KW, et al. Comparison of sensitivities of two commercial gamma interferon release assays for pulmonary tuberculosis. *J Clin Microbiol* 2008;46:1935–40.
34. Ruhwald M, Bodmer T, Maier C, et al. Evaluating the potential of IP-10 and MCP-2 as biomarkers for the diagnosis of tuberculosis. *Eur Respir J* 2008;32:1607–15.
35. Bartu V, Havelkova M, Kopecka E. QuantiFERON-TB Gold in the diagnosis of active tuberculosis. *J Int Med Res* 2008;36:434–37.
36. Raby E, Moyo M, Devendra A, et al. The effects of HIV on the sensitivity of a whole blood IFN-gamma release assay in Zambian adults with active tuberculosis. *PLoS ONE* 2008;3:e2489.
37. Aichelburg MC, Rieger A, Breitenecker F, et al. Detection and prediction of active tuberculosis disease by a whole-blood interferon-gamma release assay in HIV-1-infected individuals. *Clin Infect Dis* 2009;48:954–62.
38. Goletti D, Stefania C, Butera O, et al. Accuracy of immunodiagnostic tests for active tuberculosis using single and combined results: a multicenter TBNET-Study. *PLoS ONE* 2008;3:e3417.
39. Kampmann B, Whittaker E, Williams A, et al. Interferon-gamma release assays do not identify more children with active tuberculosis than the tuberculin skin test. *Eur Respir J* 2009;33:1374–82.
40. Lee JY, Choi HJ, Park IN, et al. Comparison of two commercial interferon-gamma assays for diagnosing *Mycobacterium tuberculosis* infection. *Eur Respir J* 2006;28:24–30.
41. Meier T, Eulenbruch HP, Wrighton-Smith P, Enders G, Regnath T. Sensitivity of a new commercial enzyme-linked immunospot assay (T SPOT-TB) for diagnosis of tuberculosis in clinical practice. *Eur J Clin Microbiol Infect Dis* 2005;24:529–36.
42. Goletti D, Carrara S, Vincenti D, et al. Accuracy of an immune diagnostic assay based on RD1 selected epitopes for active tuberculosis in a clinical setting: a pilot study. *Clin Microbiol Infect* 2006;12:544–50.
43. Ferrara G, Losi M, D'Amico R, et al. Use in routine clinical practice of two commercial blood tests for diagnosis of infection with *Mycobacterium tuberculosis*: a prospective study. *Lancet* 2006;367:1328–34.
44. Jafari C, Ernst M, Kalsdorf B, et al. Rapid diagnosis of smear-negative tuberculosis by bronchoalveolar lavage enzyme-linked immunospot. *Am J Respir Crit Care Med* 2006;174:1048–54.
45. Kang YA, Lee HW, Hwang SS, et al. Usefulness of whole-blood interferon-gamma assay and interferon-gamma enzyme-linked immunospot assay in the diagnosis of active pulmonary tuberculosis. *Chest* 2007;132:959–65.
46. Bosshard V, Roux-Lombard P, Perneger T, et al. Do results of the T-SPOT.TB interferon-gamma release assay change after treatment of tuberculosis? *Respir Med* 2009;103:30–4.
47. Wang JY, Chou CH, Lee LN, et al. Diagnosis of tuberculosis by an enzyme-linked immunospot assay for interferon-gamma. *Emerg Infect Dis* 2007;13:553–8.
48. Janssens JP, Roux-Lombard P, Perneger T, et al. Quantitative scoring of an interferon-gamma assay for differentiating active from latent tuberculosis. *Eur Respir J* 2007;30:722–8.
49. Ozekinci T, Ozbek E, Celik Y. Comparison of tuberculin skin test and a specific T-cell-based test, T-Spot.TB, for the diagnosis of latent tuberculosis infection. *J Int Med Res* 2007;35:696–703.
50. Soysal A, Torun T, Efe S, et al. Evaluation of cut-off values of interferon-gamma-based assays in the diagnosis of *M. tuberculosis* infection. *Int J Tuberc Lung Dis* 2008;12:50–6.
51. Liao CH, Chou CH, Lai CC, et al. Diagnostic performance of an enzyme-linked immunospot assay for interferon-gamma in extrapulmonary tuberculosis varies between different sites of disease. *J Infect* 2009;59:402–8.
52. Higuchi K, Kawabe Y, Mitarai S, Yoshiyama T, Harada N, Mori T. Comparison of performance in two diagnostic methods for tuberculosis infection. *Med Microbiol Immunol* 2009;198:33–7.
53. Adams LV, Waddell RD, von Reyn CF. T-SPOT.TB Test(R) results in adults with *Mycobacterium avium* complex pulmonary disease. *Scand J Infect Dis* 2008;40:196–203.
54. Pai M, Zwerling A, Menzies D. Systematic review: T-cell-based assays for the diagnosis of latent tuberculosis infection: an update. *Ann Intern Med* 2008;149:177–84.
55. Farhat M, Greenaway C, Pai M, Menzies D. False-positive tuberculin skin tests: what is the absolute effect of BCG and non-tuberculous mycobacteria? *Int J Tuberc Lung Dis* 2006;10:1192–204.
56. Wang L, Turner MO, Elwood RK, Schulzer M, FitzGerald JM. A meta-analysis of the effect of Bacille Calmette-Guerin vaccination on tuberculin skin test measurements. *Thorax* 2002;57:804–9.
57. Brock I, Weldingh K, Lillebaek T, Follmann F, Andersen P. Comparison of tuberculin skin test and new specific blood test in tuberculosis contacts. *Am J Respir Crit Care Med* 2004;170:65–9.
58. Arend SM, Thijssen SF, Leyten EM, et al. Comparison of two interferon-gamma assays and tuberculin skin test for tracing tuberculosis contacts. *Am J Respir Crit Care Med* 2007;175:618–27.
59. Choi JC, Shin JW, Kim JY, et al. The effect of previous tuberculin skin test on the follow-up examination of whole-blood interferon-gamma assay in the screening for latent tuberculosis infection. *Chest* 2008;133:1415–20.
60. Leung CC, Yam WC, Yew WW, et al. Comparison of T-Spot.TB and tuberculin skin test among silicotic patients. *Eur Respir J* 2008;31:266–72.
61. Diel R, Loddenkemper R, Meywald-Walter K, Gottschalk R, Nienhaus A. Comparative performance of tuberculin skin test, QuantiFERON-TB-Gold In Tube assay, and T-Spot.TB test in contact investigations for tuberculosis. *Chest* 2009;135:1010–8.
62. Hill PC, Jeffries DJ, Brookes RH, et al. Using ELISPOT to expose false positive skin test conversion in tuberculosis contacts. *PLoS ONE* 2007;2:e183.
63. Igari H, Watanabe A, Sato T. Booster phenomenon of QuantiFERON-TB Gold after prior intradermal PPD injection. *Int J Tuberc Lung Dis* 2007;11:788–91.
64. Leyten EM, Prins C, Bossink AW, et al. Effect of tuberculin skin testing on a *Mycobacterium tuberculosis*-specific interferon-gamma assay. *Eur Respir J* 2007;29:1212–6.
65. Naseer A, Naqi S, Kampmann B. Evidence for boosting *Mycobacterium tuberculosis*-specific IFN-gamma responses at 6 weeks following tuberculin skin testing. *Eur Respir J* 2007;29:1282–83.
66. Richeldi L, Ewer K, Losi M, et al. Repeated tuberculin testing does not induce false positive ELISPOT results. *Thorax* 2006;61:180.
67. Hill PC, Brookes RH, Fox A, et al. Longitudinal assessment of an ELISPOT test for *Mycobacterium tuberculosis* infection. *PLoS Med* 2007;4:e192.

68. Ewer K, Millington KA, Deeks JJ, et al. Dynamic antigen-specific T-cell responses after point-source exposure to *Mycobacterium tuberculosis*. *Am J Respir Crit Care Med* 2006;174:831–9.
69. Pai M, Joshi R, Dogra S, et al. Serial testing of health care workers for tuberculosis using interferon-gamma assay. *Am J Respir Crit Care Med* 2006;174:349–55.
70. Perry S, Sanchez L, Yang S, et al. Reproducibility of QuantiFERON-TB gold in-tube assay. *Clin Vaccine Immunol* 2008;15:425–32.
71. Veerapathran A, Joshi R, Goswami K, et al. T-cell assays for tuberculosis infection: deriving cut-offs for conversions using reproducibility data. *PLoS ONE* 2008;3:e1850.
72. Tsiouris SJ, Austin J, Toro P, et al. Results of a tuberculosis-specific IFN-gamma assay in children at high risk for tuberculosis infection. *Int J Tuberc Lung Dis* 2006;10:939–41.
73. Nakaoka H, Lawson L, Squire SB, et al. Risk for tuberculosis among children. *Emerg Infect Dis* 2006;12:1383–88.
74. Janssens J, Roux-Lombard P, Perneger T, et al. Contribution of a IFN-gamma assay in contact tracing for tuberculosis in a low-incidence, high immigration area. *Swiss Med Wkly* 2008;138:585–93.
75. Vynnycky E, Fine PE. Lifetime risks, incubation period, and serial interval of tuberculosis. *Am J Epidemiol* 2000;152:247–63.
76. Diel R, Loddenkemper R, Meywald-Walter K, Niemann S, Nienhaus A. Predictive value of a whole blood IFN-gamma assay for the development of active tuberculosis disease after recent infection with *Mycobacterium tuberculosis*. *Am J Respir Crit Care Med* 2008;177:1164–70.
77. Stout JE, Menzies D. Predicting tuberculosis: does the IGRA tell the tale? *Am J Respir Crit Care Med* 2008;177:1055–7.
78. Kik SV, Franken WP, Mensen M, et al. Predictive value for progression to tuberculosis by IGRA and TST in immigrant contacts. *Eur Respir J* 2009. Available at <http://erj.ersjournals.com/cgi/rapidpdf/09031936.00098509v1>. Accessed June 16, 2010.
79. Hill PC, Jackson-Sillah DJ, Fox A, et al. Incidence of tuberculosis and the predictive value of ELISPOT and Mantoux tests in Gambian case contacts. *PLoS ONE* 2008;3:e1379.
80. Newton SM, Brent AJ, Anderson S, Whittaker E, Kampmann B. Paediatric tuberculosis. *Lancet Infect Dis* 2008;8:498–510.
81. Kampmann B, Tena-Coki G, Anderson S. Blood tests for diagnosis of tuberculosis. *Lancet* 2006;368:282–3.
82. Connell TG, Curtis N, Ranganathan SC, Buttery JP. Performance of a whole blood interferon gamma assay for detecting latent infection with *Mycobacterium tuberculosis* in children. *Thorax* 2006;61:616–20.
83. Lewinsohn DA, Zalwango S, Stein CM, et al. Whole blood interferon-gamma responses to *Mycobacterium tuberculosis* antigens in young household contacts of persons with tuberculosis in Uganda. *PLoS ONE* 2008;3:e3407.
84. Bergamini BM, Losi M, Vaienti F, et al. Performance of commercial blood tests for the diagnosis of latent tuberculosis infection in children and adolescents. *Pediatrics* 2009;123:e419–24.
85. Connell TG, Ritz N, Paxton GA, et al. A three-way comparison of tuberculin skin testing, QuantiFERON-TB gold and T-SPOT.TB in children. *PLoS ONE* 2008;3:e2624.
86. Dogra S, Narang P, Mendiratta DK, et al. Comparison of a whole blood interferon-gamma assay with tuberculin skin testing for the detection of tuberculosis infection in hospitalized children in rural India. *J Infect* 2007;54:267–76.
87. Lighter J, Rigaud M, Eduardo R, Peng CH, Pollack H. Latent tuberculosis diagnosis in children by using the QuantiFERON-TB Gold In-Tube test. *Pediatrics* 2009;123:30–7.
88. Nicol MP, Davies MA, Wood K, et al. Comparison of T-SPOT.TB assay and tuberculin skin test for the evaluation of young children at high risk for tuberculosis in a community setting. *Pediatrics* 2009;123:38–43.
89. Warier A, Gunawathi S, Sankarapandian V, John KR, Bose A. T-cell assay as a diagnostic tool for tuberculosis. *Indian Pediatr* 2010;47:90–2.
90. Taylor RE, Cant AJ, Clark JE. Potential effect of NICE tuberculosis guidelines on paediatric tuberculosis screening. *Arch Dis Child* 2008;93:200–3.
91. Shingadia D, Novelli V. The tuberculin skin test: a hundred, not out? *Arch Dis Child* 2008;93:189–90.
92. Brock I, Ruhwald M, Lundgren B, et al. Latent tuberculosis in HIV positive, diagnosed by the *M. tuberculosis* specific interferon gamma test. *Respir Res* 2006;7:56.
93. Balcells ME, Perez CM, Chanqueo L, et al. A comparative study of two different methods for the detection of latent tuberculosis in HIV-positive individuals in Chile. *Int J Infect Dis* 2008;12:645–52.
94. Jones S, de Gijzel D, Wallach FR, et al. Utility of QuantiFERON-TB Gold in-tube testing for latent TB infection in HIV-infected individuals. *Int J Tuberc Lung Dis* 2007;11:1190–5.
95. Luetkemeyer AF, Charlebois ED, Flores LL, et al. Comparison of an interferon-gamma release assay with tuberculin skin testing in HIV-infected individuals. *Am J Respir Crit Care Med* 2007;175:737–42.
96. Talati NJ, Seybold U, Humphrey B, Aina A, Tapia J, et al. Poor concordance between interferon-gamma release assays and tuberculin skin tests in diagnosis of latent tuberculosis infection among HIV-infected individuals. *BMC Infect Dis* 2009;9:15.
97. Bocchino M, Matarese A, Bellofiore B, et al. Performance of two commercial blood IFN-gamma release assays for the detection of *Mycobacterium tuberculosis* infection in patient candidates for anti-TNF-alpha treatment. *Eur J Clin Microbiol Infect Dis* 2008;27:907–13.
98. Cobanoglu N, Ozcelik U, Kalyoncu U, et al. Interferon-gamma assays for the diagnosis of tuberculosis infection before using tumour necrosis factor-alpha blockers. *Int J Tuberc Lung Dis* 2007;11:1177–82.
99. Matulis G, Juni P, Villiger PM, Gadola SD. Detection of latent tuberculosis in immunosuppressed patients with autoimmune diseases: performance of a *Mycobacterium tuberculosis* antigen-specific interferon gamma assay. *Ann Rheum Dis* 2008;67:84–90.
100. Ponce de LD, Acevedo-Vasquez E, Alvizuri S, et al. Comparison of an interferon-gamma assay with tuberculin skin testing for detection of tuberculosis (TB) infection in patients with rheumatoid arthritis in a TB-endemic population. *J Rheumatol* 2008;35:776–81.
101. Mandalakas AM, Hesselink AC, Chegou NN, et al. High level of discordant IGRA results in HIV-infected adults and children. *Int J Tuberc Lung Dis* 2008;12:417–23.
102. Rangaka MX, Wilkinson KA, Seldon R, et al. Effect of HIV-1 infection on T-cell-based and skin test detection of tuberculosis infection. *Am J Respir Crit Care Med* 2007;175:514–20.
103. Stephan C, Wolf T, Goetsch U, et al. Comparing QuantiFERON-tuberculosis gold, T-SPOT tuberculosis and tuberculin skin test in HIV-infected individuals from a low prevalence tuberculosis country. *AIDS* 2008;22:2471–9.
104. Lindemann M, Dioury Y, Beckebaum S, et al. Diagnosis of tuberculosis infection in patients awaiting liver transplantation. *Hum Immunol* 2009;70:24–8.
105. Passalent L, Khan K, Richardson R, et al. Detecting latent tuberculosis infection in hemodialysis patients: a head-to-head comparison of the T-SPOT.TB test, tuberculin skin test, and an expert physician panel. *Clin J Am Soc Nephrol* 2007;2:68–73.
106. Piana F, Codecasa LR, Cavallerio P, et al. Use of a T-cell-based test for detection of tuberculosis infection among immunocompromised patients. *Eur Respir J* 2006;28:31–4.
107. Porsa E, Cheng L, Graviss EA. Comparison of an ESAT-6/CFP-10 peptide-based enzyme-linked immunospot assay to a tuberculin skin test for screening of a population at moderate risk of contracting tuberculosis. *Clin Vaccine Immunol* 2007;14:714–9.

108. Vassilopoulos D, Stamoulis N, Hadziyannis E, Archimandritis AJ. Usefulness of enzyme-linked immunospot assay (elispot) compared to tuberculin skin testing for latent tuberculosis screening in rheumatic patients scheduled for anti-tumor necrosis factor treatment. *J Rheumatol* 2008;35:1271–6.
109. MAG Mutual Healthcare Solutions, Inc. Physicians' fee and coding guide 2009. Atlanta, Georgia: MAG Mutual Healthcare Solutions, Inc.; 2009.
110. Marra F, Marra CA, Sadatsafavi M, et al. Cost-effectiveness of a new interferon-based blood assay, QuantiFERON(R)-TB Gold, in screening tuberculosis contacts. *Int J Tuberc Lung Dis* 2008;12:1414–24.
111. Sahni R, Miranda C, Yen-Lieberman B, et al. Does the implementation of an interferon-gamma release assay in lieu of a tuberculin skin test increase acceptance of preventive therapy for latent tuberculosis among healthcare workers? *Infect Control Hosp Epidemiol* 2009;30:197–8.
112. American Academy of Pediatrics. Tuberculosis. In: Pickering LK, Baker CJ, Kimberlin DW, Long SS, eds. Red book: 2009 report of the Committee on Infectious Disease. 28th ed. Elk Grove Village, IL: American Academy of Pediatrics, 2009:680–701.
113. CDC. Guidelines for preventing the transmission of *Mycobacterium tuberculosis* in health-care settings, 2005. *MMWR* 2005;54(No. RR-17).

**TABLE 1. Interpretation criteria for the QuantiFERON-TB Gold Test (QFT-G)**

Interpretation	Nil*	TB Response†	Mitogen Response§
Positive¶	Any	≥0.35 IU/ml and ≥50% of Nil	Any
Negative**	≤0.7	<0.35 IU/ml	≥0.5
Indeterminate††	≤0.7	<0.35 IU/ml	<0.5
	>0.7	<50% of Nil	Any

**Source:** Based on Cellestis Limited. QuantiFERON-TB Gold [Package insert]. Available at <http://www.cellestis.com/IRM/Company/ShowPage.aspx?CPID=1247>.

\* The interferon gamma (IFN-γ) concentration in plasma from blood incubated with saline.

† The higher IFN-γ concentration in plasma from blood stimulated with a cocktail of peptides representing early secretory antigenic target-6 (ESAT-6) or a cocktail of peptides representing culture filtrate protein 10 (CFP-10) minus Nil.

§ The IFN-γ concentration in plasma from blood stimulated with mitogen minus Nil.

¶ Interpretation indicating that *Mycobacterium tuberculosis* infection is likely.

\*\* Interpretation indicating that *M. tuberculosis* infection is not likely.

†† Interpretation indicating an uncertain likelihood of *M. tuberculosis* infection.

**TABLE 2. Interpretation criteria for the QuantiFERON-TB Gold In-Tube Test (QFT-GIT)**

Interpretation	Nil*	TB Response†	Mitogen Response§
Positive¶	≤8.0	≥0.35 IU/ml and ≥25% of Nil	Any
Negative**	≤8.0	<0.35 IU/ml or <25% of Nil	≥0.5
Indeterminate††	≤8.0	<0.35 IU/ml or <25% of Nil	<0.5
	>8.0	Any	Any

**Source:** Based on Cellestis Limited. QuantiFERON-TB Gold In-Tube [Package insert]. Available at <http://www.cellestis.com/IRM/content/pdf/QuantiFeron%20US%20VerG-Jan2010%20NO%20TRIMS.pdf>.

\* The interferon gamma (IFN-γ) concentration in plasma from blood incubated without antigen.

† The IFN-γ concentration in plasma from blood stimulated with a single cocktail of peptides representing early secretory antigenic target-6 (ESAT-6), culture filtrate protein-10 (CFP-10), and part of TB 7.7 minus Nil.

§ The IFN-γ concentration in plasma from blood stimulated with mitogen minus Nil.

¶ Interpretation indicating that *Mycobacterium tuberculosis* infection is likely.

\*\* Interpretation indicating that *M. tuberculosis* infection is not likely.

†† Interpretation indicating an uncertain likelihood of *M. tuberculosis* infection.

**TABLE 3. Interpretation criteria for the T-SPOT.TB Test (T-Spot)**

Interpretation	Nil*	TB Response†	Mitogen§
Positive¶	≤10 spots	≥8 spots	Any
Borderline**	≤10 spots	5, 6, or 7 spots	Any
Negative††	≤10 spots	≤4 spots	
Indeterminate**	>10 spots	Any	Any
	≤10 spots	<5 spots	<20 spots

**Source:** Based on Oxford Immunotec Limited. T-Spot.TB [Package insert]. Available at <http://www.oxfordimmunotec.com/USpageInsert>.

\* The number of spots resulting from incubation of PBMCs in culture media without antigens.

† The greater number of spots resulting from stimulation of peripheral blood mononuclear cells (PBMCs) with two separate cocktails of peptides representing early secretory antigenic target-6 (ESAT-6) or culture filtrate protein-10 (CFP-10) minus Nil.

§ The number of spots resulting from stimulation of PBMCs with mitogen without adjustment for the number of spots resulting from incubation of PBMCs without antigens.

¶ Interpretation indicating that *Mycobacterium tuberculosis* infection is likely.

\*\* Interpretation indicating an uncertain likelihood of *M. tuberculosis* infection.

†† Interpretation indicating that *M. tuberculosis* infection is not likely.

**TABLE 4. QuantiFERON-TB Gold-In-Tube test (QFT-GIT) sensitivity,\* by country in which study was conducted — 14 countries, 2006–2009**

Country	Subjects	Confirmed TB <sup>†</sup>				Inter-pretation criteria**	QFT-GIT results				TST <sup>¶</sup> results				% TST+ vs. QFT-GIT+ p-value <sup>††</sup>
		No. confirmed/ No. with TB diagnosis	(%)	HIV <sup>§</sup> -positive			Positive	(%)	Indeterminate		Cutoff	Positive			
				No. +/ No. tested	(%)				No. +/ No. tested	(%)		No. +/ No. tested	(%)		
South Africa <sup>§§</sup>	Children	154/154	(100)	26/41	(63)	A	100/131	(76)	23/154	(15)	Stratified	131/146	(90)	<0.01	
Germany <sup>¶¶¶</sup>	Children	28/28	(100)	NR <sup>***</sup>	NR	B	26/28	(93)	ND <sup>†††</sup>	ND	5 mm	28/28	(100)	0.49	
India <sup>§§§</sup>	Adults	58/60	(97)	3/60	(5)	A	44/60	(73)	0/60	(0)	ND	ND	ND	ND	
The Gambia <sup>¶¶¶¶</sup>	Adults	75/75	(100)	7/77	(9)	B	48/75	(64)	ND	ND	ND	ND	ND	ND	
Spain <sup>****</sup>	Adults & children	NR/42	(NR)	NR	NR	C	33/42	(79)	0/42	(0)	5 mm	40/42	(95)	0.05	
Italy <sup>††††</sup>	Mostly adults	17/17	(100)	NR	NR	C	14/17	(82)	0/17	(0)	NR	9/12	(75)	0.97	
Singapore <sup>§§§§</sup>	Adults	286/286	(100)	7/238	(3)	A	224/270	(83)	10/286	(4)	10 mm	206/217	(95)	<0.01	
											15 mm	158/217	(73)	ND	
Japan <sup>¶¶¶¶¶</sup>	Adults	100/100	(100)	1/100	(1)	D	87/94	(93)	6/100	(6)	ND	ND	ND	ND	
Denmark <sup>*****</sup>	Adults	68/80	(85)	10/56	(18)	C	65/76	(86)	4/80	(5)	10 mm	9/12 (75)	(75)	0.58	
Czech	Adults	22/22	(100)	0/22	(0)	C	19/22	(86)	0/22	(0)	5	12/22	(55)	0.05	
Republic <sup>†††††</sup>		0/31	(0)	0/31	(0)		24/28	(86)	3/31	(6)		22/31	(71)	ND	
Zambia <sup>§§§§§</sup>	Adults	0/112	(0)	59/96	(62)	A	83/96	(86)	16/112	(14)	5 mm <sup>¶¶¶¶¶</sup>	62/92 <sup>¶¶¶¶¶</sup>	(67) <sup>¶¶¶¶¶</sup>	<0.01	
											10 mm <sup>¶¶¶¶¶</sup>	48/92 <sup>¶¶¶¶¶</sup>	(52) <sup>¶¶¶¶¶</sup>	ND	
Austria <sup>*****</sup>	HIV+ adults	10/11	(91)	11/11	(100)	D	10/11	(91)	0/11	(0)	5 mm	8/10	(80)	0.92	
Multiple	Adults	121/121	(100)	3/NR	(NR)	C	99/117	(85)	4/121	(3)	10 or 15	114/136	(84)	1.0	
European <sup>†††††</sup>		0/34	(0)	0/NR	(NR)		22/34	(65)	0/34	(0)		37/41	(90)	0.02	
United Kingdom <sup>§§§§§</sup>	Children	25/25	(100)	0/35	(0)	D	20/23	(87)	2/25	(8)	10 mm	21/24	(86)	1.0	
		0/38	(0)				20/36	(56)	2/38	(5)		24/38	(63)	0.67	

\* Source: Modified from Pai M, Zwerling A, Menzies D. Systematic review: T-cell-based assays for the diagnosis of latent tuberculosis infection: an update. *Ann Intern Med* 2008;149:177–84 supplemented with additional information and compared with TST sensitivity when available.

† Tuberculosis disease was confirmed by culture and/or nucleic acid amplification test.

§ Human immunodeficiency virus.

¶ Tuberculin skin test.

\*\* "A" = QFT-GIT was interpreted as positive if Tuberculosis (TB) Response was  $\geq 0.35$  IU/mL; indeterminate if TB Response was  $< 0.35$  IU/mL and Mitogen Response was  $< 0.5$  IU/mL; and negative if TB Response was  $< 0.35$  IU/mL and Mitogen Response was  $\geq 0.5$  IU/mL. "B" = QFT-GIT was interpreted as positive if TB Response was  $\geq 0.35$  IU/mL; Mitogen Response was not measured. "C" = QFT-GIT interpretation criteria were not stated explicitly. "D" = QFT-GIT was interpreted as positive if TB Response was  $\geq 0.35$  IU/mL and Nil was  $\leq 8.0$  IU/mL; indeterminate if Nil  $\geq 8.0$  IU/mL or TB Response was  $< 0.35$  IU/mL and Mitogen Response was  $< 0.5$  IU/mL; and negative if TB Response was  $< 0.35$  IU/mL, Mitogen Response was  $\geq 0.5$  IU/mL, and Nil was  $\leq 8.0$  IU/mL.

†† Fisher's exact test was used by CDC to calculate 2-tailed p-values.

§§ Source: Tsiouris SJ, Coetzee D, Toro PL, Austin J, Stein Z, el-Sadr W. Sensitivity analysis and potential uses of a novel gamma interferon release assay for diagnosis of tuberculosis. *J Clin Microbiol* 2006;44:2844–50.

¶¶ Source: Detjen AK, Keil T, Roll S, et al. Interferon-gamma release assays improve the diagnosis of tuberculosis and nontuberculous mycobacterial disease in children in a country with a low incidence of tuberculosis. *Clin Infect Dis* 2007;45:322–8.

\*\*\* Not reported.

††† Not done.

§§§ Source: Pai M, Joshi R, Bandyopadhyay M, et al. Sensitivity of a whole-blood interferon-gamma assay among patients with pulmonary tuberculosis and variations in T-cell responses during anti-tuberculosis treatment. *Infection* 2007;35:98–103.

¶¶¶ Source: Adetifa IM, Lugos MD, Hammond A et al. Comparison of two interferon gamma release assays in the diagnosis of *Mycobacterium tuberculosis* infection and disease in The Gambia. *BMC Infect Dis* 2007;7:122.

\*\*\*\* Source: Dominguez J, Ruiz-Manzano J, De Souza-Galvao M, et al. Comparison of two commercially available gamma interferon blood tests for immunodiagnosis of tuberculosis. *Clin Vaccine Immunol* 2008;15:168–171.

†††† Source: Palazzo R, Spensieri F, Massari M, et al. Use of whole-blood samples in in-house bulk and single-cell antigen-specific gamma interferon assays for surveillance of *Mycobacterium tuberculosis* infections. *Clin Vaccine Immunol* 2008;15:327–37.

§§§§ Source: Chee CB, Gan SH, KhinMar KW, et al. Comparison of sensitivities of two commercial gamma interferon release assays for pulmonary tuberculosis. *J Clin Microbiol* 2008;46:1935–40.

¶¶¶¶ Source: Harada N, Higuchi K, Yoshiyama T, et al. Comparison of the sensitivity and specificity of two whole blood interferon-gamma assays for *M. tuberculosis* infection. *J Infect* 2008;56:348–53.

\*\*\*\*\* Source: Ruhwald M, Bodmer T, Maier C, et al. Evaluating the potential of IP-10 and MCP-2 as biomarkers for the diagnosis of tuberculosis. *Eur Respir J* 2008; 32(6):1607-1615.

††††† Source: Bartu V, Havelkova M, Kopecka E. QuantiFERON-TB Gold in the diagnosis of active tuberculosis. *J Int Med Res* 2008;36:434–7.

§§§§§ Source: Raby E, Moyo M, Devendra A, et al. The effects of HIV on the sensitivity of a whole blood IFN-gamma release assay in Zambian adults with active tuberculosis. *PLoS ONE* 2008;3:e2489. [E-pub]. Available at <http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0002489>.

¶¶¶¶¶ TST read 48–164 hours after tuberculin injection.

\*\*\*\*\* Source: Aichelburg MC, Rieger A, Breitenacker F, et al. Detection and prediction of active tuberculosis disease by a whole-blood interferon-gamma release assay in HIV-1 infected individuals. *Clin Infect Dis* 2009;48:954–62.

††††† Source: Goletti D, Stefania C, Butera O, et al. Accuracy of immunodiagnostic tests for active tuberculosis using single and combined results: a multicenter TBNET-Study. *PLoS ONE* 2008; 3:e3417. [E-published]. Available at <http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0003417>.

§§§§§ Source: Kampmann B, Whittaker E, Williams A, et al. Interferon-gamma release assays do not identify more children with active tuberculosis than the tuberculin skin test. *Eur Respir J* 2009;33:1374–82.



TABLE 5. T-SPOT.TB test (T-Spot) sensitivity,\* by country in which study was conducted —12 countries, 2005–2009

Country	Subjects	Confirmed TB†				Inter-pretation criteria**	T-Spot results				TST¶ results			% TST+ vs. QFT-GIT+ p-value††
		No. confirmed/ No. with TB diagnosis	(%)	HIV§-positive			Positive	Indeterminate		Cutoff	Positive			
				No. +/ No. tested	(%)			No. +/ No. tested	(%)		No. +/ No. tested	(%)		
Singapore§§	Adults	286/286	(100)	7/238	(3)	A	254/270	(94)	3/286	(1)	10 mm 15 mm	206/217 158/217	(95) (73)	0.84 ND¶¶
Spain***	Adults & children	NR/42	(NR)	NR†††	NR	B	36/39	(86)	3/42	(7)	5 mm	40/42	(95)	0.93
Germany§§§	Children aged 0–7 yrs	28/28	(100)	NR	NR	B	26/28	(93)	0/28	(0)	5 mm	28/28	(100)	0.49
South Korea¶¶¶¶	Adults	37/65	(57)	0/31	(0)	C	83/87	(95)	0/87	(0)	5 mm 10 mm	64/87 55/87	(74) (67)	<0.01 <0.01
Germany****	Adults	58/65	(89)	NR	NR	D	40/40	(100)	0/40	(0)	NR	35/40	(88)	0.05
Italy††††	Adults	23/23	(100)	0/23	(0)	E	21/23	(91)	NR	NR	ND	ND	ND	ND
Italy§§§§	Adults & children aged >15 yrs	13/24	(54)	NR	NR	F	20/24	(83)	0/24	(0)	5 mm	14/20	(54)	0.49
Germany¶¶¶¶¶	Adults	8/12	(67)	NR	NR	G	12/12	(100)	0/12	(0)	6 mm	8/10	(80)	0.39
South Korea*****	Adults & children aged >15 yrs	58/67	(87)	0/67	(0)	H	59/64	(92)	3/67	(4)	10 mm	45/66	(68)	<0.01
Switzerland†††††	Adults	89/89	(100)	0/89	(0)	I	61/61	(100)	1/62	(2)	ND	ND	ND	ND
Taiwan§§§§§	Adults & children aged 2–84 yrs	37/39	(95)	3/NR	(ND)	J	34/39	(87)	NR	NR	ND	ND	ND	ND
Switzerland¶¶¶¶¶¶	Adults & children aged >15 yrs	58/58	(100)	0/58	(0)	K	57/58	(98)	0/58	(2)	ND	ND	ND	ND
Turkey*****	Adults	NR/28	NR/28	NR	NR	B	26/28	(93)	NR	NR	10 mm	23/28	(82)	0.42
Turkey††††††	Adults & children aged >15 yrs	100/100	(100)	0/100	(0)	L	80/96	(83)	4/100	(4)	10 mm	80/99	(81)	0.79
Multiple European§§§§§§§	Adults	69/69 0/19	(100) (0)	3/NR 0/NR	(NR) (NR)	B	62/69 13/19	(90) (68)	0/69 0/19	(0) (0)	10 or 15	114/136 37/41	(84) (90)	0.06 0.09
Taiwan¶¶¶¶¶¶¶	Adults with extra-pulmonary TB	50/50 0/39	(100) (0)	2/NR	(NR)	M	40/50 31/39	(80) (79)	NR	NR	ND	ND	ND	ND
United Kingdom*****	Children	25/25 0/38	(100) (0)	0/35	(0)	F	14/24 17/34	(58) (50)	1/25 4/38	(8) (11)	10 mm	21/24 24/38	(86) (63)	0.05 0.38
Japan†††††††	Adults	49/49	(100)	NR	NR	N	47/47	(100)	2/49	(4)	ND	ND	ND	ND

See Table 5 footnotes on the following page.

**TABLE 5. (Continued) T-SPOT.TB (T-Spot) sensitivity\* results, by country in which study was conducted —12 countries, 2005–2009**

\* **Source:** Modified from Pai M, Zwerling A, Menzies D. Systematic review: T-cell-based assays for the diagnosis of latent tuberculosis infection: an update. *Ann Intern Med* 2008;149:177–84 supplemented with additional information and compared with TST specificity when available.

† Tuberculosis. Confirmed by culture and/or nucleic acid amplification test.

§ Human immunodeficiency virus.

¶ Tuberculin skin test.

\*\* "A" = T-Spot was interpreted as positive if a test well (with either early secretory antigenic target-6 [ESAT-6] or culture filtrate protein culture filtrate protein [CFP-10]) contained 6 spots or more than the negative control well and had at least twice the spots as the negative control well, and the negative control well had  $\leq 10$  spots; indeterminate if not "positive" and the mitogen control well had  $< 20$  spots or the negative control well had  $> 10$  spots. "B" = T-Spot interpretation criteria were not explicitly stated. "C" = T-Spot was interpreted as positive if a test well (with either ESAT-6 or CFP-10) contained 5 spots or more than the negative control well and had at least twice the spots as the negative control well and the negative control well had  $\leq 10$  spots and as indeterminate if the negative control well had  $> 10$  spots. "D" = T-Spot was interpreted as positive if a test well (with either ESAT-6 or CFP-10) contained 5 spots or more than the negative control well and had at least twice the spots as the negative control well and the mitogen control well had  $> 20$  spots and indeterminate if the mitogen control well had  $\leq 20$  spots. "E" = T-Spot was interpreted as positive if the well with ESAT-6 contained at least twice the average number of spots as the negative control well or the well with CFP-10 contained at least 4 times the average number of spots as the negative control well. "F" = T-Spot was interpreted as positive if a test well (with either ESAT-6 or CFP-10) contained 6 spots or more than the negative control well and had at least twice the spots as the negative control well and the negative control well had  $< 10$  spots, as indeterminate if not "positive" and the mitogen control well had  $< 20$  spots and the negative control well had  $< 10$  spots, as negative if not positive and spots in the negative control well were  $< 10$  and the spots in the mitogen control were  $\geq 20$ , and as technical error if the negative control well had  $\geq 10$  spots. "G" = T-Spot was interpreted as positive if a test well (with either ESAT-6 or CFP-10) contained 5 spots or more than the negative control well and had at least twice the spots as the negative control well; wells contained 200,000 PBMCs instead of 250,000 PBMCs as recommended by the manufacturer. "H" = T-Spot was interpreted as positive if a test well (with either ESAT-6 or CFP-10) contained 5 spots or more than the negative control well and had at least twice the spots as the negative control well; reported indeterminate results but did not explicitly state criteria; wells contained 200,000 PBMCs instead of 250,000 PBMCs as recommended by the manufacturer. "I" = T-Spot was interpreted as positive if a test well (with either ESAT-6 or CFP-10) contained 6 spots more than the negative control well and had at least twice the spots as the negative control well and the negative control well had  $\leq 10$  spots and as indeterminate if not "positive" and the mitogen control well had  $< 20$  spots or the negative control well had  $> 10$  spots. "J" = T-Spot was interpreted as positive if the mean number of spots in duplicate test wells (with either ESAT-6 or CFP-10) was 10 or more than the mean number of spots in duplicate negative control wells and at least twice the mean number of spots in the negative control wells; other criteria were not explicitly stated. "K" = T-Spot was interpreted as indeterminate if the mitogen control well had  $< 20$  spots and as positive if not indeterminate and a test well (with either ESAT-6 or CFP-10) contained  $> 6$  spots more than the negative control well. "L" = T-Spot was interpreted as indeterminate if the mitogen control well had  $\leq 20$  spots or the negative control well had  $\geq 10$  spots and as positive if not indeterminate and a test well (either ESAT-6 or CFP-10) contained 6 spots or more than the negative control well and had at least twice the number of spots as the negative control well. "M" = T-Spot was interpreted as positive if a test well (with either ESAT-6 or CFP-10) had  $\geq 10$  spots (when the negative control well had  $< 5$  spots), or at least twice the number of spots in the negative control well (when the negative control well had  $\geq 5$  spots). "N" = T-Spot was interpreted as positive if the Nil well had 0–5 spots and a test well (with either ESAT-6 or CFP-10) had  $\geq 6$  spots more than the Nil well or if the Nil well had 6–10 spots and a test well had at least twice the number of spots as the negative control well; test is indeterminate if the number of spots in the Nil well is  $> 10$  or the number of spots in the mitogen well is  $< 20$  and neither test well is positive.

†† Fisher's exact test was used by CDC to calculate 2-tailed p-values.

§§ **Source:** Chee CB, Gan SH, KhinMar KW, et al. Comparison of sensitivities of two commercial gamma interferon release assays for pulmonary tuberculosis. *J Clin Microbiol* 2008;46:1935–40.

¶¶ Not done.

\*\*\* **Source:** Dominguez J, Ruiz-Manzano J, De Souza-Galvao M, et al. Comparison of two commercially available gamma interferon blood tests for immunodiagnosis of tuberculosis. *Clin Vaccine Immunol* 2008;15:168–71.

††† Not reported.

§§§ **Source:** Detjen AK, Keil T, Roll S et al. Interferon-gamma release assays improve the diagnosis of tuberculosis and nontuberculous mycobacterial disease in children in a country with a low incidence of tuberculosis. *Clin Infect Dis* 2007;45:322–8.

¶¶¶ **Source:** Lee JY, Choi HJ, Park IN, et al. Comparison of two commercial interferon-gamma assays for diagnosing *Mycobacterium tuberculosis* infection. *Eur Respir J* 2006;28:24–30.

\*\*\*\* **Source:** Meier T, Eulenbruch HP, Wrighton-Smith P, Enders G, Regnath T. Sensitivity of a new commercial enzyme-linked immunospot assay (T SPOT-TB) for diagnosis of tuberculosis in clinical practice. *Eur J Clin Microbiol Infect Dis* 2005;24:529–36.

†††† **Source:** Goletti D, Carrara S, Vincenti D, et al. Accuracy of an immune diagnostic assay based on RD1 selected epitopes for active tuberculosis in a clinical setting: a pilot study. *Clin Microbiol Infect* 2006;12:544–50.

§§§§ **Source:** Ferrara G, Losi M, D'Amico R, et al. Use in routine clinical practice of two commercial blood tests for diagnosis of infection with *Mycobacterium tuberculosis*: a prospective study. *Lancet* 2006;367:1328–34.

¶¶¶¶ **Source:** Jafari C, Ernst M, Kalsdorf B, et al. Rapid diagnosis of smear-negative tuberculosis by bronchoalveolar lavage enzyme-linked immunospot. *Am J Respir Crit Care Med* 2006;174:1048–54.

\*\*\*\*\* **Source:** Kang YA, Lee HW, Hwang SS, et al. Usefulness of whole-blood interferon-gamma assay and interferon-gamma enzyme-linked immunospot assay in the diagnosis of active pulmonary tuberculosis. *Chest* 2007;132:959–65.

††††† **Source:** Bosshard V, Roux-Lombard P, Perneger T, et al. Do results of the T-SPOT.TB interferon-gamma release assay change after treatment of tuberculosis? *Respir Med* 2009;103:30–4.

§§§§§ **Source:** Wang JY, Chou CH, Lee LN, et al. Diagnosis of tuberculosis by an enzyme-linked immunospot assay for interferon-gamma. *Emerg Infect Dis* 2007;13:553–8.

¶¶¶¶¶ **Source:** Janssens JP, Roux-Lombard P, Perneger T, Metzger M, Vivien R, Rochat T. Quantitative scoring of an interferon-gamma assay for differentiating active from latent tuberculosis. *Eur Respir J* 2007;30:722–8.

\*\*\*\*\* **Source:** Ozekinci T, Ozbek E, Celik Y. Comparison of tuberculin skin test and a specific T-cell-based test, T-Spot.TB, for the diagnosis of latent tuberculosis infection. *J Int Med Res* 2007;35:696–703.

††††† **Source:** Soysal A, Torun T, Efe S, Gencer H, Tahaoglu K, Bakir M. Evaluation of cut-off values of interferon-gamma-based assays in the diagnosis of *M. tuberculosis* infection. *Int J Tuberc Lung Dis* 2008;12:50–6.

§§§§§ **Source:** Goletti D, Stefania C, Butera O, et al. Accuracy of immunodiagnostic tests for active tuberculosis using single and combined results: a multicenter TBNET-Study. *PLoS ONE* 2008;3:e3417. [E-published]. Available at <http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0003417>.

¶¶¶¶¶ **Source:** Liao CH, Chou CH, Lai CC, et al. Diagnostic performance of an enzyme-linked immunospot assay for interferon-gamma in extrapulmonary tuberculosis varies between different sites of disease. *J Infect* 2009;59:402–8.

\*\*\*\*\* **Source:** Kampmann B, Whittaker E, Williams A, et al. Interferon-gamma release assays do not identify more children with active tuberculosis than the tuberculin skin test. *Eur Respir J* 2009;33:1374–82.

††††† **Source:** Higuchi K, Kawabe Y, Mitarai S, Yoshiyama T, Harada N, Mori T. Comparison of performance in two diagnostic methods for tuberculosis infection. *Med Microbiol Immunol* 2009;198:33–7.

**TABLE 6. QuantiFERON-TB Gold In-Tube test (QFT-GIT) specificity,\* by country in which study was conducted — four countries, 2007–2008**

		BCG <sup>†</sup> -vaccinated		HIV <sup>§</sup> -positive		QFT-GIT results					TST <sup>¶</sup> Results			% TST- vs. % QFT-GIT- p-value <sup>††</sup>
						Negative			Indeterminate		Negative			
		No. vaccinated/ No. evaluated	(%)	No. +/ No. tested	(%)	Inter- pretation criteria <sup>**</sup>	No. +/ No. valid	(%)	No. +/ No. tested	(%)	Cutoff	No. +/ No. tested	(%)	
Country	Subjects													
Germany <sup>§§</sup>	Children aged 0–11 yrs w/ lymphadenitis	0/23	(0)	NR <sup>¶¶</sup>	NR	A	19/19	(100)	ND <sup>***</sup>	ND	5	2/23	(9)	<0.01
											10	5/23	(22)	<0.01
Germany <sup>§§</sup>	Children aged 0–7 yrs w/ respi-rator infection	0/22	(0)	NR	NR	A	21/21	(100)	ND	ND	5	22/22	(100)	1.0
											10	22/22	(100)	1.0
Japan <sup>†††</sup>	Adult students	140/168	(83)	0/168	(0)	B	158/160	(99)	6/168	(4)	ND	ND	ND	ND
Denmark <sup>§§§</sup>	High school students & staff	38/124	(31)	0/124	(0)	C	124/124	(100)	0	(0)	10	116/124	(94)	<0.01
Italy <sup>¶¶¶</sup>	Mostly adults	1/14	(7)	0/14	(0)	C	14/14	(100)	0/14	(0)	NR	8/8	(100)	ND

\* **Source:** Modified from Pai M, Zwerling A, Menzies D. Systematic review: T-cell-based assays for the diagnosis of latent tuberculosis infection: an update. *Ann Intern Med* 2008;149:177–84 supplemented with additional information and compared with TST specificity when available.

<sup>†</sup> *Bacillus Calmette-Guérin*.

<sup>§</sup> Human immunodeficiency virus.

<sup>¶</sup> Tuberculin skin test.

\*\* "A" indicates that QFT-GIT was interpreted as positive if Tuberculosis (TB) Response was  $\geq 0.35$  IU/mL; Mitogen Response was not measured. "B" indicates that QFT-GIT was interpreted as positive if TB Response was  $\geq 0.35$  IU/mL and Nil was  $\leq 8.0$  IU/mL, as indeterminate if Nil  $\geq 8.0$  IU/mL or the TB Response was  $< 0.35$  IU/mL and the Mitogen Response was  $< 0.5$  IU/mL, and as negative if the TB Response was  $< 0.35$  IU/mL, the Mitogen Response was  $\geq 0.5$  IU/mL, and Nil was  $\leq 8.0$  IU/mL. "C" indicates that QFT-GIT interpretation criteria were not explicitly stated.

†† Fisher's exact test was used by CDC to calculate 2-tailed p-values.

§§ **Source:** Detjen AK, Keil T, Roll S, et al. Interferon-gamma release assays improve the diagnosis of tuberculosis and nontuberculous mycobacterial disease in children in a country with a low incidence of tuberculosis. *Clin Infect Dis* 2007;45:322–8.

¶¶ Not reported.

\*\*\* Not done.

††† **Source:** Harada N, Higuchi K, Yoshiyama T, et al. Comparison of the sensitivity and specificity of two whole blood interferon-gamma assays for *M. tuberculosis* infection. *J Infect* 2008;56:348–53.

§§§ **Source:** Ruhwald M, Bodmer T, Maier C, et al. Evaluating the potential of IP-10 and MCP-2 as biomarkers for the diagnosis of tuberculosis. *Eur Respir J* 2008;32:1607–15.

¶¶¶ **Source:** Palazzo R, Spensieri F, Massari M, et al. Use of whole-blood samples in in-house bulk and single-cell antigen-specific gamma interferon assays for surveillance of *Mycobacterium tuberculosis* infections. *Clin Vaccine Immunol* 2008;15:327–37.



TABLE 7. T-SPOT.TB test (T-Spot) specificity,\* by country in which study was conducted — three countries, 2006–2008

Country	Subjects	BCG <sup>†</sup> -vaccinated		HIV <sup>§</sup> status	Inter-pretation criteria <sup>**</sup>	T-Spot results				TST <sup>¶</sup> results			% TST-vs. % T-Spot-p-value <sup>††</sup>
		No. vaccinated/ No. evaluated	(%)			Negative		Indeterminate		Cutoff	Negative		
						No. +/ No. valid	(%)	No. +/ No. tested	(%)		No. +/ No. tested	(%)	
Germany <sup>§§</sup>	Children aged 0–11 yrs w/ lymphadenitis	0/19	(0)	NR <sup>¶¶</sup>	A	18/19	(95)	4/23	(17)	5 10	2/23 5/23	(9) (22)	<0.01 <0.01
Germany <sup>***</sup>	Children aged 0–7 yrs w/ other respiratory infection	0/21	(0)	NR	A	21/21	(100)	1/22	(5)	5 10	22/22 22/22	(100) (100)	1.0 1.0
South Korea <sup>†††</sup>	High school students	131/131	(100)	NR	B	111/ 131	(85)	0/131	(0)	10 15	103/131 125/131	(79) (95)	0.26 <0.01
United States <sup>§§§</sup>	Adults with & w/o prior MAC <sup>¶¶¶</sup> disease	0/18	(0)	NR	C	17/18	(94)	0/18	(0)	ND	ND	ND	ND

\* **Source:** Modified from Pai M, Zwerling A, Menzies D. Systematic review: T-cell-based assays for the diagnosis of latent tuberculosis infection: an update. *Ann Intern Med* 2008;149:177–84 supplemented with additional information and compared with TST specificity when available.

<sup>†</sup> *Bacillus Calmette-Guerin*.

<sup>§</sup> Human immunodeficiency virus.

<sup>¶</sup> Tuberculin skin test.

\*\* "A" indicates that T-Spot interpretation criteria were not explicitly stated. "B" indicates that T-Spot was interpreted as positive if a test well (with either early secretory antigenic target-6 [ESAT-6] or culture filtrate protein-10 [CFP-10]) contained 5 spots or more than the negative control well and had at least twice the spots as the negative control well and the negative control well had ≤10 spots and as indeterminate if the negative control well had >10 spots. "C" indicates that T-Spot was interpreted as positive if a test well (with either ESAT-6 or CFP-10) contained 6 spots or more than the negative control well and had at least twice the spots as the negative control well and as indeterminate if not "positive" and the mitogen control well had <20 spots.

†† Fisher's exact test was used by CDC to calculate 2-tailed p-values.

§§ **Source:** Detjen AK, Keil T, Roll S et al. Interferon-gamma release assays improve the diagnosis of tuberculosis and nontuberculous mycobacterial disease in children in a country with a low incidence of tuberculosis. *Clin Infect Dis* 2007;45:322–8.

¶¶ Not reported.

\*\*\* **Source:** Detjen AK, Keil T, Roll S et al. Interferon-gamma release assays improve the diagnosis of tuberculosis and nontuberculous mycobacterial disease in children in a country with a low incidence of tuberculosis. *Clin Infect Dis* 2007;45:322–8.

††† **Source:** Lee JY, Choi HJ, Park IN et al. Comparison of two commercial interferon-gamma assays for diagnosing *Mycobacterium tuberculosis* infection. *Eur Respir J* 2006;28:24–30.

§§§ **Source:** Adams LV, Waddell RD, von Reyn CF. T-SPOT.TB Test results in adults with *Mycobacterium avium* complex pulmonary disease. *Scand J Infect Dis* 2008;40:196–203.

¶¶¶ *Mycobacterium avium* complex.

**TABLE 8. Summary of findings of published studies evaluating QuantiFERON-TB Gold-In-Tube test (QFT-GIT) and/or T-SPOT.TB test (T-Spot) in tuberculosis contacts compared with tuberculin skin test (TST) when available, by country in which study was conducted — seven countries, 2006–2008**

Country	Subjects	BCG vaccinated*		TST cutoff	Findings
		No. vaccinated/No. evaluated	(%)		
South Africa†	Children aged 5–15 yrs	115/174	(66)	10 mm	QFT-GIT and TST results were associated with older age but not with recent or remote household contact.
Nigeria§	Child contacts & controls aged 1–14 yrs	187/207	(90)	10 mm	QFT-GIT and TST results were associated with acid-fast bacillus (AFB) status of source and age for children living with AFB-negative persons and controls. +TST/-QFT-GIT discordance was more common in controls and children living with AFB-negative persons. -TST/+ QFT-GIT were more common in children living with AFB-positive persons.
Denmark¶	Adult contacts w/out BCG	0/785	(0)	10 mm	TST results were associated with age but not with estimates of exposure. T-Spot results were associated with an estimate of exposure (cumulative shopping time). QFT-GIT (without mitogen) was associated with cumulative shopping time more so than T-Spot.
The Gambia**	Adult & child contacts aged ≥15 yrs	84/194	(43)	10 mm	TST more strongly associated with exposure gradient than QFT-GIT (without mitogen). For contacts sleeping in the same room as compared with those sleeping in different houses, the odds ratio for a positive TST was 4.8 (95% confidence interval [CI] = 1.3–17.1) as compared with 3.8 (CI = 1.2–12.5) for QFT-GIT.
Switzerland††	Adult & child contacts aged 16–83 yrs	238/295	(81)	10 mm	Both TST & T-Spot results were associated with age, gender, BCG, and incidence of tuberculosis in country of origin, but not to any of 5 exposure scores.
Germany§§	Adult & child contacts w/ TST >5 mm	453/812	(56)	NA	Both QFT-GIT & T-Spot results were associated with age, AFB + or coughing source, cumulative exposure time, and foreign origin. Associations with TST results were not assessed.
Spain¶¶	Adults & children	128/270	(47)	5 mm	TST results were associated with BCG. QFT-GIT & T-Spot results were not associated with BCG. Association of test results with incidence of tuberculosis in country of origin was not assessed.

\* Bacillus Calmette-Guerin.

† **Source:** Tsiouris SJ, Austin J, Toro P et al. Results of a tuberculosis-specific IFN-gamma assay in children at high risk for tuberculosis infection. *Int J Tuberc Lung Dis* 2006;10:939–41.

§ **Source:** Nakaoka H, Lawson L, Squire SB, et al. Risk for tuberculosis among children. *Emerg Infect Dis* 2006;12:1383–8.

¶ **Source:** Arend SM, Thijsen SF, Leyten EM, et al. Comparison of two interferon-gamma assays and tuberculin skin test for tracing tuberculosis contacts. *Am J Respir Crit Care Med* 2007;175:618–27.

\*\* **Source:** Adetifa IM, Lugos MD, Hammond A, et al. Comparison of two interferon gamma release assays in the diagnosis of *Mycobacterium tuberculosis* infection and disease in The Gambia. *BMC Infect Dis* 2007;7:122.

†† **Source:** Janssens J, Roux-Lombard P, Perneger T, Metzger M, Vivien R, Rochat T. Contribution of a IFN-gamma assay in contact tracing for tuberculosis in a low-incidence, high immigration area. *Swiss Med Wkly* 2008;138:585–93.

§§ **Source:** Diel R, Loddenkemper R, Meywald-Walter K, Gottschalk R, Nienhaus A. Comparative performance of tuberculin skin test, QuantiFERON-TB-Gold In Tube assay, and T-Spot.TB test in contact investigations for tuberculosis. *Chest* 2009;135:1010–8.

¶¶ **Source:** Dominguez J, Ruiz-Manzano J, De Souza-Galvao M, et al. Comparison of two commercially available gamma interferon blood tests for immunodiagnosis of tuberculosis. *Clin Vaccine Immunol* 2008;15:168–71.

**TABLE 9. QuantiFERON-TB Gold In-Tube (QFT-GIT) test results in immunosuppressed persons compared with tuberculin skin test (TST) results when available — 10 countries, 2006–2008**

Country	Subjects	HIV* status	QFT-GIT results				TST results			% TST+ vs. % QFT-GIT+† p-value†
			Positive		Indeterminate		Cutoff	Positive		
			No. +/ No. valid	(%)	No. +/ No. tested	(%)		No. +/ No. tested	(%)	
Denmark§	607 adults	607 HIV+	27/570	(4.7)	20/590	(3.4)	ND¶	ND	ND	ND
Chile**	116 adults	116 HIV+	17/115	(15)	0/115	(0)	5 mm	12/110	(11)	0.50
United States††	207 adults	207 HIV+	11/191	(6)	10/201	(5)	5 mm	13/201	(7)	0.94
United States§§	294 adults	294 HIV+	25/279	(9)	15/294	(5)	5 mm	19/205	(9)	0.99
Zambia¶¶	112 adults with smear + TB	59 HIV+	37/49	(76)	10/59	(17)	5 mm	26/47	(55)	0.06
		37 HIV-	31/32	(97)	5/37	(14)		25/31	(81)	0.09
		16 not tested	15/15	(100)	1/16	(6)		0/14	(0)	<0.01
South Africa***	154 adults with Culture + TB	26 HIV+	17/21	(81)	5/26	(19)	10 mm	22/26	(85)	0.99
		15 HIV-	11/15	(73)	0/15	(0)	5 mm	15/15	(100)	0.09
		113 not tested	72/95	(76)	18/113	(16)	10 mm	67/113	(59)	0.02
Austria†††	8 adults w/TB at baseline	8 HIV+	7/8	(88)	0/8	(0)	5 mm	8/8	(100)	ND
Austria†††	822 adults w/o TB at baseline	822 HIV+	37/775	(5)	47/822	(6)	5 mm	23/34§§§	(74)	ND
United States¶¶¶	336 adults	336 HIV+	9/330	(3)	6/336	(2)	4 mm	7/278	(3)	0.92
Italy****	69 TNFi†††† candidates	69 HIV-	22/67	(33)	2/69	(3)	5 mm	18/69	(26)	0.49
Turkey	68 adult TNFi candidates	68 unknown	9/61	(15)	7/68	(10)	10 mm	37/61	(61)	<0.01
Switzerland§§	142 adults with autoimmune disease	142 unknown	17/134	(13)	8/142	(6)	5 mm	46/115	(40)	<0.01
Peru¶¶¶¶	106 adults with rheumatoid arthritis	106 unknown	45/104	(43)	2/106	(1)	5 mm	27/101	(27)	0.02

\* Human immunodeficiency virus.

† Fisher's exact test was used by CDC to calculate 2-tailed p-values.

§ **Source:** Brock I, Ruhwald M, Lundgren B, Westh H, Mathiesen LR, Ravn P. Latent tuberculosis in HIV positive, diagnosed by the *M. tuberculosis* Specific Interferon Gamma test. *Respir Res* 2006;7:56.

¶ Not done.

\*\* **Source:** Balcells ME, Perez CM, Chanqueo L et al. A comparative study of two different methods for the detection of latent tuberculosis in HIV-positive individuals in Chile. *Int J Infect Dis* 2008;12:645–52.†† **Source:** Jones S, de Gijzel D, Wallach FR, Gurtman AC, Shi Q, Sacks H. Utility of QuantiFERON-TB Gold in-tube testing for latent TB infection in HIV-infected individuals. *Int J Tuberc Lung Dis* 2007;11:1190–5.§§ **Source:** Luetkemeyer AF, Charlebois ED, Flores LL, et al. Comparison of an interferon-gamma release assay with tuberculin skin testing in HIV-infected individuals. *Am J Respir Crit Care Med* 2007;175:737–42.¶¶ **Source:** Raby E, Moyo M, Devendra A, et al. The effects of HIV on the sensitivity of a whole blood IFN-gamma release assay in Zambian adults with active tuberculosis. *PLoS ONE* 2008;3:e2489. [E-pub]. Available at <http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0002489>.\*\*\* **Source:** Tsiouris SJ, Coetzee D, Toro PL, Austin J, Stein Z, el-Sadr W. Sensitivity analysis and potential uses of a novel gamma interferon release assay for diagnosis of tuberculosis. *J Clin Microbiol* 2006;44:2844–50.††† **Source:** Aichelburg MC, Rieger A, Breitenacker F, et al. Detection and prediction of active tuberculosis disease by a whole-blood interferon-gamma release assay in HIV-1-infected individuals. *Clin Infect Dis* 2009;48:954–62.

§§§ Tuberculin skin testing was offered only to subjects with a positive QFT-GIT.

¶¶¶ **Source:** Talati NJ, Seybold U, Humphrey B, et al. Poor concordance between interferon-gamma release assays and tuberculin skin tests in diagnosis of latent tuberculosis infection among HIV-infected individuals. *BMC Infect Dis* 2009;9:15.\*\*\*\* **Source:** Bocchino M, Matarese A, Bellofiore B, et al. Performance of two commercial blood IFN-gamma release assays for the detection of *Mycobacterium tuberculosis* infection in patient candidates for anti-TNF-alpha treatment. *Eur J Clin Microbiol Infect Dis* 2008;27:907–13.†††† Tumor necrosis factor  $\alpha$  inhibitor.§§§§ **Source:** Cobanoglu N, Ozcelik U, Kalyoncu U, et al. Interferon-gamma assays for the diagnosis of tuberculosis infection before using tumour necrosis factor-alpha blockers. *Int J Tuberc Lung Dis* 2007;11:1177–82.¶¶¶¶ **Source:** Matulis G, Juni P, Villiger PM, Gadola SD. Detection of latent tuberculosis in immunosuppressed patients with autoimmune diseases: performance of a *Mycobacterium tuberculosis* antigen-specific interferon gamma assay. *Ann Rheum Dis* 2008;67:84–90.\*\*\*\* **Source:** Ponce de LD, Acevedo-Vasquez E, Alvizuri S, et al. Comparison of an interferon-gamma assay with tuberculin skin testing for detection of tuberculosis (TB) infection in patients with rheumatoid arthritis in a TB-endemic population. *J Rheumatol* 2008;35:776–81.

**TABLE 10. Published studies evaluating T-SPOT.TB test (T-Spot) among immunosuppressed persons compared with tuberculin skin test (TST) when available — eight countries, 2006–2008**

Country	Subjects	HIV* Status	T-Spot results				TST results			% TST+ vs. % T-Spot+ p-value†
			Positive		Indeterminate		Cutoff	Positive		
			No. +/ No. valid	(%)	No. +/ No. tested	(%)		No. +/ No. tested	(%)	
South Africa§	20 HIV+ adults	20 HIV+	13/18	(72)	2/20	(10)	5 mm	10/16	(63)	0.81
	23 HIV+ children	23 HIV+	12/23	(52)	0/23	(0)		6/23	(26)	0.13
South Africa¶	160 adults at HIV screening clinic	74 HIV+	38/73	(52)	1/74	(1)	5 mm	35/67	(52)	0.99
		86 HIV-	51/86	(59)	0/86	(0)		66/77	(86)	<0.01
Germany**	286 HIV+ outpatients	286 HIV+	66/267	(25)	8/275	(3)	5 mm	33/275	(12)	<0.01
United States††	336 HIV+ adults	336 HIV+	14/289	(5)	47/336	(14)	5 mm	7/278	(2.5)	0.21
Italy§§	69 HIV- TNFi¶¶ candidates	69 HIV-	21/65	(32)	4/69	(6)	5 mm	18/ 69	(26)	0.55
Hong Kong***	134 adults w/ silicosis	134 unknown	86/128	(67)	6†††/134	(5)	10 mm	92/134	(69)	0.90
Germany§§§	48 patients awaiting liver transplant	48 unknown	4/48	(8)	0/48	(0)	5 mm	6/47	(13)	0.71
Canada¶¶¶	203 patients on hemodialysis	203 unknown	72/189	(38)	14/203	(7)	10 mm	19/203	(9)	<0.01
Italy****	138 patients w/ hematologic disease	138 HIV-	61/129	(47)	6/135	(4)	5 mm	24/122	(20)	<0.01
United States††††	49 inmates w/ hx IVDU§§§§ (of 390 total in study)	49 unknown	17/49	(35)	0/49	(0)	10 mm	6/49	(12)	0.02
Greece¶¶¶¶	70 HIV- TNFi candidates	70 HIV-	16/70	(23)	0/70	(0)	5 mm	27/70	(39)	0.07

\* Human immunodeficiency virus.

† Fisher's exact test was used by CDC to calculate 2-tailed p-values.

§ **Source:** Mandalakas AM, Hesselning AC, Chegou NN, et al. High level of discordant IGRA results in HIV-infected adults and children. *Int J Tuberc Lung Dis* 2008;12:417–23.¶ **Source:** Rangaka MX, Wilkinson KA, Seldon R, et al. Effect of HIV-1 infection on T-Cell-based and skin test detection of tuberculosis infection. *Am J Respir Crit Care Med* 2007;175:514–20.\*\* **Source:** Stephan C, Wolf T, Goetsch U, et al. Comparing QuantiFERON-tuberculosis gold, T-SPOT tuberculosis and tuberculin skin test in HIV-infected individuals from a low prevalence tuberculosis country. *AIDS* 2008;22:2471–9.†† **Source:** Talati NJ, Seybold U, Humphrey B, et al. Poor concordance between interferon-gamma release assays and tuberculin skin tests in diagnosis of latent tuberculosis infection among HIV-infected individuals. *BMC Infect Dis* 2009;9:15.§§ **Source:** Bocchino M, Matarese A, Bellofiore B, et al. Performance of two commercial blood IFN-gamma release assays for the detection of *Mycobacterium tuberculosis* infection in patient candidates for anti-TNF-alpha treatment. *Eur J Clin Microbiol Infect Dis* 2008;27:907–13.¶¶ Tumor necrosis factor  $\alpha$  inhibitor.\*\*\* **Source:** Leung CC, Yam WC, Yew WW, et al. Comparison of T-Spot.TB and tuberculin skin test among silicotic patients. *Eur Respir J* 2008;31:266–72.

††† Reclassified with second test.

§§§ **Source:** Lindemann M, Dioury Y, Beckebaum S, et al. Diagnosis of tuberculosis infection in patients awaiting liver transplantation. *Hum Immunol* 2009;70:24–8.¶¶¶ **Source:** Passalent L, Khan K, Richardson R, Wang J, Dedier H, Gardam M. Detecting latent tuberculosis infection in hemodialysis patients: a head-to-head comparison of the T-SPOT.TB test, tuberculin skin test, and an expert physician panel. *Clin J Am Soc Nephrol* 2007;2:68–73.\*\*\*\* **Source:** Piana F, Codecasa LR, Cavallerio P, et al. Use of a T-cell-based test for detection of tuberculosis infection among immunocompromised patients. *Eur Respir J* 2006;28:31–4.†††† **Source:** Porsa E, Cheng L, Graviss EA. Comparison of an ESAT-6/CFP-10 peptide-based enzyme-linked immunospot assay to a tuberculin skin test for screening of a population at moderate risk of contracting tuberculosis. *Clin Vaccine Immunol* 2007;14:714–9.

§§§§ Intravenous- drug user.

¶¶¶¶ **Source:** Vassilopoulos D, Stamoulis N, Hadziyannis E, Archimandritis AJ. Usefulness of enzyme-linked immunospot assay (elispot) compared to tuberculin skin testing for latent tuberculosis screening in rheumatic patients scheduled for anti-tumor necrosis factor treatment. *J Rheumatol* 2008;35:1271–6.

### **IGRA Expert Committee Members** **Membership as of August 2008**

**Chair:** Neil Schluger, MD, Columbia University, New York, New York

**Moderator:** John Seggerson, Stop TB USA, Atlanta, GA

**Members:** Paul Barnicott, U.S. Air Force School of Aerospace Medicine, San Antonio, Texas; John Bernardo, MD, Boston University School of Medicine, Boston, Massachusetts; Henry M. Blumberg, MD, Emory University School of Medicine, Atlanta, Georgia; Helene Calvet, MD, Long Beach Dept. of Health and Human Services, Long Beach, California; Charles Daley, MD, National Jewish Medical and Research Center, Denver, Colorado; Susan Dorman, MD, Johns Hopkins University School of Medicine, Baltimore, Maryland; Edward Graviss, PhD, Baylor College of Medicine, Houston, Texas; Tiffany Harris, PhD, New York City Dept. of Health and Mental Hygiene, New York, New York; Philip Hill, MD, University of Otago School of Medicine, Dunedin, New Zealand; Masae Kawamura, MD, San Francisco Department of Public Health, San Francisco, California; Lisa Keep, MD, Uniformed Services Univ. of the Health Sciences, Bethesda, Maryland; Stephen Kralovic, MD, Cincinnati VA Medical Center, Cincinnati, Ohio; Michael Leonard, MD, Georgia Department of Human Resources, Atlanta, Georgia; David Lewinsohn, MD, PhD, Oregon Health and Sciences University, Portland VA Medical Center, Deborah Lewinsohn, MD, Oregon Health and Sciences University, Portland, Oregon; Kathleen Moser, MD, San Diego County Department of Health, Poway, California; Edward Nardell, MD, Brigham and Women's Hospital, Boston, Massachusetts; Masa Narita, MD, Seattle and King County Public Health, Seattle, Washington; Richard O'Brien, MD, Foundation for Innovative New Diagnostics, Geneva, Switzerland; Randall Reves, MD, Denver Public Health Department, Denver, Colorado; Luca Richeldi, MD, PhD, University of Modena and Reggio Emilia, Modena, Italy; Kim Connelly Smith, MD, University of Texas Health Science Center, Jeffery Starke, MD, Texas Children's Hospital, Baylor College of Medicine, Houston, Texas; David Warshauer, PhD, Wisconsin State Laboratory of Hygiene, Madison, Wisconsin; Gail Woods, MD, Central Arkansas Veterans Healthcare System, Little Rock, Arkansas.

### **IGRA Expert Committee Presenters** **Membership as of August 2008**

**Members:** Sandra Arend, MD, PhD, Leiden University Medical Center, Leiden, The Netherlands; John Bernardo, MD, Boston University School of Medicine, Boston, Massachusetts; Henry M. Blumberg, MD, Emory University School of Medicine, Atlanta, Georgia; Charles Daley, MD, National Jewish Medical and Research Center, Denver, Colorado; Roland Diel, MD, University of Düsseldorf, School of Public Health, Düsseldorf, Germany; Edward Graviss, MD, Baylor College of Medicine, Houston, Texas; Tiffany Harris, PhD, New York City Dept. of Health and Mental Hygiene, New York, New York; Anthony Hawkrigde, MD, Aeras Global TB Vaccine Foundation, Cape Town, South Africa; Philip Hill, MD, University of Otago School of Medicine, Dunedin, New Zealand; Masae Kawamura, MD, San Francisco Department of Public Health, San Francisco, California; Deborah Lewinsohn, MD, Portland VA Medical Center, David Lewinsohn, MD, PhD, Oregon Health and Sciences University, Portland, Oregon; Hassan Mahomed, Mmed, University of Cape Town, Cape Town, South Africa; Freddie Poole, MS, Center for Devices and Radiological Health, Food and Drug Administration, Rockville, Maryland; Luca Richeldi, MD, PhD, University of Modena and Reggio Emilia, Modena, Italy; James Rothel, PhD, Cellectis Limited, Carnegie, Victoria, Australia; Neil Schluger, MD, Columbia University, New York, New York; John Seggerson, STOP TB USA, Atlanta, Georgia; Kim Connelly Smith, MD, University of Texas Health Science Center, Houston, Texas; Peter Wrighton-Smith, DPhil, Oxford Immunotec, Inc., Oxford, United Kingdom; Jean-Pierre Zellweger, MD, Swiss Lung Association, Lausanne, Switzerland; Kenneth Castro, MD, John Jereb, MD, Gerald Mazurek, MD, CDC, Atlanta, Georgia.

The *Morbidity and Mortality Weekly Report (MMWR)* Series is prepared by the Centers for Disease Control and Prevention (CDC) and is available free of charge in electronic format. To receive an electronic copy each week, visit *MMWR*'s free subscription page at <http://www.cdc.gov/mmwr/mmwrsubscribe.html>. Paper copy subscriptions are available through the Superintendent of Documents, U.S. Government Printing Office, Washington, DC 20402; telephone 202-512-1800.

Data presented by the Notifiable Disease Data Team and 122 Cities Mortality Data Team in the weekly *MMWR* are provisional, based on weekly reports to CDC by state health departments. Address all inquiries about the *MMWR* Series, including material to be considered for publication, to Editor, *MMWR* Series, Mailstop E-90, CDC, 1600 Clifton Rd., N.E., Atlanta, GA 30333 or to [mmwrq@cdc.gov](mailto:mmwrq@cdc.gov).

All material in the *MMWR* Series is in the public domain and may be used and reprinted without permission; citation as to source, however, is appreciated.

Use of trade names and commercial sources is for identification only and does not imply endorsement by the U.S. Department of Health and Human Services.

References to non-CDC sites on the Internet are provided as a service to *MMWR* readers and do not constitute or imply endorsement of these organizations or their programs by CDC or the U.S. Department of Health and Human Services. CDC is not responsible for the content of these sites. URL addresses listed in *MMWR* were current as of the date of publication.

# Study Guide

## Session 5

### Evaluating Clinical Practice Guidelines

Allen F. Shaughnessy, PharmD, MMedEd

**The aims** of this session are to:

- 1) Introduce you to the different types of clinical practice guidelines available to practicing physicians;
- 2) Present some issues that threaten the validity of recommendations in practice guidelines
- 3) Practice how to quickly and accurately evaluate clinical practice guidelines to determine their validity

**Specific Objectives:** By completing the initial reading and participating in class, students should be able to:

- 1) Quickly identify a guideline as being expert-based, evidence-based, or evidence-linked
- 2) Use the National Guideline Clearinghouse to find relevant clinical guidelines
- 3) Explain financial and intellectual conflicts of interest and how they can affect guideline recommendations
- 4) Use as set of questions to quickly evaluate a clinical practice guideline for threats to validity.

This study guide provides an outline of the concepts necessary to meet these objectives. It contains hyperlinks to short videos, web pages, or articles that explain the concepts in other ways or in greater detail. You can follow these hyperlinks if the explanations and examples I've given you are not sufficient to help you understand and to help you complete the readiness assessment test.



For more information, read (optional): [Standards for Developing Trustworthy Clinical Practice Guidelines, Institute of Medicine](#) or [How NICE clinical guidelines are developed: an overview for stakeholders, the public, and the NHS.](#)



## What are clinical practice guidelines?

A **clinical practice guideline** aims to guide decisions by providing criteria regarding diagnosis, management, and treatment in specific areas of healthcare.

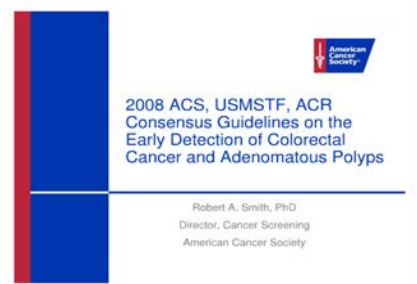
Guidelines have been produced throughout the history and have been based on tradition or authority. For example, a leading book on the treatment of sexually transmitted infections, written by a venerated venereal authority in 1859, states:

*iodide* of mercury. In my opinion, mercury should never be discarded except when there is a very decided repulsion on the part of the system, and even then it should not be wholly thrown aside. I have seen, indeed, patients at first decidedly anti-pathic to mercury, and in whom it produced unpleasant effects in the abdomen and mouth; some of these were debilitated by the smallest doses of mercury, and yet after having been strengthened by the preparations of iodine or iron, these same patients could take and tolerate the mercury, and it finally became the means of their cure. I have already mentioned these practical facts. I repeat, that I prefer to begin with small doses of mercury, for I fear that large doses will compel me to suspend the treatment, which would be an unfortunate circumstance, tending to favor the relapses and to make the subsequent treatment more difficult. However, I must acknowledge that I do sometimes discontinue the use of mercury even when it is well borne, but it is only when it has been used for a long time, as for two months, without arresting the progress of the disease. Then, if the patient retain his strength, I administer,

(Vidal, A-T. [\*A Treatise on venereal diseases.\*](#))

There is a natural tendency to turn to experts for information and authority-based guidelines continue to flourish. They are often called “consensus guidelines.” Frequently, this type of guidelines comes from professional groups or societies:

**International Consensus Conferences  
in Intensive Care Medicine: Noninvasive  
Positive Pressure Ventilation in Acute  
Respiratory Failure**  
Organized Jointly by the American Thoracic  
Society, the European Respiratory Society, the  
European Society of Intensive Care Medicine, and  
the Société de Réanimation de Langue Française,  
and approved by the ATS Board of Directors,



The goal today is to have guidelines based on evidence, and most guidelines are now labeled as being evidence-based. *Evidence-based*, as we will discuss, does not necessarily mean “trustworthy.”



## Who produces clinical practice guidelines?

There are more than 3,700 guidelines from 39 countries in the Guidelines International Network database. These come from:

- Government agencies, e.g., The U.S. Preventive Services Task Force
- Medical associations, e.g., The American Thoracic Society
- Managed care organizations, i.e., insurers
- Other organizations, e.g., The Global Initiative for Chronic Obstructive Lung Disease
- Foundations, e.g., The National Osteoporosis Foundation
- Advocacy groups, e.g., The American Diabetes Association
- Commercial organizations
- *Ad hoc* groups, often funded by the pharmaceutical industry, e.g., [treating pain due to peripheral neuropathy](#)
- Local healthcare systems, e.g., Tufts Medical Center
- Individual practices

## There are three general categories of guidelines:

### 1. Authority-based guidelines: BOGSATS



These are often called “consensus guidelines.” Frequently, this type of guidelines comes from professional groups or societies. These are developed by bringing together a group of experts who decide how to write the guideline. Evidence is

likely used by the group in some way, but there is no indication in the guideline regarding how they found, evaluated, and interpreted the evidence. As a result, the [quality of the recommendations](#) is low.

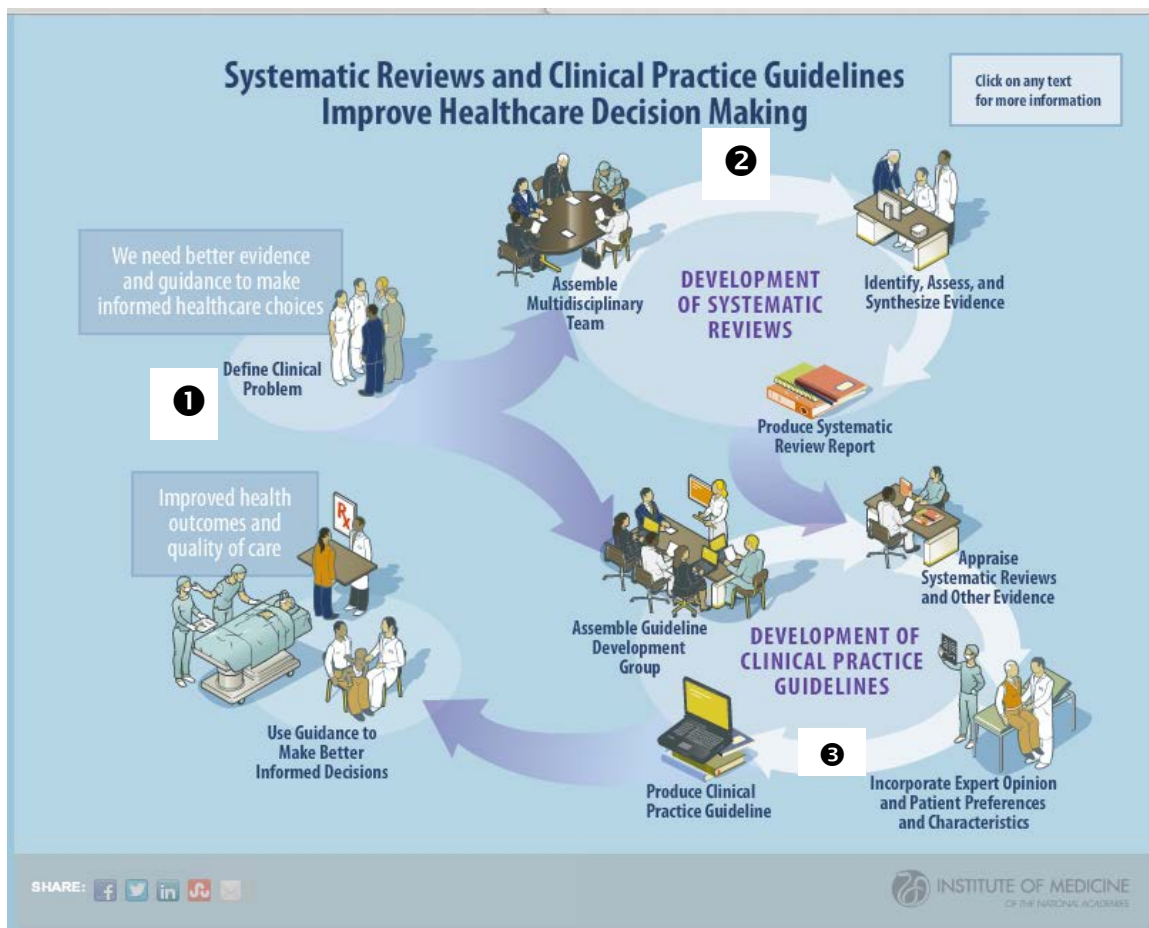
Less kindly, authority-based guidelines are called “**BOGSATS**,” which stands for a “bunch of old guys (or gals) sitting around a table.”

*“Consensus: A process by which a group agrees to something which no individual in the group believes to be appropriate.”*

-- Unknown

## 2. “Evidence-based” guidelines: Trust us, we have the evidence

The goal now is to have guidelines based on evidence, and most guidelines are now labeled as being “evidence-based.” These guidelines start by identifying the clinical questions to be addressed by the guideline (❶, below). A group assembles all the pertinent evidence available to answer the questions into a systematic review (❷). A guideline development group evaluates the information, weighs the benefits and risks of different interventions, and develops a practice guideline (❸)



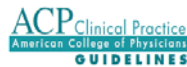
In contrast to evidence-linked guidelines, discussed below, evidence-based guidelines are limited by the lack of explicitness. They ask clinicians to be assured that they have used the evidence appropriately. Trust, though, is not part of EBM. Later in this section we will discuss how to identify guidelines written at this level.

### 3. Evidence-linked guidelines: Here is what we think and here is the evidence

Evidence-linked guidelines can be identified by a methods section explaining how the evidence was determined, how it was graded, and how it was used to inform the decision-making behind the guidelines. Frequently this type of guideline has two documents: one document contains the recommendations that are linked to a second document that reports the evidence supporting these guidelines.

#### Recommendations

##### CLINICAL GUIDELINES



### Screening for Osteoporosis in Men: A Clinical Practice Guideline from the American College of Physicians

Amir Qaseem, MD, PhD, MHA; Vincenza Snow, MD; Paul Shekelle, MD, PhD; Robert Hopkins Jr., MD; Mary Ann Forciea, MD; and Douglas K. Owens, MD, MS, for the Clinical Efficacy Assessment Subcommittee of the American College of Physicians\*

**Description:** The American College of Physicians developed this guideline to present the available evidence on risk factors and screening tests for osteoporosis in men.

**Methods:** Published literature on this topic was identified by using MEDLINE (1990 to July 2007). Reference mining was done on the retrieved articles, references of previous reviews, and solicited articles from experts. The inclusion criteria for the studies were measuring risk factors for low bone mineral density or osteoporotic fracture in men or comparing 2 different methods of assessment for the presence of osteoporosis in men. This guideline grades the evidence and recommendations by using the American College of Physicians' clinical practice guidelines grading system.

**Recommendation 1:** The American College of Physicians recommends that clinicians periodically perform individualized assessment

of risk factors for osteoporosis in older men (Grade: strong recommendation; moderate-quality evidence).

**Recommendation 2:** The American College of Physicians recommends that clinicians obtain dual-energy x-ray absorptiometry for men who are at increased risk for osteoporosis and are candidates for drug therapy (Grade: strong recommendation; moderate-quality evidence).

**Recommendation 3:** The American College of Physicians recommends further research to evaluate osteoporosis screening tests in men.

Ann Intern Med. 2008;148:680-684.  
For author affiliations, see end of text.

www.acp.org

#### Evidence Report

##### Annals of Internal Medicine

##### CLINICAL GUIDELINES

### Screening for Osteoporosis in Men: A Systematic Review for an American College of Physicians Guideline

Hui Liu, MD, MBA, MPH; Neil M. Paige, MD, MSHS; Caroline L. Goldzweig, MD, MSHS; Elaine Wong, MD; Annie Zhou, MS; Marika J. Suttorp, MS; Brett Munjas, BA; Eric Orwoll, MD; and Paul Shekelle, MD, PhD

**Background:** Screening for low bone mineral density (BMD) by dual-energy x-ray absorptiometry (DXA) is the primary way to identify asymptomatic men who might benefit from osteoporosis treatment. Identifying men at risk for low BMD and fracture can help clinicians determine which men should be tested.

**Purpose:** To identify which asymptomatic men should receive DXA BMD testing, this systematic review evaluates 1) risk factors for osteoporotic fracture in men that may be mediated through low BMD and 2) the performance of non-DXA tests in identifying men with low BMD.

**Data Sources:** Studies identified through the MEDLINE database (1990 to July 2007).

**Study Selection:** Articles that assessed risk factors for osteoporotic fracture in men or evaluated a non-DXA screening test against a gold standard of DXA.

**Data Extraction:** Researchers performed independent dual abstractions for each article, determined performance characteristics of screening tests, and assessed the quality of included articles.

**Data Synthesis:** A published meta-analysis of 167 studies evaluating risk factors for low BMD-related fracture in men and women

found high-risk factors to be increased age (>70 years), low body weight (body mass index <20 to 25 kg/m<sup>2</sup>), weight loss (>10%), physical inactivity, prolonged corticosteroid use, and previous osteoporotic fracture. An additional 102 studies assessing 15 other proposed risk factors were reviewed; most had insufficient evidence in men to draw conclusions. Twenty diagnostic study articles were reviewed. At a T-score threshold of -1.0, calcaneal ultrasonography had a sensitivity of 75% and specificity of 66% for identifying DXA-determined osteoporosis (DXA T-score, -2.5). At a risk score threshold of -1, the Osteoporosis Self-Assessment Screening Tool had a sensitivity of 81% and specificity of 68% to identify DXA-determined osteoporosis.

**Limitation:** Data on other screening tests, including radiography, and bone geometry variables, were sparse.

**Conclusion:** Key risk factors for low BMD-mediated fracture include increased age, low body weight, weight loss, physical inactivity, prolonged corticosteroid use, previous osteoporotic fracture, and androgen deprivation therapy. Non-DXA tests either are too insensitive or have insufficient data to reach conclusions.

Ann Intern Med. 2008;148:685-701.  
For author affiliations, see end of text.

www.annals.org

## What are (potential) problems with guidelines?

*Fundamentally, however, it is now nearly impossible for all stakeholders to be confident of [clinical practice guideline] quality.*

-- [Institute of Medicine](#), 2011, p. 191

Guidelines can differ markedly in their quality. [A recent study](#) showed that less than ½ of the 130 guidelines they evaluated met more than 50% of the requirements for good guideline development. Problems with guidelines include:

- **Lack of transparency.** There should be a clear path from the recommendations back to the evidence.
- **Guidelines that don't guide.** Sometimes, recommendations are written in a general form that makes it hard for clinicians to understand exactly what is being recommended. For example, a recent depression guideline suggested that, "it is not unreasonable to try exercise in certain patients."
- **Lack of systematic review of the literature.** As a result, guidelines sometimes suffer from selective citing of research that supports a certain position, ignoring research that doesn't support this position.

### 3.6 Identifying the Evidence

To identify the relevant evidence, a team of methodologists and medical librarians at the Oregon Health & Science University Evidence-based Practice Center conducted literature searches of Medline, the Cochrane Library, and the Database of Abstracts of Reviews of Effects. For each article, the team conducted a search for systematic reviews and another for original studies encompassing the main populations and interventions for that article. These searches included studies indexed from week 1, January 2005, forward because AT8 searches were carried out up to that date (search strategies are available on request). Many articles supplemented these searches with more-focused searches addressing specific clinical questions. When clinical questions had not been covered in AT8, searches commenced at a date relevant to each intervention.

Titles and abstracts retrieved from bibliographic database searches generally were screened in duplicate, and full-text articles were retrieved for further review. Consensus on whether individual studies fulfilled inclusion criteria was achieved for each study between two reviewers. If consensus could not be achieved, the topic editor and other topic panelists were brought into the discussion. Deputy editors reviewed lists of included studies from the database searches in order to identify any potentially missed studies. Additional studies identified were then retrieved for further evaluation.

Topic panels also searched the same bibliographic databases for systematic reviews addressing each PICO question. The quality of reviews was assessed using principles embodied in prior instruments addressing methodologic quality of systematic reviews,<sup>9,10</sup> and wherever possible, current high-quality systematic reviews were used as the source of summary estimates. Reviews were also used to identify additional studies to complement the database searches.

Outline of a thorough search of the literature, with a rigorous method of deciding which research should be included.

- **Conflicts of interest.** These can be financial or intellectual.
  - A financial conflict (duality) of interest can occur when a guideline developer receives research funding by a manufacturer of a product affected by the guideline, or if he or she is a paid speaker or consultant to a manufacturer or entity. A financial conflict of interest can also occur if a recommendation would result in financial benefit or harm to an organization or its members since a professional society has a primary responsibility to promote its members' interests.<sup>1</sup> These financial arrangements may result in a conscious or unconscious bias.



[Practice Guidelines: More Harm than Good?](#) (watch only about the first three minutes)

- Intellectual conflicts occur when one's research, profession, or experiences unduly influence one's ability to evaluate other types of evidence. It produces a sort of "tunnel vision." See [déformation professionnelle](#).



[KevinMD.com](#): *Conflicts of Interest Don't Always Involve Money*

## How do these problems affect decision-making?

Clinicians are often confused when guidelines markedly differ from one another. For example, regarding the treatment of newborn infants with elevated bilirubin levels, the [U.S. Preventive Services Task Force](#) concludes:

*"that the evidence is insufficient to recommend screening infants for hyperbilirubinemia to prevent chronic bilirubin encephalopathy."*

The [American Academy of Pediatrics statement](#) was published *the same month and year*, had the opposite recommendation:

*"...We recommend universal predischARGE bilirubin screening, which helps to assess the risk of subsequent severe hyperbilirubinemia. These recommendations represent a consensus of expert opinion based on the available evidence, and they are supported by several independent reviewers."*

---

<sup>1</sup> Quanstrum KH, Hayward RA. Lessons from the mammography wars. N Engl J Med 2010; 363:1076-1079.



*Think about factors that might explain why these guidelines present such different guidance.*



# Finding Clinical Guidelines: The National Guideline Clearinghouse

<http://ngc.gov/>

The National Guideline Clearinghouse provides physicians and other health professionals, health care providers, health plans, integrated delivery systems, purchasers, and others an accessible mechanism for obtaining objective, detailed information on clinical practice guidelines and to further their dissemination, implementation, and use. The Clearinghouse actively searches for guidelines as well as accepts submissions from guideline development groups.

Guidelines can be searched by keyword, MeSH, or by topic or organization.

The screenshot shows the homepage of the National Guideline Clearinghouse (NGC). At the top, there is a header for the AHRQ Agency for Healthcare Research and Quality, with the tagline 'Advancing Excellence in Health Care'. Below this, there is a navigation bar with links to 'Visit: National Quality Measures Clearinghouse | Health Care Innovations Exchange | AHRQ Home' and a 'Sign In' button. The main content area features a search bar with a 'Search' button, and links to 'Search Tips', 'Advanced Search', and 'About Search'. A sidebar on the left contains a 'Guidelines' section with a 'Browse' dropdown menu. The 'Browse' menu is open, showing options: 'By Topic', 'By Organization', 'Guidelines in Progress', 'Guideline Index', 'Guideline Archive', and 'Related NQMC Measures'. A red arrow points from the 'By Topic' option to the 'Guidelines by Topic' section. The 'Guidelines by Topic' section contains a brief description of the MeSH hierarchy and a 'Create Topic E-mail Alerts' button. Below this, there are three columns of topic lists: 'Disease/Condition', 'Treatment/Intervention', and 'Health Services Administration'. Each column lists various topics with their respective counts. A second red arrow points from the 'About' link in the sidebar to the 'About' link in the 'Guidelines' section.

Users can create accounts and store guidelines or searches, as well as receive notification of updates.

There is some quality control regarding which guidelines are included in the Clearinghouse. The NGC has specific criteria for including guidelines (these are in the process of being updated):

- 1) The clinical practice guideline contains systematically developed statements, strategies, or information for specific clinical circumstances;
- 2) The clinical practice guideline was produced under the auspices of medical specialty associations; relevant professional societies, public or private organizations, government agencies at the Federal, State, or local level; or health care organizations or plans.
- 3) There is evidence of a systematic literature search.

- 4) The full text of the guideline is available and has been developed, reviewed, or revised in the past 5 years.

Guidelines are summarized using a template, and guidelines can be **compared** to determine differences in methodology and recommendations.

Home	Compare Guidelines >
Guidelines	Guideline Comparison
Expert Commentaries	
Guideline Syntheses	
Guideline Resources	
Annotated Bibliographies	
Compare Guidelines	
FAQ	
Submit Guidelines	
About	
My NGC	

<b>Guideline Title</b>	Hypertension in pregnancy. The management of hypertensive disorders during pregnancy.	Treatment of the hypertensive disorders of pregnancy. In: Diagnosis, evaluation and management of the hypertensive disorders of pregnancy.
<b>Date Released</b>	2010 Aug	2008 Mar
<b>Adaptation</b>	Not applicable: The guideline was not adapted from another source.	Not applicable: The guideline was not adapted from another source.
<b>Guideline Developer(s)</b>	National Collaborating Centre for Women's and Children's Health - National Government Agency [Non-U.S.]	Society of Obstetricians and Gynaecologists of Canada - Medical Specialty Society
<b>Source(s) of Funding</b>	National Institute for Health and Clinical Excellence (NICE)	Society of Obstetricians and Gynaecologists of Canada
<b>Composition of Group That Authored the Guideline</b>	<b>Guideline Development Group Members:</b> Chris Barry, Portfolio GP, Swindon, Wiltshire; Rachel Fielding, Deputy Director of Midwifery, North Bristol NHS Trust; Pauline Green, Consultant in Obstetrics and Gynaecology, Wirral University Teaching Hospital; Jane Hawdon, Consultant Neonatologist, University College London Hospitals NHS Foundation Trust; David James (from December 2009), Clinical Co-Director, National Collaborating Centre for Women's and Children's Health; Rajesh Khanna (until May 2009), Senior Research Fellow, National Collaborating Centre for Women's and Children's Health; Surbhi Malhotra, Consultant Anaesthetist, St Mary's Hospital, London; Fiona Milne, Patient and carer representative, Action on Pre-eclampsia; Susan Mitchinson, Patient and carer representative; Moira Mugglestone (from May 2009), Director of Guideline Development, National Collaborating Centre for Women's and Children's Health; Lynda Mulhair, Consultant Midwife, Guy's and St Thomas' NHS Foundation Trust, London; Leo Nherera, Health Economist, National Collaborating Centre for Women's and Children's Health; Adam North, Senior Paediatric Pharmacist, Royal Brompton and Harefield NHS Foundation Trust, London; Derek Tuffnell, Consultant Obstetrician, Bradford Royal Infirmary; James Walker, Professor in Obstetrics and Gynaecology, University of Leeds; Stephen Walkinshaw (Chair), Consultant in Maternal and Fetal Medicine, Liverpool Women's Hospital; David Williams, Consultant Obstetric Physician, University College London Hospitals NHS Foundation Trust, London.	<b>Principal Authors:</b> Laura A. Magee, MD, Vancouver BC; Michael Helewa, MD, Winnipeg MB; Jean-Marie Moutquin, MD, Sherbrooke QC; Peter von Dadelzen, MBChB, Vancouver BC. <b>Hypertension Guideline Committee Members:</b> Savannah Cardew, MD, Vancouver BC; Anne-Marie Côté, MD, Sherbrooke QC; Myrtle Joanne Douglas, MD, Vancouver BC; Tabassum Firoz, MD, Vancouver BC; Paul S. Gibson, MD, Calgary AB; Andrée Gruslin, MD, Ottawa ON; Ian Lange, MD, Calgary AB; Line Leduc, MD, Montreal QC; Alexander G. Logan, MD, Toronto ON; Evelyn Rey, MD, Montreal QC; Vyta Senikas, MD, Ottawa ON; Graeme N. Smith, MD, Kingston ON. <b>Strategic Training Initiative in Research in the Reproductive Health Sciences (STRIRHS)</b>

Higher quality guidelines can be found by using the **Advanced Search** to limit results to guidelines with more methodologic rigor.

Help | RSS | Subscribe to weekly e-mail | Site map | Contact us | For web developers

National Guideline Clearinghouse

acupuncture Search Search Tips Advanced Search About Search

Home Guidelines Expert Commentaries Guideline Syntheses Guideline Resources Annotated Bibliographies Compare Guidelines FAQ Submit Guidelines About My NGC

**Advanced Search**

Select up to 10 of the lists below to create a targeted search of summaries. As you make your selections, the number of available results will automatically update.

Keyword:

Search indexing keywords only: ☐ Disease or Condition ☐ Treatment or Intervention ☐ Health Services Administration

**Make selections in the checklists to target your search.**

Your selections below will yield **12 results**

[Clear all selections](#)

**Age of Target Population:**

☐ Adolescent (13 to 18 years)  
☐ Adult (19 to 44 years)  
☐ Aged (65 to 79 years)  
☐ Aged, 80 and over  
☐ Child (2 to 12 years)

**Clinical Specialty:**

☐ Allergy and Immunology  
☐ Anesthesiology  
☐ Cardiology  
☐ Chiropractic  
☐ Colon and Rectal Surgery

**Guideline Category:**

☐ Assessment of Therapeutic Effectiveness  
☐ Counseling  
☐ Diagnosis  
☐ Evaluation  
☐ Management

**Implementation Tools:**

☐ Clinical Guidelines

**Methods Used to Assess the Quality and Strength of the Evidence:**

☐ Expert Consensus  
☐ Expert Consensus (Committee)  
☐ Expert Consensus (Delphi Method)  
☐ Subjective Review  
☒ Weighting According to a Rating Scheme (Scheme)

**Methods Used to Formulate the Recommendations:**

☒ Balance Sheets  
☐ Expert Consensus  
☐ Expert Consensus (Consensus Development Conference)  
☐ Expert Consensus (Delphi)

**Only include guidelines that have/incorporate:**

☐ Patient Resources  
☐ A Formal Cost Analysis  
☐ An Implementation Plan  
☐ A Clinical Algorithm

**Organizations:**

☐ National Guideline Clearinghouse

# Evaluating Clinical Practice Guidelines

There are three characteristics of guidelines that must be evaluated:

1. The methodology used to identify and use the evidence.
2. The quality of the available evidence.
3. The presence of conflicts of interest.

You can evaluate them by following the questions below. There is a worksheet that can be used to keep track of the answers. Often these questions can be answered by finding the appropriate section of the National Guideline Clearinghouse template.

## Evaluating the validity of the process: Identifying flaws in the methodology used to identify and use the evidence

The first step is to evaluate the process the guideline developers used. Evidence-based guidelines will describe an explicit process to finding, evaluating, and interpreting the evidence.

### 1. Evidence: Are the guidelines linked to a separate, systematic review of the evidence?

Recommendations in a guideline should be based on a systematic review performed by methodologists, not researchers in the field or the guideline development group. The guideline statements [should be linked](#) to this systematic review rather than requiring readers to trust the developers. This linkage can be quickly identified by finding *evidence tables, balance sheets, or indicators of the strength of recommendations*. Simply quoting selected evidence is not enough; there should be an explicit link between each recommendation and the corresponding aspect of the systematic review.

CLINICAL GUIDELINES

Screening for Type 2 Diabetes Mellitus in Adults: Recommendations and Rationale

U.S. Preventive Services Task Force<sup>a</sup>

This document summarizes the current U.S. Preventive Services Task Force (USPSTF) recommendations on screening for type 2 diabetes in adults and updates for USPSTF recommendations on this topic. The complete USPSTF recommendation and rationale statement on this topic, which includes a brief review of the supporting evidence, is available through the USPSTF Web site ([www.preventiveservices.org](http://www.preventiveservices.org)) and the National Guideline Clearinghouse ([www.guideline.gov](http://www.guideline.gov)) and is posted through the Agency for Healthcare Research and Quality (AHRQ) Publications Clearinghouse ([pubs.aahr.org](http://pubs.aahr.org)) or email ([dissemination@ahrq.gov](mailto:dissemination@ahrq.gov)). The complete information on which this statement is based, including evidence tables and references, is available in the accompanying article in this issue and in the summary of the evidence and systematic evidence review on this topic on the Web site already mentioned. The summary of the evidence is also available in print through the AHRQ Publications Clearinghouse.

CLINICAL GUIDELINES

Screening Adults for Type 2 Diabetes: A Review of the Evidence for the U.S. Preventive Services Task Force

David N. Kohn, MD, MPH; William Bracken, MD, MPH; Jeff S. Raftery, MPH; Paul F. Pines, MD; Steven H. Woolf, MD, MPH; and Robert H. Loh, PhD

Table 1. Randomized, Controlled Trials of Tight Glycemic Control<sup>a</sup>

Study, Year (Reference)	Quality	Length of Study, y	Groups (Patients)	Glycemic Control
UGDP, 1971 (4b), 1978 (4b)	Fair	8.75	Placebo (n = 204) Insulin variable (n = 198) Conventional therapy (n = 1108) Intensive therapy (n = 2129)	22.8% increase vs. 13.5% decrease†
UNPDS 33, 1998 (4b)	Good	10	Conventional therapy (n = 1108) Intensive therapy (n = 2129)	7.9% vs. 7.0%‡
UNPDS 34, 1998 (4c)	Good	10.7	Conventional therapy, primarily diet (n = 417) Intensive therapy with metformin (n = 342) Conventional therapy (n = 50) Intensive therapy (n = 52)	8.0% vs. 7.4%§
Kumamoto, 1995 (5), 2000 (5)	Fair	8	Standard therapy (n = 78) Intensive therapy (n = 75)	9.4% vs. 7.1%§
VA CSDM, 1997 (52), 1996 (54), 1999 (56), 1999 (55), 2000 (57)	Fair	2.25	Standard therapy (n = 80) Intensive therapy (n = 80)	9.2% vs. 7.1%§
Steno 2, 1999 (53)	Fair	3.8	Standard therapy (n = 80) Intensive therapy (n = 80)	9.0% vs. 7.6%§

Example of an Evidence Table

NGC advanced search

Methods Used to Formulate the Recommendations:

- ☒ Balance Sheets
- ☒ Expert Consensus
- ☐ Expert Consensus (Consensus Development Conference)
- ☐ Expert Consensus (Delphi)

Balance sheet  
of benefits,  
risks and  
other aspects  
of decision-  
making

#### Key Action Statement 5B

Clinicians may offer tympanostomy tubes for recurrent AOM (3 episodes in 6 months or 4 episodes in

1 year, with 1 episode in the preceding 6 months). (Evidence Quality: Grade B, Rec. Strength: Option)

Strength of  
recommendation

#### Key Action Statement Profile: KAS 5B

Aggregate evidence quality	Grade B
Benefits	Decreased frequency of AOM. Ability to treat AOM with topical antibiotic therapy.
Risks, harms, cost	Risks of anesthesia or surgery. Cost. Scarring of TM, chronic perforation, cholesteatoma. Otorrhea.
Benefits-harms assessment	Equilibrium of benefit and harm.
Value judgments	None.
Intentional vagueness	Option based on limited evidence.
Role of patient preferences	Joint decision of parent and clinician.
Exclusions	Any contraindication to anesthesia and surgery.
Strength	<b>Option</b>

## 2. Chain of logic: Was an explicit, sensible, and transparent process used to weigh the risks and benefits associated with the recommendation?

Since all guidelines involve the application of values and judgments to the available evidence, this process should be described. The guideline should explain the target conditions, target populations, practice settings, and audience to which the recommendations apply. The outcomes should be specific – “clinically effective” is not a suitable outcome.

How the American College of Chest Physicians, sensibly, explicitly, and transparently selected outcomes and weighed risks and benefits.

#### 3.2 Patient-Important and Surrogate Outcomes

The outcomes for each clinical question were chosen by the topic editors and their panel members and were generally consistent across articles. Outcomes were restricted to those of importance to patients.<sup>4</sup> Panels considered the burden of anticoagulation therapy as a patient-important outcome when its consideration could tip the balance of benefits and harms. If we found no data for an outcome considered at the outset as patient-important, we nevertheless included uncertainty about the effects of the intervention on that outcome when weighing its benefits and harms.

In the absence of data on patient-important outcomes, surrogates could contribute to the estimation of the effect of an intervention on the outcomes that are important. Examples of surrogate outcomes include asymptomatic venous thrombosis detected by venographic or ultrasound surveillance and the percentage of time that an international normalized ratio was in therapeutic range (used as a surrogate for bleeding and thrombosis in the assessment of the effectiveness of centralized anticoagulation services).

The issue of asymptomatic thrombosis detected by venographic or ultrasound surveillance presented particular challenges to the articles addressing VTE prevention in orthopedic and nonorthopedic surgery populations, an article addressing nonsurgical prophylaxis, and an article addressing stroke prevention. We were explicit in considering the trade-offs between VTE and bleeding events. An article by Guyatt et al<sup>5</sup> in this supplement addresses these issues in some detail.

## Evaluating the Validity of the Supporting Research

### ***3. Is the supporting research primarily randomized controlled trials, systematic reviews, or meta-analyses?***

The next step is to review the evidence report to evaluate the quality of the research on which the guidelines are based. The guideline developers should describe their process for assigning levels of evidence. For example, some groups consider meta-analysis and systematic review to be lower quality evidence on the evidence hierarchy, a view not in line with current evidence-based medicine thinking. Guidelines based on low quality evidence should be clearly identifiable, either through levels of evidence or strength of recommendation ratings.

The American Psychiatric Association's evidence hierarchy, with meta-analysis near the bottom of the list.

- [A] *Randomized, double-blind clinical trial.* A study of an intervention in which subjects are prospectively followed over time; there are treatment and control groups; subjects are randomly assigned to the two groups; and both the subjects and the investigators are "blind" to the assignments.
- [A-] *Randomized clinical trial.* Same as above but not double blind.
- [B] *Clinical trial.* A prospective study in which an intervention is made and the results of that intervention are tracked longitudinally. Does not meet standards for a randomized clinical trial.
- [C] *Cohort or longitudinal study.* A study in which subjects are prospectively followed time without any specific intervention.
- [D] *Control study.* A study in which a group of patients and a group of control subject identified in the present and information about them is pursued retrospectively or forward in time.
- [E] *Review with secondary data analysis.* A structured analytic review of existing data, a meta-analysis or a decision analysis.
- [F] *Review.* A qualitative review and discussion of previously published literature with quantitative synthesis of the data.
- [G] *Other.* Opinion-like essays, case reports, and other reports not categorized above.

Examples of good grading taxonomies:

[United States Preventive Services Task Force Grades of Evidence](#)

[Strength of Recommendation Taxonomy \(SORT\)](#)

## Evidence of Bias

As previously mentioned, intellectual bias, profession-based tunnel vision, and financial conflicts of interest can influence recommendations; [especially those from professional advocacy groups](#). Conflicts of interest can have conscious or unconscious effects on the decision-making process. From [physicians](#) to [U.S. Supreme Court Justices](#), most people with conflicts of interest do not recognize the effect on their judgments. Merely declaring or acknowledging conflicts does not mitigate their effects. Both [social science](#) and [neuroscience](#) literature demonstrate that transparency alone is an insufficient solution because bias is often implicit and unintentional. Moreover, disclosure may not only normalize conflicts of interest but may also [worsen bias](#): *“disclosure can actually lead doctors to give biased advice, either through strategic exaggeration (whereby more biased advice is provided to counteract anticipated discounting), or “moral licensing” such that advice is legitimized because advisees “have been warned” (that is, caveat emptor or “buyer beware”).”*

### ***4. Financial conflict of interest: Are most of the guideline developers free of declared financial conflicts of interest, especially the committee's chair?***

Guidelines should contain a statement explaining the financial conflicts of interest of the guideline developers. The [Institute of Medicine](#) suggests that a minority of the development group should have conflicts and that the guideline chair should not have any financial ties.

Example of a guideline development group with most members declaring conflicts of interest, including the chair.

#### Financial Disclosures/Conflicts of Interest

##### Chair

Dr. Nelson B. Watts reports that he has received speaker honoraria from Amgen Inc., the International Society for Clinical Densitometry, Novartis AG, and Warner Chilcott; consultant honoraria from Amgen Inc., Arena Pharmaceuticals, Inc., Baxter, Intekrin Therapeutics Inc., Johnson & Johnson Services, Inc., Medpace, Merck & Co., Inc., NPS Pharmaceuticals, Orexigen Therapeutics, Inc., Pfizer Inc, sanofi-aventis U.S. LLC, Takeda, VIVUS, Inc., and Warner Chilcott; and research grant support through the University of Cincinnati [Osteoporosis Center](#) from Amgen Inc., Eli Lilly and Company, Merck & Co., Inc., Novartis AG, and NPS Pharmaceuticals.

##### Task Force Members

Dr. John P. Bilezikian reports that he has received speaker honoraria from Amgen Inc., Eli Lilly and Company, and Novartis AG and consultant honoraria from Amgen Inc., Eli Lilly and Company, Merck & Co., Inc., and Warner Chilcott.

Dr. Pauline M. Camacho reports that she has received research grant support for her role as principal investigator from Eli Lilly and Company and Procter & Gamble.

Dr. Susan L. Greenspan reports that she has received consultant honoraria from Amgen Inc. and Merck & Co., Inc. and research grant support for her role as principal investigator from the Alliance for Better Bone Health (Procter & Gamble/sanofi-aventis U.S. LLC), Eli Lilly and Company, and Warner Chilcott.

Dr. Steven T. Harris reports that he has received speaker honoraria from Amgen Inc., Genentech, Inc., Gilead, GlaxoSmithKline plc, F. Hoffmann-La Roche Ltd, Eli Lilly and Company, Novartis AG, Procter & Gamble, sanofi-aventis U.S. LLC, and Warner Chilcott and consultant honoraria from Amgen Inc., Gilead, GlaxoSmithKline plc, F. Hoffmann-La Roche Ltd, Eli Lilly and Company, Merck & Co., Inc., and Novartis AG.

Dr. Stephen F. Hodgson reports that he does not have any relevant financial relationships with any commercial interests.

Dr. Michael Kleerekoper reports that he has received speaker honoraria from Amgen Inc. and Eli Lilly and Company and speaker and consultant honoraria from F. Hoffmann-La Roche Ltd Diagnostics.

Dr. Marjorie M. Luckey reports that she has received speaker honoraria and consultant fees from Amgen Inc. and Novartis AG.

Dr. Michael R. McClung reports that he has received research grant support, consulting fees, and/or speakers' bureau honoraria from Amgen Inc., Eli Lilly and Company, Merck & Co., Inc., Novartis AG, and Warner Chilcott.

Dr. Rachel Pessah Pellack reports that she does not have any relevant financial relationships with any commercial interests.

Dr. Steven M. Petak reports that he has received speaker honoraria from Amgen Inc. and the International Society for Clinical Densitometry.

<http://ngc.gov/content.aspx?id=34968&search=osteoporosis>



**5. Intellectual conflict of interest:** Are the developers from a range of specialties, and include patients and other stakeholders?

Guidelines, especially from specialty societies are at risk of an intellectual conflict of interest when all group members are from the same profession and are members of the professional society promoting and developing the guidelines.

See: <http://tinyurl.com/d5gtznv>.

Even if the guidelines are not self-serving, there is a tunnel vision that occurs when all guideline developers are looking at evidence through the same prism of experience. Guideline development groups should have representatives from different specialties and, where possible, patients or consumer advocacy groups. Ideally, a methodologist should be part of the group as well.

Active researchers, while a source of cutting edge information, also risk being unduly influenced by their own research findings to the exclusion of other evidence. Evaluating for this bias can be accomplished by checking the author affiliations to determine whether they are active researchers, as well as by searching the guideline citations for publications by guideline development group members.

**6. Professional Conflict of Interest:** Were the guidelines approved or modified by a board, executive committee, or consensus committee of a professional society before their release?

Professional societies may not be able to provide unbiased guidance. As mentioned in a [recent editorial](#),

“... Although it is true that individual medical providers care deeply about their patients, the guild of health care professionals — including their specialty societies — has a primary responsibility to promote its members' interests. . . But it is a fool's dream to expect the guild of any service industry to harness its self-interest and to act according to beneficence alone — to compete on true value when the opportunity to inflate perceived value is readily available.”

Recommendations are suspect if they are written by a specialty group and the recommendations would benefit that specialty group's members. To quote an old adage, “Never ask a barber if you need a haircut.”

[Back to instructions](#)

## A Worksheet for Evaluating Practice Guidelines

### Determine *Relevance*

*Is this guideline worth considering? If the answer to any of these questions is No, it may be better to read other articles first.*

- A. Are the recommendations based on research on outcomes that patients would care about? (Be careful to avoid results that require extrapolation to an outcome that truly matters to patients)

Yes (go on )

No (**stop**)

- B. Does the guideline address problems common to your practice and suggest feasible interventions?

Yes (go on )

No (**stop**)

- C. Will this information, if true, require you to *change* your current practice?

Yes (go on )

No (**move on to the next guideline**)

[Back to instructions](#)

### Determine *Validity*:

#### D. Validity of the process:

- |   |     |    |
|---|-----|----|
| 1. <i>Evidence</i> : Are the guidelines linked to a separate, systematic review of the evidence?  | Yes | No |
| 2. <i>Chain of logic</i> : Was an explicit, sensible and transparent process used to weigh the risks and benefits associated with the recommendation? | Yes | No |

#### E. Validity of the supporting research

- |   |     |    |
|---|-----|----|
| 3. <i>Validity</i> : Is the supporting research primarily randomized controlled trials, systematic reviews, or meta-analyses? | Yes | No |
|---|-----|----|

#### F. Evidence of Bias

- |  |     |    |
|--|-----|----|
| 4. <i>Financial conflict of interest</i> : Are most of the guideline developers free of declared financial conflicts of interest, especially the committee chair?                              | Yes | No |
| 5. <i>Intellectual conflict of interest</i> : Are the developers from a range of specialties, and include patients and other stakeholders?   | Yes | No |
| 6. <i>Professional Conflict of Interest</i> : Were the guidelines approved or modified by a board, executive committee, or consensus committee of a professional society before their release? | Yes | No |

[Back to instructions](#)