

Competing interests: MH, JH, and DI have received fees for teaching acupuncture in courses of professional societies. All other authors: none declared.

Ethical approval: The protocol was approved by all relevant local ethics review boards.

- International Headache Society. ICH-10 guide for headaches. *Cephalalgia* 1997;17(suppl 19):1-82.
- Melchart D, Linde K, Fischer P, White A, Allais G, Vickers A, et al. Acupuncture for recurrent headaches: a systematic review of randomized controlled trials. *Cephalalgia* 1999;19:779-86.
- Melchart D, Linde K, Streng A, Reitmayr S, Hoppe A, Brinkhaus B, et al. Acupuncture randomized trials (ART) in patients with migraine or tension-type headache—design and protocols. *Forsch Komplementärmed Klass Naturheilkd* 2003;10:179-84.
- Pfaffenrath V, Brune K, Diener HC, Gerber WD, Göbel H. [Treatment of tension-type headache. Recommendation of the German Migraine and Headache Society.] *Schmerz* 1998;12:156-70.
- Nagel B, Gerbershagen HU, Lindena G, Pflingsten M. [Development and evaluation of the multidimensional German pain questionnaire.] *Schmerz* 2002;16:263-70.
- Vincent C. Credibility assessments in trials of acupuncture. *Compl Med Res* 1990;4:8-11.
- Karst M, Reinhard M, Thum P, Wiese B, Rollnik J, Fink M. Needle acupuncture in tension-type headache: a randomized, placebo-controlled study. *Cephalalgia* 2001;21:637-42.
- Tavola T, Gala C, Conte G, Invernizzi G. Traditional Chinese acupuncture in tension-type headache: a controlled study. *Pain* 1992;48:325-9.
- Carlsson J, Fahlcrantz A, Augustinsson LE. Muscle tenderness in tension headache treated with acupuncture or physiotherapy. *Cephalalgia* 1990;10:131-41.
- Hansen PE, Hansen JH. Acupuncture treatment of chronic tension headache—a controlled cross-over trial. *Cephalalgia* 1985;5:137-42.
- White AR, Resch KL, Chan JCK, Norris CD, Modi SK, Patel JN, et al. Acupuncture for episodic tension-type headache: a multicentre randomized controlled trial. *Cephalalgia* 2000;20:632-7.
- Xue CCL, Dong L, Polus B, English RA, Zheng Z, da Costa C, et al. Electroacupuncture for tension-type headache on distal acupoints only: a randomized, controlled cross-over trial. *Headache* 2004;44:333-41.
- Schoenen J. Guidelines for trials of drug treatments in tension-type headache. *Cephalalgia*. 1995;15:165-79.
- Kapchuk TJ. The placebo effect in alternative medicine: can the performance of a healing ritual have clinical significance? *Ann Intern Med* 2002;136:817-25.
- Bogaards MC, ter Kuile MM. Treatment of recurrent tension headache: a meta-analytic review. *Clin J Pain* 1994;10:174-90.
- Hrobjartsson A, Gøtzsche PC. Is the placebo powerless? An analysis of clinical trials comparing placebo with no treatment. *N Engl J Med* 2001;344:1594-602.

(Accepted 27 May 2005)

doi 10.1136/bmj.38512.405440.8F

## Variability in interpretation of chest radiographs among Russian clinicians and implications for screening programmes: observational study

Y Balabanova, R Coker, I Fedorin, S Zakharova, S Plavinskij, N Krukov, R Atun, F Drobniowski

### Abstract

**Objective** To determine variability in interpretation of chest radiographs among tuberculosis specialists, radiologists, and respiratory specialists.

**Design** Observational study.

**Setting** Tuberculosis and respiratory disease services, Samara region, Russian Federation.

**Participants** 101 clinicians involved in the diagnosis and management of pulmonary tuberculosis and respiratory diseases.

**Main outcome measures** Interobserver and intraobserver agreement on the interpretation of 50 digital chest radiographs, using a scale of poor to very good agreement ( $\kappa$  coefficient:  $\leq 0.20$  poor, 0.21-0.40 fair, 0.41-0.60 moderate, 0.61-0.80 good, and 0.81-1.00 very good).

**Results** Agreement on the presence or absence of an abnormality was fair only ( $\kappa = 0.380$ , 95% confidence interval 0.376 to 0.384), moderate for localisation of the abnormality (0.448, 0.444 to 0.452), and fair for a diagnosis of tuberculosis (0.387, 0.382 to 0.391). The highest levels of agreement were among radiologists. Level of experience (years of work in the specialty) influenced agreement on presence of abnormalities and cavities. Levels of intraobserver agreement were fair.

**Conclusions** Population screening for tuberculosis in Russia may be less than optimal owing to limited agreement on interpretation of chest radiographs, and may have implications for radiological screening programmes in other countries.

### Introduction

Radiological examination plays an important part in the diagnosis and monitoring of tuberculosis, particularly in the Russian Federation, yet the control of tuberculosis in Russia remains a challenge and an economic burden.<sup>1</sup> Case finding is based on fluorographic screening of the population, and diagnosis may be made on the basis of radiological abnormalities without bacteriological confirmation.<sup>2 3</sup>

We determined interobserver and intraobserver variability in interpretation of chest radiographs among a group of Russian clinicians from the disciplines of radiology, respiratory medicine, and tuberculosis.

### Methods

Our study was carried out in Samara, a Russian city about 1000 km south east of Moscow (population 1.2 million). We invited to take part in our study all specialists in tuberculosis, respiratory physicians from the two main local general hospitals, radiologists specialising in tuberculosis, and general radiologists.

The study material consisted of 50 high resolution digital posterior-anterior chest radiographs selected from the archives at King's College

Health Protection Agency National Mycobacterium Reference Unit, Department of Microbiology and Infection, Guy's, King's, and St Thomas' Medical School, London  
Y Balabanova  
research associate  
F Drobniowski  
professor

Samara Regional Tuberculosis Service, Samara Oblast Dispensary, Samara, Russia  
I Fedorin  
chief physician

Samara City Tuberculosis Service, Samara, Russia  
S Zakharova  
chief physician

College for Public Health, St Petersburg Academy for Postgraduate Sciences, Russia  
S Plavinskij  
professor

continued over

BMJ 2005;331:379-82



Table showing levels of experience is on bmj.com



This is an abridged version; the full version is on bmj.com

Department of  
Internal Medicine,  
Samara State  
Medical University,  
Russia

N Krukov  
*professor*

Department of  
Public Health and  
Policy, London  
School of Hygiene  
and Tropical  
Medicine, London

R Coker  
*senior lecturer*

Centre for Health  
Management,  
Tanaka Business  
School, Imperial  
College, London

R Atun  
*reader*

Correspondence to:  
F Drobniowski,  
Health Protection  
Agency National  
Mycobacterium  
Reference Unit,  
Institute of Cell and  
Molecular Sciences,  
Queen Mary's  
School of Medicine,  
London E1 2AT  
francis.drobniowski@  
kcl.ac.uk

Hospital, London. Thirty seven of the radiographs showed an abnormality and 13 were reported as normal.

To assess intraobserver agreement, we randomly repeated 10 pairs of radiographs in the set. The participants were familiar with the digital format, as both conventional film radiographs and digital radiographs are used in Russia. We converted these series of digital images into a high resolution slide presentation (Microsoft Powerpoint), which was reviewed by each participant in a darkened room during a single viewing session, independently from the other participants. Abnormal and normal images were randomly mixed and each participant reviewed them in the same order. Each image was reviewed for two minutes, a period determined from a pilot study and one which approximates to the time spent reviewing images in population screening. No clinical information was provided.

The participants recorded their interpretation of each radiograph on a structured questionnaire, using a five point scale<sup>4</sup>: 1 = normal; 2 = abnormal but not clinically important; 3 = not certain, warrants further diagnostic evaluation; 4 = abnormal diagnosis uncertain, warrants further diagnostic evaluation; and 5 = abnormal—diagnosis apparent but warrants appropriate clinical management.

The questionnaire also included categorical questions on the localisation of an abnormality and the presence of cavities. The participants were asked whether the radiographic findings were consistent with a diagnosis of tuberculosis and, if so, which form (according to the Russian classification system) and whether it was likely to be active. If observers suspected another diagnosis, they were asked to state the most likely diagnosis.

### Statistical analysis

We generated a receiver operating curve for three subgroups: tuberculosis specialists, general radiologists, and respiratory specialists. To decrease the subjectivity of a single expert decision and to limit bias due to differences in professional practice between UK and Russian clinicians, we took a reference standard from a majority decision of the specialist radiologists on the question of whether the findings were consistent with tuberculosis. We used this standard to compare the performance of the other participants with that of the specialist radiologists. The participants were blind to the reference standard.

To assess interobserver agreement among the participants and within the three subgroups, we used  $\kappa$  statistics for multiple observers ( $\kappa_m$ ), which is a measure of agreement beyond the level of agreement expected by chance alone. We also used  $\kappa$  statistics to measure intraobserver agreement between the two reports of radiographs that had been repeated. We adopted the guidelines for interpretation of  $\kappa$  coefficients from Altman: <0.20, poor agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80 good agreement; and 0.81–1.00 very good agreement<sup>5</sup>; we also calculated 95% confidence intervals.<sup>6</sup> By averaging the  $\kappa$  values of each lung zone, we calculated the mean interobserver and intraobserver  $\kappa$  statistics for localisation of an abnormality.

## Results

Overall, 61 of 80 (76%) tuberculosis specialists agreed to participate in our study, as did 15 of 18 (83%) respiratory specialists, all 12 specialist radiologists, and all 13 general radiologists (see table on [bmj.com](http://bmj.com)).

Overall agreement on the presence or absence of an abnormality on chest radiographs was fair only ( $\kappa_m=0.380$ ). Interobserver agreement was highest when we compared both normal findings and abnormal but not clinically important findings with the other responses (not certain, warrants further diagnostic evaluation; abnormal diagnosis uncertain, warrants further diagnostic evaluation; and abnormal—diagnosis apparent but warrants appropriate clinical management), although even then agreement was only moderate (0.479).

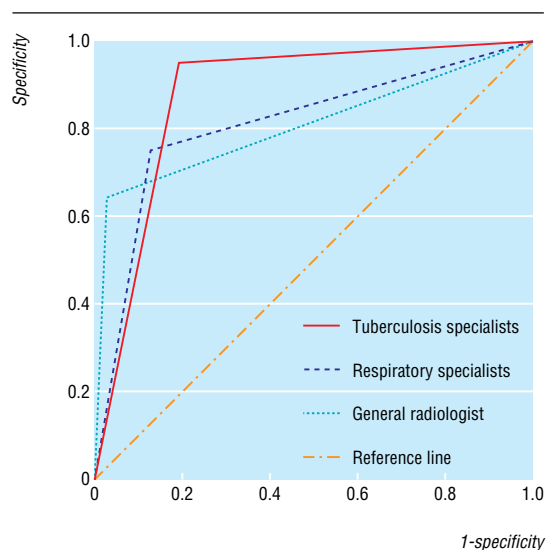
Agreement on localisation of abnormalities was moderate only (0.448; range 0.351–0.547) and agreement on determining a diagnosis of tuberculosis was fair only (0.387). For each of the 50 radiographs reviewed, tuberculosis was offered as a diagnosis by at least one participant. Agreement was highest among the radiologists, but still only moderate (0.448; see [bmj.com](http://bmj.com)).

When we combined normal findings with abnormal but not clinically important findings, the more experienced participants showed greater agreement on presence or absence of abnormalities (0.388, 95% confidence interval 0.383 to 0.393 v 0.355, 0.316 to 0.353) and detection of cavities (0.450, 0.444 to 0.456 v 0.354, 0.331 to 0.376), but not when we took all five responses into account. Level of experience made little difference to agreement on localisation of an abnormality and tuberculosis as a diagnosis.

We analysed agreement between the general radiologists and the specialist radiologists separately. The specialist radiologists showed higher levels of agreement on the four main questions posed: is a clinically important abnormality present, is a cavity present, are radiographic findings consistent with tuberculosis, and is the tuberculosis active? When comparing the majority decision among participating groups of doctors against the reference standard (figure), the areas under the receiver operating curve were: tuberculosis specialists, 0.88 (95% confidence interval 0.78 to 0.98); respiratory specialists, 0.81 (0.68 to 0.94); and general radiologists, 0.81 (0.67 to 0.95), illustrating no statistically significant variation in the performance of respiratory specialists or general radiologists from the reference opinion of whether the chest radiograph showed possible tuberculosis. The majority opinion of tuberculosis specialists was significantly closer to the opinion of the reference group than to the opinions of the other two groups.

Intraobserver agreement for all responses on repeated radiographs was fair to moderate only (see [bmj.com](http://bmj.com)). The radiologists had the highest levels of agreement (moderate to good;  $\kappa$  range 0.529–0.627).

Between doctors with less than five years' experience and those with five or more years' experience, the largest difference in intraobserver agreement was in assessing whether an abnormality



Receiver operating curve for the question "Are findings consistent with tuberculosis?"

was present (0.423 v 0.465). Experience did not seem to play an important part in interobserver agreement for presence of abnormalities (0.215 v 0.219), being low overall.

## Discussion

The interpretation of chest radiographs by Russian clinicians involved in the screening for and treatment of tuberculosis in Samara region is highly subjective and agreement was often low.

As Samara is a typical Russian city we believe that our findings may be generalisable throughout the Russian Federation. Levels of agreement were similar to other reports,<sup>7-14</sup> but these studies were not carried out in settings where mass population screening is routine practice, or in a post-Soviet environment.

In our study, professional experience had some influence on the ability to detect abnormalities, including cavities, which may be a prerequisite for any successful method for screening populations. In general, the effect of professional seniority on levels of diagnostic agreement was limited. Intraobserver agreement was not high overall, with radiologists showing most consistency in agreeing with their previous opinions on chest radiographs.

The effectiveness of the Russian model of screening (general population screening is mandatory and annual targets are set) depends on the tools used (radiology) and the interpretation of findings. Given the relatively low intraobserver and interobserver agreement we found in the interpretation of chest radiographs, the implications are profound as a significant number of the general population may be wrongly told that they have tuberculosis. This has repercussions both for the individual and for the tuberculosis programme, as considerable scarce resources may be used to exclude a diagnosis of tuberculosis. Under-capacity in microbiological laboratory services means that refuting a putative diagnosis of tuberculosis is prone to error. It seems likely that many people are potentially wrongly diagnosed as having

### What is already known on this topic

Radiological screening is an important tool in diagnosing tuberculosis

### What this study adds

The interpretation of chest radiographs among health professionals is limited

In the absence of symptoms, population screening programmes for tuberculosis have a low positive predictive value

tuberculosis. Moreover, many patients with tuberculosis may not be identified.

Our study was limited in two ways. Firstly, owing to the small number of chest radiographs selected for second review, the  $\kappa$  values for intraobserver agreement had wide confidence intervals. Secondly, the presence and type of abnormality was based on only one plain posterior-anterior chest radiograph. Therefore care should be taken in extrapolating results to routine clinical practice if clinical history, results of physical examination, and other radiographs are available.

Our study highlights the subjective nature of interpreting radiographs and the problems that such subjectivity has on management decisions for patients and on the effectiveness of an active post-Soviet screening programme. Clinical diagnoses and monitoring of progress should, whenever possible, be supported by the submission of pathological material for bacteriological or molecular examination.

We assessed the effectiveness of a screening programme provided by radiologists in Samara region. This region has an adult population of two million and an estimated prevalence of tuberculosis of 80 per 100 000. The positive predictive value (assuming sensitivity of 63% and specificity of 97%) is likely to be in the order of 1.7%; a maximum of 60 000 people without tuberculosis potentially would be subjected to unnecessary further investigations. The Russian government should be strongly advised to revise their screening policy and make better use of limited healthcare resources.

We thank Ekaterina Dodonova for statistical advice and R D Barker for help in selecting radiographs.

Contributors: See bmj.com

Funding: UK Department for International Development and a European Respiratory Society fellowship to YB.

Competing interests: None declared.

Ethical approval: Not required.

- 1 WHO, Division of emerging and other communicable diseases surveillance and control strategic plan 1996-2000. Geneva: World Health Organization.
- 2 Drobniewski F, Tayler E, Ignatenko N, Paul J, Connolly M, Nye P, et al. Tuberculosis in Siberia 2. Diagnosis, chemoprophylaxis and treatment. *Tuber Lung Dis* 1996;77:297-301.
- 3 Coker RJ, Dimitrova B, Drobniewski F, Samyshkin Y, Balabanova Y, Kuznetsov S, et al. Tuberculosis control in Samara Oblast, Russia: institutional and regulatory environment. *Int J Tuberc Lung Dis* 2003;7:920-32.
- 4 Potchen EJ, Cooper TG, Sierra AE, Aben GR, Potchen MJ, Potter MG, et al. Measuring performance in chest radiography. *Radiology* 2000;217:456-9.
- 5 Altman D. *Practical statistics for medical research*. London: Chapman and Hall, 1991.
- 6 Brealey S, Scally AJ. Bias in plain film reading performance studies. *Br J Radiol* 2001;74:307-16.

- 7 Albaum MN, Hill LC, Murphy M, Li YH, Fuhrman CR, Britton CA, et al. Interobserver reliability of the chest radiograph in community-acquired pneumonia. PORT Investigators. *Chest* 1996;110:343-50.
- 8 Alland D, Kalkut GE, Moss AR, McAdam RA, Hahn JA, Bosworth W, et al. Transmission of tuberculosis in New York city. An analysis by DNA fingerprinting and conventional epidemiologic methods. *N Engl J Med* 1994;330:1710-6.
- 9 Coblenz CL, Babcock CJ, Alton D, Riley BJ, Norman G. Observer variation in detecting the radiologic features associated with bronchiolitis. *Invest Radiol* 1991;26:115-8.
- 10 Du Toit G, Swingle G, Itoni K. Observer variation in detecting lymphadenopathy on chest radiography. *Int J Tuberc Lung Dis* 2002;6:814-7.
- 11 Tudor GR, Finlay DB. Error review: can this improve reporting performance? *Clin Radiol* 2001;56:751-4.
- 12 Kwong JS, Carignan S, Kang EY, Muller NL, FitzGerald JM. Miliary tuberculosis. Diagnostic accuracy of chest radiography. *Chest* 1996;110:339-42.
- 13 Dhingra R, Finlay DB, Robinson GD, Liddicoat AJ. Assessment of agreement between general practitioners and radiologists as to whether a radiation exposure is justified. *Br J Radiol* 2002;75:136-9.
- 14 Zitting AJ. Prevalence of radiographic small lung opacities and pleural abnormalities in a representative adult population sample. *Chest* 1995;107:126-31.

(Accepted 22 June 2005)

## Temporal trends in multiple births after in vitro fertilisation in Sweden, 1982-2001: a register study

Bengt Källén, Orvar Finnström, Karl Gösta Nygren, Petra Otterblad Olausson

Centre for Reproduction Epidemiology, Tornblad Institute, University of Lund, Biskopsgatan 7, S-223 62 Lund, Sweden

Bengt Källén  
professor

Department of Paediatrics, University Hospital, S-581 85 Linköping, Sweden

Orvar Finnström  
professor

IVF Clinic, Sophiahemmet, S-114 86 Stockholm, Sweden  
Karl Gösta Nygren  
associate professor

Centre for Epidemiology, National Board of Health and Welfare, Stockholm, Sweden  
Petra Otterblad Olausson  
head of register

Correspondence to: Bengt Källén  
embryol@embryol.lu.se

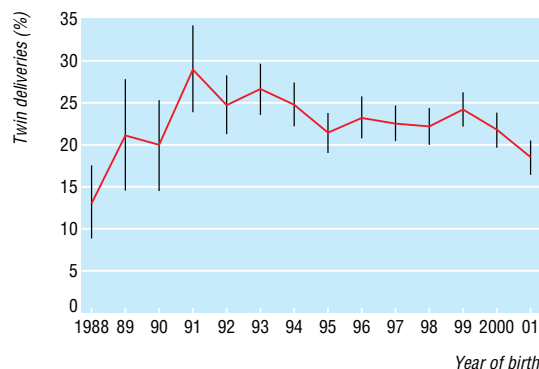
BMJ 2005;331:382-3

During the 20 years since the first child was born in Sweden after in vitro fertilisation, clinical policy has changed. During the early 1990s, the clinics performing in vitro fertilisation and the National Board of Health and Welfare agreed to reduce the number of embryos transferred to the uterus during in vitro fertilisation from three to two. Here, we describe the changes in multiple birth rates that have resulted from this change and their impact on the rate of preterm birth. In future papers we intend to describe maternal and child characteristics in greater detail.

### Participants, methods, and results

In Sweden 17 hospitals or private clinics perform in vitro fertilisation. The National Board of Health and Welfare requested information from these laboratories on all women who had undergone in vitro fertilisation and who had had a baby or whose pregnancy outcome was not known. By linking these data with the Swedish medical birth register, we identified infants born from 1982 to 2001.<sup>1 2</sup> This register covers nearly all deliveries in Sweden and is based on copies of medical documents from the antenatal care, the delivery, and the paediatric examination of newborn infants.

We compared the infants born after in vitro fertilisation and identified in the registry with all infants



Changes in percentage of deliveries after in vitro fertilisation that resulted in birth of twins, by year of birth (vertical bars are 95% confidence intervals)

### What is already known on this topic

Pregnancies occurring after in vitro fertilisation are characterised by high rates of multiple births and, as a consequence, high rates of preterm births

During the 1990s Swedish fertility clinics agreed to reduce from three to two the number of embryos transferred to the uterus during in vitro fertilisation

### What this study adds

The rate of multiple births after in vitro fertilisation increased to a maximum of 29% in 1991 but fell to 18.5% by 2001, resulting in a 70% reduction of preterm births

born in Sweden and recorded in that registry (2 039 943 during 1982-2001). We performed statistical analyses using the Mantel-Haenszel technique, with adjustment for various putative confounders. We expressed risks as odds ratios and calculated 95% confidence intervals with a test-based method (according to Miettinen).

We studied a total of 13 261 births after in vitro fertilisation, which resulted in 16 280 infants registered. The figure shows the changes in percentage of twin deliveries according to year of birth, with the first seven years added because of low numbers. The twinning rate increased to a maximum of 29% in 1991 and then decreased steadily to 18.5% in 2001. The annual number of triplet deliveries after in vitro fertilisation varied between 11 and 32 during 1992-6, and between three and seven after 1997. All seven sets of quadruplets were born before 1994.

The impact of the reduction of multiple births on the rate of preterm births is marked. In 1991, the crude odds ratio for preterm birth was 9.41 (95% confidence interval 7.58 to 11.67) and the odds ratio adjusted for maternal age (in 5 year groups, <20, 20-24, etc), parity

This article was posted on [bmj.com](http://bmj.com) on 13 May 2005: <http://bmj.com/cgi/doi/10.1136/bmj.38443.595046.E0>