BMJ open
accessible medical research

Empirical evidence that disease prevalence may affect the performance of diagnostic tests with an implicit threshold: a cross-sectional study

Brian H Willis

To cite: Willis BH. Empirical evidence that disease prevalence may affect the performance of diagnostic tests with an implicit threshold: a cross-sectional study. *BMJ Open* 2012;**2**: e000746. doi:10.1136/bmjopen-2011-000746

► Prepublication history and additional appendix for this paper are available online. To view these files please visit the journal online (http://bmjopen.bmj.com).

Received 9 December 2011
Accepted 20 December 2011

Department of Biostatistics, University of Manchester, Manchester, UK

Correspondence to
Dr Brian H Willis;
b.h.willis@doctors.org.uk

ABSTRACT

Objective: To investigate the effects that prevalence has on the diagnostic performance of junior doctors in interpreting x-rays.

Design: Two-armed cross-sectional design using systematic sampling.

Setting: Emergency department in the UK.

Participants: From a sample of 2593 patients (1434 men and 1159 women) taken from an unselected attending cohort between January and April 2002, 967 x-rays were analysed. The sex distribution was 558 men and 409 women, and the mean age of those receiving an x-ray was 34.6.

Interventions: The interpretation of x-rays by junior doctors after their triage into high- and low-prevalence populations by radiographers.

Main outcome measures: Sensitivity, specificity, likelihood ratios, diagnostic odds ratios and receiver operator characteristic curve.

Results: There were statistically significant differences in the performance characteristics of junior doctors when interpreting high-probability and low-probability x-rays. For the high- and low-probability populations, respectively, the sensitivities were 95.8% (95% CI 91.1% to 98.1%) and 78.3% (95% CI 65.7% to 87.2%) and the specificities were 56.0% (95% CI 41.9% to 69.2%) and 92.3% (95% CI 90.0% to 94.2%). Hierarchical logistic regression showed that the sensitivity did depend on the type of x-ray being interpreted but the diagnostic odds ratios did not vary significantly with prevalence, suggesting that doctors were changing their implicit threshold between the two populations along a common receiver operator characteristic curve.

Conclusions: This study provides evidence on how the prevalence may affect the performance of diagnostic tests with an implicit threshold and potentially includes the clinical history and examination. This has implications both for clinicians applying research findings to their practice and the design of future studies.

ARTICLE SUMMARY

Article focus

- The sensitivity, specificity and likelihood ratios of a diagnostic test are often assumed to remain constant even when the prevalence (pre-test probability) of disease changes.
- There is a lack of research in the literature on the effects that the prevalence has on the performance of diagnostic tests particularly those tests with implicit thresholds such as when doctors interpret x-rays.
- This study investigates the effects that prevalence has on the diagnostic performance of junior doctors in interpreting x-rays.

Key messages

- This study provides empirical evidence that the sensitivity, specificity and likelihood ratios may change with prevalence in diagnostic tests that require subjective interpretation, as in the case of junior doctors examining x-rays.
- The most plausible explanation for the effect seems to be doctors modifying their threshold for an abnormal diagnosis based on the results of previous tests.
- These results suggest that likelihood ratios and other test accuracy statistics derived from clinical studies have the potential to be misleading when applying them in practice.

Strengths and limitations of this study

- The study models a large data set collected from a real-life clinical setting and is representative of everyday clinical practice.
- The findings are likely to extend beyond the clinical tests analysed here.
- There is a potential for review bias owing to a lack of blinding between the test and reference standard.

INTRODUCTION

It is convenient in the assessment of diagnostic tests to divide the study population into two disjoint subpopulations consisting of those with disease and those without. Leading from this observation, many authors have asserted that performance characteristics, such as the sensitivity and specificity, which are derived from one or other of these

populations but not both, are independent of the prevalence of disease.[1–4]

This assertion has been questioned by some authors,[5–8] and circumstances in which a change in prevalence may affect the sensitivity and specificity have been described.[5 6] A Bayesian approach to diagnostic medicine relies on the reported values of the sensitivity, specificity and, hence, likelihood ratios being reproducible in practice. For the evidence-based clinician hoping to apply likelihood ratios reported in published studies to their practice, the potential for them to vary with the prior probability could have a profound effect on the reliability of applying diagnostic test research.

Despite its potential importance, currently there are few studies[6 8–13] which have considered the effects of prevalence on a test's performance. Ideally, this would be demonstrated by a study design, which has at least two arms, where prior testing has modified the pre-test probabilities so that they are different for each arm, before the test under investigation is applied. There are instances where this has been done for diagnostic tests, which have an explicit (fixed) threshold for a positive result.[14–16]

By contrast, in tests which have an implicit threshold, such as examining an MRI scan, the operator sets the level of the threshold, usually based on prior training and experience, but potentially in response to prior test results. This latter point seems to have received little attention in the literature. While there are examples of studies which have evaluated the performance of tests combined sequentially,[17–20] due to limitations in design,[17–20] the effect that each of the different outcomes of a test may have on the performance of a subsequent test has rarely been estimated.[21 22]

To help address this, the example used here investigates the effect the pre-test probability has on the performance of junior doctors in interpreting plain x-rays in an emergency department (ED) setting, before considering the implications for similar diagnostic tests. This study was part of a larger investigation, which has been published elsewhere.[23] Although the data were collected in 2002, the lack of research in this area and the continuing relevance of the findings underline the importance of research in this field. Note that pre-test probability and prevalence are used interchangeably.

METHOD

Between January and April 2002, systematic sampling was used to collect data on an unselected attending cohort of patients at the ED of the Horton Hospital in the UK.

All patients seeing an ED junior doctor underwent a clinical examination to determine whether treatment or further investigation was necessary. As part of their evaluation, some patients were required to have an x-ray, where the type of x-ray received depended on the results of the clinical examination.

Before the junior doctors viewed any x-rays, they were first interpreted then triaged, on the basis of their findings, by one of the departmental radiographers (radiologic technologists). Thus, those x-rays considered abnormal or 'high-probability' x-rays were marked with a red dot by the radiographer otherwise they were left unmarked. All the radiographers had received in-house training in interpreting x-rays.

Each x-ray was then interpreted by one of the ED junior doctors (each with similar training of at least 1-year experience post-qualification). All x-rays were then verified by a radiologist and this was the reference standard.

The data collected included the date, patient's age, x-ray type (eg, scaphoid), radiographer's triage result, junior doctor's diagnosis and reference diagnosis. The x-rays were classified by the part of the body irradiated (x-ray type), such as chest x-rays.

Features considered abnormal on an x-ray depended on the x-ray type and included fractures (skeletal x-rays), cardiomegaly (chest x-rays) and dilated bowel (abdominal x-rays), thus covering a range of target disorders and are detailed elsewhere.[23] In the high-probability (red dot) x-rays, the prevalence of abnormal findings was 77% compared with 13% in the 'low-probability' x-rays. Although the junior doctors were aware that a red dot indicated a higher probability of an abnormality, they were not aware of how high this probability was.

Statistical analysis

Two by two tables were derived for each of the high- and low-prevalence subpopulations. The sensitivity, specificity, likelihood ratios and diagnostic odds ratios (DOR) were used for comparison and a receiver operator characteristic (ROC) curve was constructed.[1 24–27]

While performance statistics, such as the sensitivity and specificity, could be calculated from pooling the data across all the junior doctors, this does not take into account variation in the performances between junior doctors. Furthermore, it does not allow for the effects of the x-ray type on the performances of individual junior doctors.

Hence, a hierarchical logistic regression model[27 28] was constructed to study the effects of different covariates on the dependent variables, logit sensitivity and logit specificity. Junior doctors were included in the model as a random effect, and covariates on prevalence, x-ray types and broader groupings of x-ray types were also included.

As any effects of prevalence on performance may be explained by differences in performance across different x-ray types, the interaction between the prevalence and x-ray type was evaluated. Cross-level interactions between explanatory variables were also investigated by allowing the slope to vary across individual doctors. Models were compared using the log likelihood ratio test statistic (LRT), which has an asymptotic χ^2 distribution with degrees of freedom (df).[27 28] All analyses were completed using the statistical software R, and statistical significance was set at $p<0.05$. A full description of the model may be found in the online appendix.

Table 1 Contingency tables showing the summary totals in each of the cells after pooling all the junior doctors

Pooled data for the junior doctors						
	High-prevalence population (77%)			Low-prevalence population (13%)		
	Reference standard			Reference standard		
	Positive	Negative		Positive	Negative	
Doctor's diagnosis						
Positive	159	22	181	72	50	122
Negative	10	28	38	24	602	626
Totals	169	50	219	96	652	748

Note x-rays in the high/low-prevalence population were those interpreted by the radiographer as having a high/low probability of an abnormal feature. The true prevalence is determined by the reference standard.

The type of x-ray a patient receives is, in part, indicative of their morbidity. Thus, the distributions of x-ray types were inspected to give some indication on whether the mix of patients (or patient-mix) varied between the high- and low-prevalence populations. If some x-ray types are more difficult to interpret than others (such as abdominal x-rays compared with tibial x-rays), then differences in the relative proportions of these may explain differences in the performance characteristics.

RESULTS

There were 1053 x-rays interpreted by 26 ED junior doctors following triage by a radiographer. Eighty-six were excluded due to incomplete information on the radiographers' triage result (28), junior doctors' diagnosis (10) and reference diagnosis (48). The remaining 967 x-rays are analysed in table 1.

The striking feature of these results is the change in sensitivity, specificity and positive likelihood ratio between the low- and high-prevalence populations (table 2). The differences are statistically significant and provide evidence against the null hypothesis that the performance characteristics of junior doctors at interpreting x-rays do not vary with prevalence.

In contrast, the DOR for each of the high- and low-prevalence populations were not statistically significantly different, being very close to each other at 37.3 (95% CI 3.6 to 101.3) and 36.1 (95% CI 21.0 to 62.3), respectively. This is consistent with the null hypothesis that the DOR is constant, which has a bearing on the shape of the ROC curve. A common DOR generates a symmetrical ROC curve,[24–26] and observing how closely the points are to the curve, this informs a possible cause to the variation in the sensitivity and the specificity, namely a change in the implicit threshold for test positives as applied by the junior doctors (figure 1).

The x-ray distributions for each of the normal and abnormal populations are shown in figures 2 and 3. On inspection, the distributions are broadly similar for the high- and low-prevalence populations in each case, with only chest x-rays being an outlier in figure 3. This would suggest that any differences in performance between the high- and low-prevalence populations are unlikely to be due to differences in the relative proportions of x-ray type.

Table 2 Summary performance estimates given for the independent significant covariate, prevalence. Also given are the estimates of sensitivity for each level of the covariate x-ray group, which was significant for the dependent variable logit (sensitivity)

Model estimates of performance characteristics in significant covariates		
	High prevalence	Low prevalence
Sensitivity (%)		
Soft tissue x-rays	93.7 (79.5 to 98.3)	68.3 (44.3 to 85.3)
Appendicular x-rays	97.3 (93.3 to 99.0)	84.0 (70.3 to 92.2)
Axial skeletal x-rays	58.6 (17.3 to 90.5)	17.0 (2.4 to 63.1)
Summary	95.8 (91.1 to 98.1)	78.3 (65.7 to 87.2)
Specificity (%)		
Summary	56.0 (41.9 to 69.2)	92.3 (90.0 to 94.2)
Positive likelihood ratio		
Summary	2.2 (1.6 to 3.0)	10.2 (7.6 to 13.8)
Negative likelihood ratio		
Summary	0.07 (0.03 to 0.17)	0.23 (0.14 to 0.38)
Diagnostic Odds ratio		
Summary	37.3 (3.6 to 101.3)	36.1 (21.0 to 62.3)

All estimates are derived from the hierarchical regression model and take into account variation in performance between individual doctors and different x-ray groups. The covariate x-ray group has three levels: soft tissue (chest and abdominal x-rays), appendicular (limbs, hands and feet) and axial (skull, spine and sacrum). Interaction terms were not significant. 95% CIs are shown in brackets.

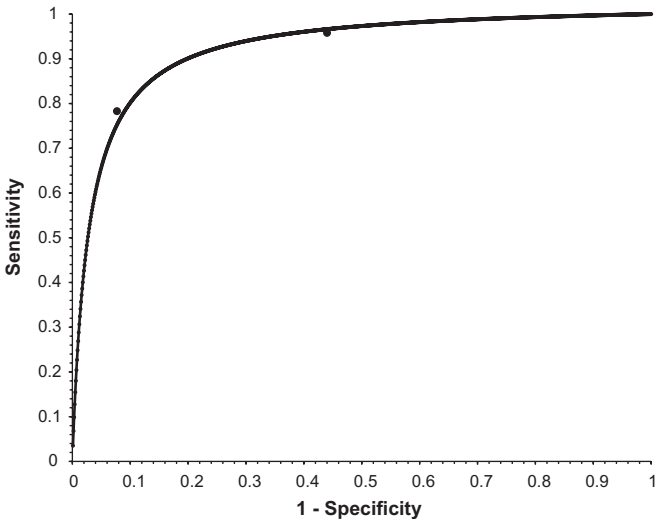


Figure 1 Symmetrical receiver operator characteristic curve (weighted mean diagnostic odds ratios (DOR)) for the average junior doctor. Weighted mean DOR (36.4) was derived from weighting model estimates of DORs for high-prevalence population (37.3) and low-prevalence population (36.1). Point estimates of sensitivity and 1− specificity for both populations are also given.

The effects that the change in x-ray distributions may have on performance between the two subpopulations was modelled using hierarchical logistic regression. Unsurprisingly, prevalence was a significant covariate for each of the dependent variables logit sensitivity (LRT=20.6, df=1, $p \sim 10^{-5}$) and logit specificity (LRT=42.8, df=1, $p \sim 10^{-11}$).

In contrast, x-ray type was not a significant covariate for either logit sensitivity (LRT=34.4, df=24, p=0.078) or logit specificity (LRT=23.3, df=33, p=0.89). Owing to the number of levels to the factor x-ray type (34), this could be due to insufficient data. Therefore, x-ray types were grouped into three broad mutually exclusive groups: skeletal x-rays that were subdivided anatomically into appendicular (limbs, hands and feet) and axial (skull, facial and spine)[29] and soft tissue x-rays (chest and abdomen). The x-ray group was a significant independent covariate for logit sensitivity (LRT=10.88, df=2, p=0.0043) but not for logit specificity (LRT=2.74, df=2, p=0.26) (table 2). However, interaction terms between prevalence and x-ray group and across levels between x-ray groups and junior doctors were not significant for either dependent variable.

As chest x-rays were a potential outlier (figure 3), a sensitivity analysis was performed to investigate the effects of this category on the statistical significance of covariates, by including and excluding this category from the model. No significant effects were found.

DISCUSSION

This study demonstrated statistically significant differences in the sensitivities, specificities and positive likelihood ratios between the high- and low-prevalence populations (table 2), providing evidence that the diagnostic performance of junior doctors in interpreting x-rays does vary with prevalence. There was evidence that the sensitivity depended on the x-rays being interpreted, and although such dependence could not be demonstrated for individual x-ray types (due to sample size), it was demonstrated for broader categories of x-rays. Since the x-ray type is an indicator of the type of target disorder and therefore patient, this implies that the diagnostic performance does depend to some degree on both the type of x-ray being interpreted and the target disorder being sought.

However, this was an independent effect: analysis of the interaction between prevalence and x-ray group was not significant. The effect of the junior doctors' performance varying with prevalence occurred irrespective of the type of x-ray being interpreted or target disorder being sought. Although performance was evaluated over different types of x-rays and multiple target disorders, these findings suggest the potential of observing such

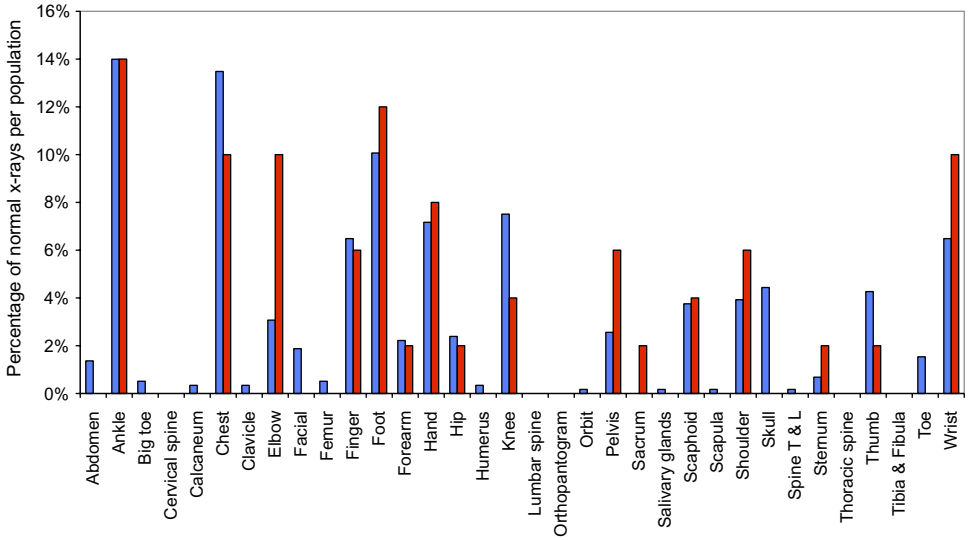


Figure 2 Distribution of x-rays with a normal diagnosis in the two populations: high prevalence (red) and low prevalence (blue). Shown are the percentage of normal x-rays in each population (high or low prevalence), which are of a particular type. For example, 10% of x-rays diagnosed normal in the high-prevalence (red) population were of elbows. Differences in the distributions between the high- and low-prevalence populations could potentially account for differences in the specificity between the respective populations. Note that the normal diagnosis refers to the reference standard diagnosis. T & L, thoracic and lumbar.

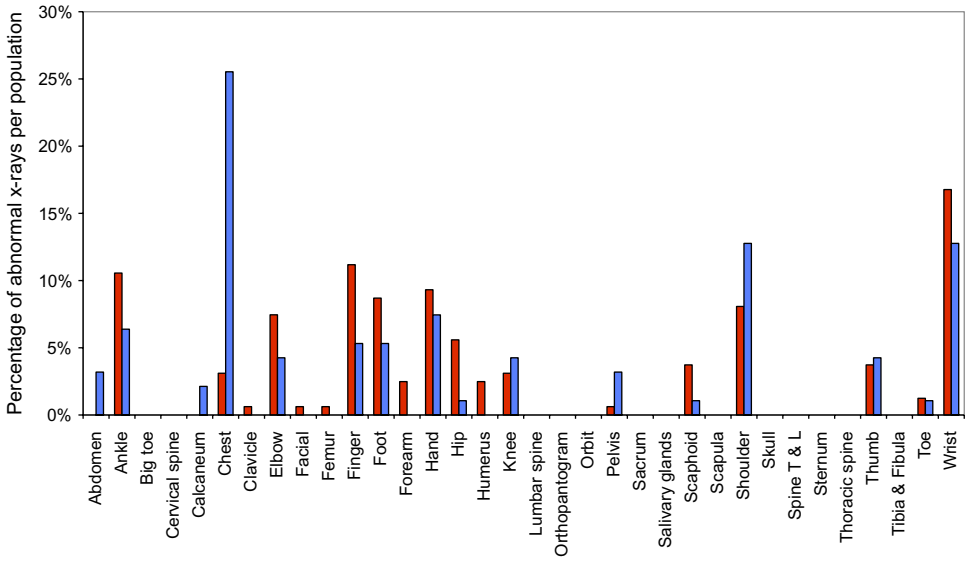


Figure 3 Distribution of x-rays with an abnormal diagnosis in the two populations: high prevalence (red) and low prevalence (blue). Shown are the percentage of abnormal x-rays in each population (high or low prevalence), which are of a particular type. For example, 10.5% of x-rays diagnosed abnormal in the high-prevalence (red) population were of ankles. Differences in the distributions between the high- and low-prevalence populations could potentially account for differences in the sensitivity between the respective populations. Note that the abnormal diagnosis refers to the reference standard diagnosis. T & L, thoracic and lumbar.

prevalence effects when only a single target condition is of interest.

Given there was insufficient evidence to reject the hypothesis of a common DOR and considering the closeness of the (sensitivity, 1− specificity) pairs to the ROC curve (figure 1), the doctors' performance seems to change along a single symmetrical ROC curve. This is consistent with the junior doctors changing their implicit threshold for an abnormal diagnosis on the basis of the radiographer's triage result. This does seem plausible when it is noted that the doctors both had knowledge of the previous test's (radiographer's triage) results and could change their subjective threshold for a positive test result on the basis of this information. It is possible that this latter point was amplified by the relative lack of experience in the participating doctors, with more experienced clinicians being expected to exhibit such threshold effects to a lesser degree. Clearly, this study does not answer this latter point.

Other explanations are still possible: the ROC curve may not be unique or symmetrical[24–26]; differences in the patient spectrum between the two populations may affect the different performance characteristics observed.[30 31] For instance, the initial triaging by the radiographers into high- and low-probability x-rays is dependent on their ability to spot abnormal features. Severe cases, where the abnormal features are more striking, are more easily identified and more likely to be allocated to the high-prevalence (probability) population. Thus, the differences in performance between the high and low populations could be a reflection of differences in severity.

This cannot be discounted and almost certainly explains part of the effect of the prevalence on performance. However, the circumstantial evidence in favour of junior doctors changing their implicit threshold seems more extensive, suggesting that this is likely to be the most important factor.

The question that is raised by this example is whether the effects observed may be generalised to other diagnostic tests? An example where these may occur is in the dynamic process of taking a clinical history or examining a patient, where information from previous tests such as the response to a particular question is available to inform future tests. During this process, the clinician may adjust their threshold for a positive result on the basis of the previous test results. The strength of expectation generated by the previous test results is likely to play a role in how far the clinician adjusts this threshold. Thus, a sequence of four positive responses to directed questions in a history might influence a clinician to lower their threshold for the next question, thereby increasing the sensitivity and decreasing the specificity, compared with if the four previous responses had been negative.

In this study, the test was evaluated in two separate subpopulations in which the main difference was the prevalence of abnormality. This has obvious advantages over two separate studies by controlling for a number of factors that may affect the test performance: the same junior doctors, same radiographers, same reference standard and similar patient-mixes.

Nonetheless, there are two principal limitations relating to the quality of the reference standard (a single radiologist's opinion) and a lack of blinding between the test and the reference standard, raising the potential for review bias.[32] It is difficult to gauge the effect a lower quality reference standard would have on performance estimates, but it is unlikely to have a differential bias between the high- and low-prevalence populations. Equally, the effects of review bias are likely to inflate estimates of the sensitivity and specificity in both the high- and low-prevalence populations and given it is differences between these performance statistics that are important to demonstrate the principle, inflated estimates in both subpopulations are less of a problem.

In the regression model, the sensitivity and specificity were treated as independent variables. A bivariate random effects model would maintain the association between the sensitivity and specificity, and individual patient data models have been suggested.[33] While such advanced approaches may augment the analysis, they would not change the broad findings demonstrated here.

In summary, the diagnostic performance of junior doctors in interpreting x-rays does vary with pre-test probability and this seems to be predominantly based on changing the implicit threshold in response to previous test results. Furthermore, it is unlikely that these findings are confined to the example analysed here. As such, it is an area deserving of further research to establish the extent by which it affects those tests in which there is a subjective element in the execution of the test.

Acknowledgements I would like to thank Dr Shyamaly Sur, MRCOG, in helping collect some of the original data for the study, Professor Chris Hyde, FFPHM, Professor Aneez Esmail, PhD, and Professor Graham Dunn, PhD, for comments on the manuscript. I had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Contributors BHW conceived the study, interpreted the data and wrote this manuscript.

Funding This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. The author was in receipt of a Medical Research Council fellowship during the conduct of this study.

Competing interests The author has completed the Unified Competing Interest form at http://www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declares that (1) BHW had support from a Medical Research Council fellowship during the conduct of this work, (2) has no relationship with any companies that might have an interest in the submitted work in the previous 3 years, (3) spouse, partner or children have no financial relationships that may be relevant to the submitted work and (4) have no non-financial interests that may be relevant to the submitted work.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data are available.

REFERENCES

1. Zhou X, Obuchowski NA, McClish DK. *Statistical Methods in Diagnostic Medicine*. New York: John Wiley and Sons, 2002:21.
2. Feinstein AR. *Clinical Epidemiology: The Architecture of Clinical Research*. Philadelphia: Saunders, 1985:434–9.
3. Kramer MS. *Clinical Epidemiology and Biostatistics: A Primer for Clinical Investigators and Decision-Makers*. Berlin: Springer, 1988:211–16.
4. Lalkhen AG, McCluskey A. Clinical tests: sensitivity and specificity, continuing education in anaesthesia. *Crit Care Pain* 2008;8:221–3.
5. Boyko EJ. Re: "Meta-analysis of Pap test accuracy" (Letter). *Am J Epidemiol* 1996;143:406–7.
6. Leeflang M, Bossuyt P, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. *J Clin Epidemiol* 2009;62:5–12.
7. Gianrossi R, Detrano R, Colombo A, *et al*. Cardiac fluoroscopy for the diagnosis of coronary artery disease: a meta analytic review. *Am Heart J* 1990;120:1179–88.
8. Lachs MS, Nachamkin I, Edelstein PH, *et al*. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. *Ann Intern Med* 1992;117:135–40.
9. Lee KH, Hashimoto SA, Hooge JP, *et al*. Magnetic resonance imaging of the head in the diagnosis of multiple sclerosis: a prospective 2-year follow-up with comparison of clinical evaluation, evoked potentials, oligoclonal banding, and CT. *Neurology* 1991;41:657–60.
10. O'Connor PW, Tansey CM, Detsky AS, *et al*. The effect of spectrum bias on the utility of magnetic resonance imaging and evoked potentials in the diagnosis of suspected multiple sclerosis. *Neurology* 1996;47:140–4.
11. Dagnelie CF, Bartelink ML, Van Der Graaf Y, *et al*. Towards a better diagnosis of throat infections (with group A beta-haemolytic streptococcus) in general practice. *Br J Gen Pract* 1998;48:959–62.
12. DiMattteo LA, Lowenstein SR, Brimhall B, *et al*. The relationship of pharyngitis and the sensitivity of a rapid antigen test: evidence of spectrum bias. *Ann Emerg Med* 2001;38:648–52.
13. Hall MC, Kieke B, Gonzales R, *et al*. Spectrum bias for a rapid antigen detection test for group A β-haemolytic streptococcus pharyngitis in a paediatric population. *Pediatrics* 2004;114:182–6.
14. Wells PS, Anderson DR, Rodger M, *et al*. Evaluation of D-dimer in the diagnosis of suspected deep-vein thrombosis. *N Engl J Med* 2003;349:1227–35.
15. Oudega R, Hoes AW, Moons KG. The Wells rule does not adequately rule out deep venous thrombosis in primary care patients. *Ann Intern Med* 2005;143:100–7.
16. Sultana RV, Zalstein S, Cameron P, *et al*. Dipstick urinalysis and the accuracy of the clinical diagnosis of urinary tract infection. *J Emerg Med* 2001;20:13–19.
17. Jahnke C, Bauer E, Hengge UR, *et al*. Accuracy of diagnosis of pediculosis capitis: visual inspection vs wet combing. *Arch Dermatol* 2009;145:309–13.
18. Hudelist G, Oberwinkler KH, Singer CF, *et al*. Combination of transvaginalsonography and clinical examination for preoperative diagnosis of pelvic endometriosis. *Hum Reprod* 2009;24:1018–24.
19. Taylor KJ, Merritt C, Piccoli C, *et al*. Ultrasound as a complement to mammography and breast examination to characterise breast masses. *Ultrasound Med Biol* 2002;28:19–26.
20. Aguilar C, del Villar V. Combined D-dimer and clinical probability are useful for exclusion of recurrent deep venous thrombosis. *Am J Hematol* 2007;82:41–4.
21. Sostman HD, Miniati M, Gottschalk A, *et al*. Sensitivity and specificity of perfusion scintigraphy combined with chest radiography for acute pulmonary embolism in PIOPED II. *J Nucl Med* 2008;49:1741–8.
22. Egglin TK, Feinstein AR. Context bias reference: a problem in diagnostic radiology. *JAMA* 1996;274:1752–5.
23. Willis BH, Sur SD. How good are emergency department senior house officers at interpreting x-rays following radiographers' triage? *Eur J Emerg Med* 2007;14:6–13.
24. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. *Stat Med* 1993;12:1293–316.
25. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Med Decis Making* 1993;13:313–21.
26. Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Stat Med* 2002;21:1237–56.
27. Collett D. *Modelling Binary Data*. 2nd edn. London: Chapman and Hall/CRC, 2003:269–301.
28. Gelman A, Hill J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, UK: Cambridge University Press, 2007:301–23.
29. Rogers AW. *Textbook of Anatomy*. London, UK: Churchill Livingstone, 1992:20–8.
30. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978;299:926–30.
31. Willis BH. Spectrum bias—why clinicians need to be cautious when applying diagnostic test studies. *Fam Pract* 2008;25:390–6.
32. Whiting P, Rutjes AWS, Dinnes J, *et al*. Development and validation of methods for assessing the quality of diagnostic accuracy studies. *Health Technol Assess* 2004;8:63.
33. Riley RD, Dodd SR, Craig JV, *et al*. Meta-analysis of diagnostic test studies using individual patient data and aggregate data. *Stat Med* 2008;27:6111–36.

STARD checklist for reporting of studies of diagnostic accuracy

(version January 2003)

Section and Topic	Item #		On page #
TITLE/ABSTRACT/ KEYWORDS	1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading 'sensitivity and specificity').	1
INTRODUCTION	2	State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.	4
METHODS			
Participants	3	The study population: The inclusion and exclusion criteria, setting and locations where data were collected.	5
	4	Participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?	5
	5	Participant sampling: Was the study population a consecutive series of participants defined by the selection criteria in item 3 and 4? If not, specify how participants were further selected.	5
	6	Data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?	5
Test methods	7	The reference standard and its rationale.	5 and 11
	8	Technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard.	5
	9	Definition of and rationale for the units, cut-offs and/or categories of the results of the index tests and the reference standard.	5 and reference [23]
	10	The number, training and expertise of the persons executing and reading the index tests and the reference standard.	5
	11	Whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers.	11
Statistical methods	12	Methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals).	6
	13	Methods for calculating test reproducibility, if done.	N/A
RESULTS			
Participants	14	When study was performed, including beginning and end dates of recruitment.	5
	15	Clinical and demographic characteristics of the study population (at least information on age, gender, spectrum of presenting symptoms).	6
	16	The number of participants satisfying the criteria for inclusion who did or did not undergo the index tests and/or the reference standard; describe why participants failed to undergo either test (a flow diagram is strongly recommended).	7 and reference [23]
Test results	17	Time-interval between the index tests and the reference standard, and any treatment administered in between.	N/A
	18	Distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.	8-9
	19	A cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard.	8-9 and reference [23]
	20	Any adverse events from performing the index tests or the reference standard.	N/A
Estimates	21	Estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals).	8
	22	How indeterminate results, missing data and outliers of the index tests were handled.	7
	23	Estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done.	8
	24	Estimates of test reproducibility, if done.	N/A
DISCUSSION	25	Discuss the clinical applicability of the study findings.	9-12

APPENDIX

Hierarchical logistic regression model

When the data are aggregated using simple pooling, variation in performance between the individual junior doctors is not accommodated. This may be modeled by allowing the effect that each junior doctor has on the overall performance, to be a random effect. In each of the models below, the j^{th} doctor modifies the aggregate performance by an amount δ_j where

$$\delta_j \sim N(0, \sigma_A^2), \text{ for some variance } \sigma_A^2$$

For each patient i the test response y_{ij} is a Bernoulli variable, such that

$$y_{ij} \sim Bernoulli(\pi_{ij})$$

where depending on whether the diseased group or the non-diseased group is analyzed π_{ij} refers to the sensitivity or the specificity of junior doctor j interpreting an x-ray on patient i. In this analysis the sensitivity and specificity were considered independent.

The sampling error ε_{ij} for each observation has a normal distribution given by

$$\varepsilon_{ij} \sim N(0, v^2) \text{ for some variance } v^2$$

In the base model (model 0), other than sampling error, the performance π_{ij} depends only on the individual doctors' performances. Covariates are then added incrementally to the base model and retained if their effect on the performance is significant.

In model 1, the additional effect of the binary variable, prevalence (high/low), is considered. The effect of the type of x-ray as an independent covariate is considered in model 2. For both the sensitivity and the specificity, the x-ray type was not a significant covariate (see below). However, $Xtype_i$ coded for 34 different types of x-

rays and there were only 219 observations in the diseased group and 748 in the non-diseased group. Hence, insignificant results could result from there being too few observations. To allow for this, the x-ray types were combined into three broad categories of soft tissue, axial skeleton and appendicular skeleton. The factor Xbd_i, codes for these 3 categories. As the incremental effect that prevalence has on the performance may vary across the different x-ray categories this is modeled by including the interaction between the prevalence and the *Xbd* (model 4). Finally, model 5 allows for the individual performance of each of the junior doctors to vary with x-ray category.

The fit of the models may be evaluated by comparing any of the goodness of fit statistics, Akaike information criterion (AIC), Bayesian information criterion or the Likelihood ratio test statistic (LRT), all of which are based on the log-likelihood function (LogLik) and have χ^2 distributions.

Thus, when comparing two models, the Likelihood ratio test statistic (= twice the difference of the LogLik) is compared to the χ^2 distribution with Δdf degrees of freedom. The results of comparisons of the different models on the sensitivity and the specificity as the performance statistics are given below.

Model 0: $logit(\pi_{ij}) = \alpha + \delta_j + \varepsilon_{ij}$

Model 1: $logit(\pi_{ij}) = \alpha + \beta_1 Prev_i + \delta_j + \varepsilon_{ij}$

Model 2: $logit(\pi_{ij}) = \alpha + \beta_1 Prev_i + \beta_2 Xtype_i + \delta_j + \varepsilon_{ij}$

Model 3: $log\ it(\pi_{ij}) = \alpha + \beta_1 Prev_i + \beta_2 Xbd_i + \delta_j + \varepsilon_{ij}$

Model 4: $log\ it(\pi_{ij}) = \alpha + \beta_1 Prev_i + \beta_2 Xbd_i + \beta_3 Xbd_i \times Prev_i + \delta_j + \varepsilon_{ij}$

Model 5: $log\ it(\pi_{ij}) = \alpha + \beta_1 Prev_i + (\delta_j + Xbd_i \times \gamma_j) + \varepsilon_{ij}$

where $\gamma_j \sim N(0, \sigma_B^2)$

Note in model 5, the junior doctor modifies the aggregate performance by an amount $\delta_j + Xbd_i \times \gamma_j$, the latter term varying with the category of x-ray (soft tissue, axial skeleton and appendicular skeleton).

1. Effects of covariates on the sensitivity.

Model	df	LogLik	LRT (χ^2)	Δdf	Pr(>Chi)
0	2	-100.7			
1	3	-90.4	20.603	1	$\sim 10^{-6}$ **
1	3	-90.4			
2	27	-73.2	34.410	24	0.07765
1	3	-90.4			
3	5	-84.9	10.876	2	0.0043**
3	5	-84.9			
4	7	-84.0	1.8154	2	0.4035
3	5	-84.9			
5	10	-84.6	0.6736	5	0.9844

** indicate significant with $p<0.05$.

Thus model 3 provided the best fit of the data when estimating the effects of different covariates on the sensitivity. Both the prevalence and the broad category of x-ray (*Xbd*) were significant.

Coefficients for model 3.

Covariate	Coefficient Estimate	Standard error
Intercept	0.7666	0.4976
Prev	1.9311	0.4963
Xbd		
Appendicular	0.8946	0.5418
Axial	-2.3516	1.1103

Model prediction (example)

The logit sensitivity in x-rays of the axial skeleton in the high prevalence population is given by

Logit(sensitivity) = 0.7666 + 1.9311+ -2.3516 = 0.3461

Hence the sensitivity =exp(0.461)/(1+exp(0.3461) = 58.57%

Note for soft tissue x-rays the coefficient =0

2. Effects of covariates on the specificity.

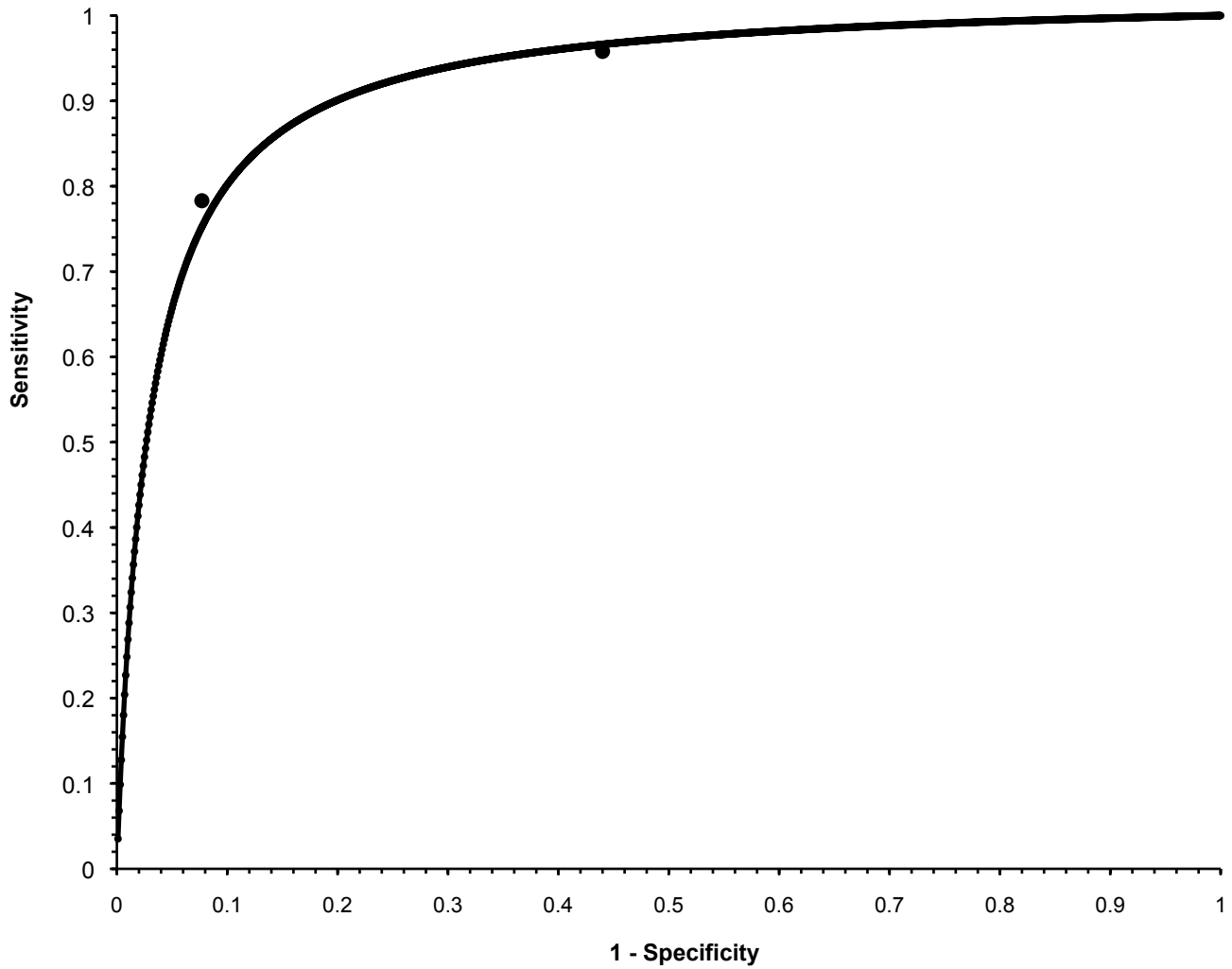
Model	df	LogLik	LRT (χ^2)	Δdf	Pr(>Chi)
0	2	-232.1			
1	3	-210.7	42.817	1	$\sim 10^{-11}$ **
1	3	-210.7			
2	36	-199.1	23.302	33	0.8946
1	3	-210.7			
3	5	-209.4	2.7372	2	0.2545
1	3	-210.7			
4	7	-206.5	8.3761	4	0.07873
1	3	-210.7			
5	8	-210.7	$\sim 10^{-08}$	5	1

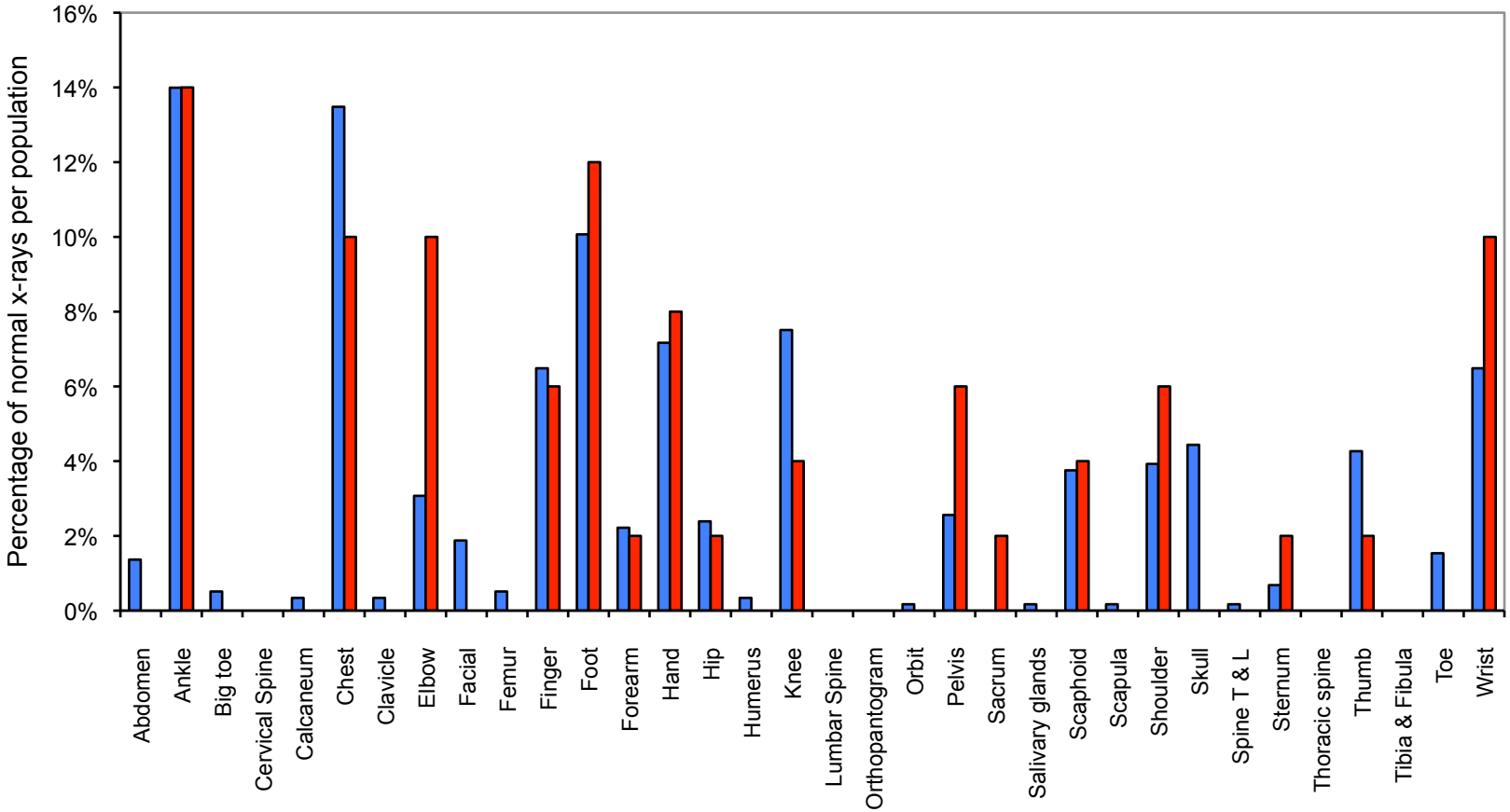
** indicate significant with $p<0.05$.

Thus model 1 provided the best fit of the data when estimating the effects of different covariates on the specificity. Only the prevalence was significant

Coefficients for model 1.

Covariate	Coefficient estimate	Standard error
Intercept	2.4882	0.1472
Prev	-2.2472	0.3207


Sensitivity
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1 - Specificity


Percentage of normal x-rays per population
16%
14%
12%
10%
8%
6%
4%
2%
0%
Abdomen
Ankle
Big toe
Cervical Spine
Calcaneum
Chest
Clavicle
Elbow
Facial
Femur
Finger
Foot
Forearm
Hand
Hip
Humerus
Knee
Lumbar Spine
Orthopantogram
Orbit
Pelvis
Sacrum
Salivary glands
Scaphoid
Scapula
Shoulder
Skull
Spine T & L
Sternum
Thoracic spine
Thumb
Tibia & Fibula
Toe
Wrist

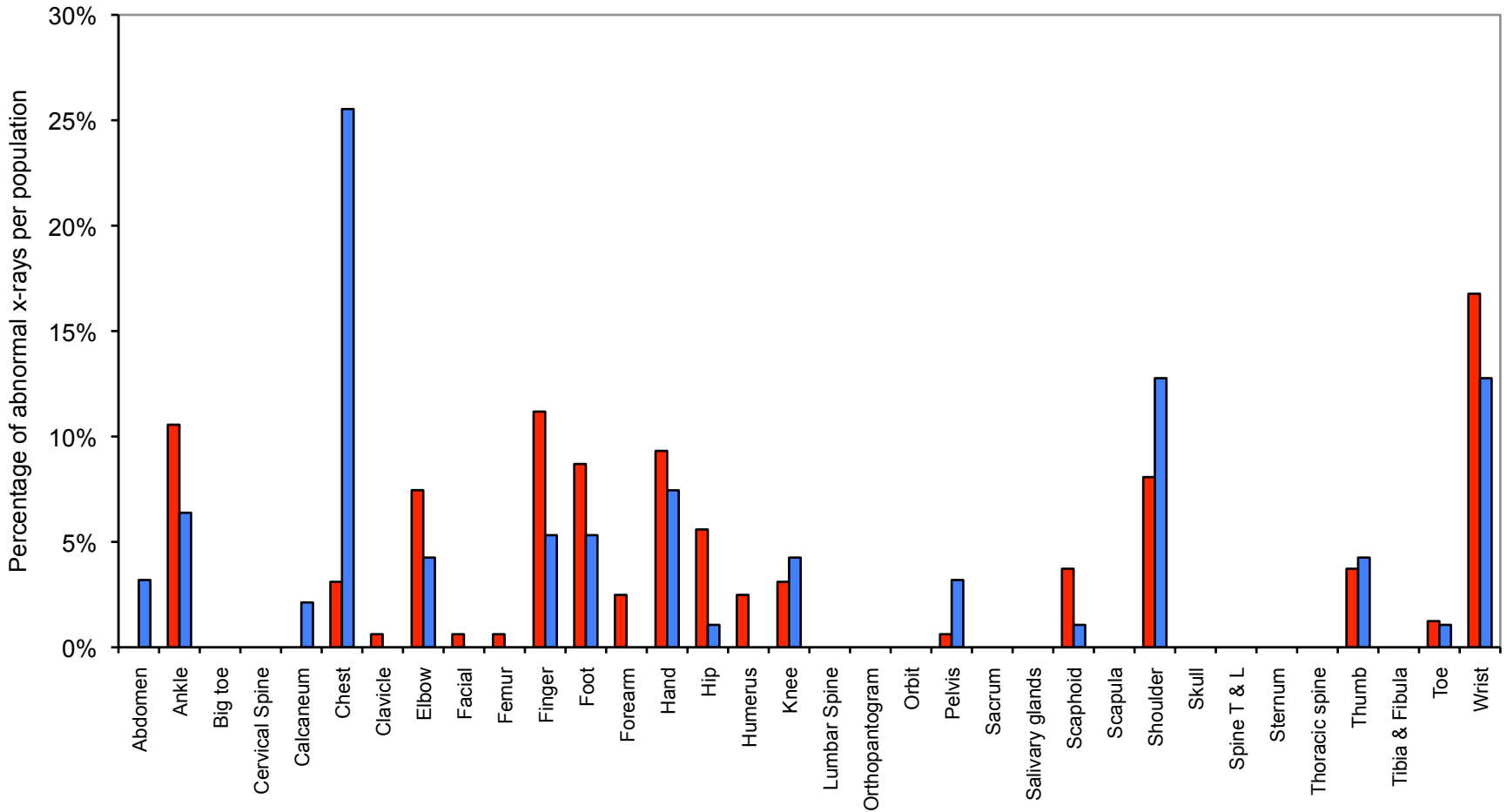

Percentage of abnormal x-rays per population
30%
25%
20%
15%
10%
5%
0%
Abdomen
Ankle
Big toe
Cervical Spine
Calcaneum
Chest
Clavicle
Elbow
Facial
Femur
Finger
Foot
Forearm
Hand
Hip
Humerus
Knee
Lumbar Spine
Orthopantogram
Orbit
Pelvis
Sacrum
Salivary glands
Scaphoid
Scapula
Shoulder
Skull
Spine T & L
Sternum
Thoracic spine
Thumb
Tibia & Fibula
Toe
Wrist

Figure Legends

Figure 1. Symmetrical ROC curve (weighted mean DOR) for the average junior doctor.
Weighted mean DOR (36.4) was derived from weighting model estimates of DORs for high prevalence population (37.3) and low prevalence population (36.1). Point estimates of sensitivity and 1- specificity for both populations are also given.

Figure 2. Distribution of x-rays with a normal diagnosis in the two populations: high prevalence (red), low prevalence (blue)
Shown are the percentage of normal x-rays in each population (high or low prevalence) which are of a particular type. For example 10% of x-rays diagnosed normal in the high prevalence (red) popualtion were of elbows. Differences in the distributions between the high and low prevalence populations could potentially account for differences in the **specificity** between the respective populations. Note, the normal diagnosis refers to the reference standard diagnosis. Abbreviation T & L = thoracic and lumbar.

Figure 3. Distribution of x-rays with an abnormal diagnosis in the two populations: high prevalence (red), low prevalence (blue)
Shown are the percentage of abnormal x-rays in each population (high or low prevalence) which are of a particular type. For example 10.5% of x-rays diagnosed abnormal in the high prevalence (red) popualtion were of ankles. Differences in the distributions between the high and low prevalence populations could potentially account for differences in the **sensitivity** between the respective populations. Note, the abnormal diagnosis refers to the reference standard diagnosis. Abbreviation T & L refers to thoracic and lumbar