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