Manuscript title

400 or more participants needed for stable contingency table estimates of clinical prediction rule performance

Authors

Peter Kent[1,2], Eleanor Boyle[2,3], Jennifer L Keating[4], Hanne B. Albert[5], Jan Hartvigsen[2,6].

[1]School of Physiotherapy and Exercise Science, Curtin University, Perth, Australia
[2]Clinical Biomechanics Research Unit, Department of Sports Science and Clinical Biomechanics, University of Southern Denmark, Odense, Denmark.
[3]Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada
[4]Department of Physiotherapy, Faculty of Medicine Nursing and Health Sciences, Monash University, Melbourne, Australia
[5]The Modic Clinic, Odense, Denmark
[6]Nordic Institute of Chiropractic and Clinical Biomechanics, Odense, Denmark

Correspondence:
Peter Kent
School of Physiotherapy and Exercise Science, Curtin University, Kent Street, Bently, Perth, Western Australia 6102, Australia
Phone: (+61) 8 9266 3629

Abstract

Objective: To quantify variability in the results of statistical analyses based on contingency tables and discuss the implications for the choice of sample size for studies that derive clinical prediction rules.

Study Design and Setting: An analysis of three pre-existing sets of large cohort data (n= 4,062 to 8,674) was performed. In each dataset, repeated random-sampling of various sample sizes, from n=100 up to n=2,000, was performed 100 times at each sample size and the variability in estimates of sensitivity, specificity, positive and negative likelihood ratios, post-test probabilities, odds ratios and risk/prevalence ratios, for each sample size was calculated.

Results: There were very wide, and statistically significant, differences in estimates derived from contingency tables from the same dataset when calculated in sample sizes below 400 people, and typically this variability stabilized in samples of 400 to 600 people. Although estimates of prevalence also varied significantly in samples below 600 people, that relationship only explains a small component of the variability in these statistical parameters.

Conclusion: To reduce sample-specific variability, contingency tables should consist of 400 participants or more when used to derive clinical prediction rules or test their performance.

Key words:
Clinical prediction rule, sample size, reproducibility of results, epidemiologic research design, predictive value of tests, decision support techniques.

What is new?

- There is a lack of information about appropriate sample sizes for studies that derive or test clinical prediction rules using contingency tables.
- We found very wide and statistically significant variability in estimates derived from contingency tables (sensitivity, specificity, positive and negative likelihood ratios, post-test probabilities, odds ratios and risk/prevalence ratios) when calculated in sample sizes of 100 or 200 people, which typically stabilized in samples of 400 to 600 or more people.
- Although estimates of prevalence also varied significantly in samples below 600 people, in less than 15% of occasions was there less variability in samples extracted with a fixed prevalence than in samples with a varying prevalence.
- Sample sizes in studies that derive prediction rules, or measure prediction rule performance, using contingency tables should consist of 400 participants or more.

Manuscript

1. Introduction

Clinical prediction rules are tools that define the relationship between multiple predictors (e.g. from an individual patient's history, physical examination, and/or test results), and likely diagnosis, prognosis or treatment response [1, 2]. They can be used to identify clinically relevant subgroups of patients. There is growing interest in clinical prediction rules, as seen in a recent study that identified more than 400 unique prediction rules across a range of health conditions that had been derived and published between 1965 and 2009, with the 80% of them published since the year 2000 [1].

Clinical prediction rules are derived from multivariable prediction models. The typical sequence is that candidate predictor variables are formed into prediction models using a variety of statistical methods, a final model is chosen based on its performance measures and then that prediction model is transformed into a prediction rule [3]. Although the derivation of the rule from the model can also occur using a variety of statistical approaches, they often involve the use of statistics based on dichotomization of data into 2 x 2 contingency tables.

The 2 x 2 contingency table represents a dichotomized predictor variable and dichotomized outcome variable (the numbers of people who have/do not have a clinical characteristic present who also have/do not have a particular outcome). Dichotomized predictor and outcome variables in a contingency table enable the estimation of sensitivity, specificity, likelihood ratios, odds ratios, risk or prevalence ratios, and pre-test and post-test probabilities. The clinical use of post-test probabilities is considered to be a high level application of evidence-based care for the diagnosis of, and treatment selection for, individual patients [3].

Contingency tables have been used at various stages in the derivation of prediction rules. For example in the case of the Flynn prediction for spinal manipulation in people with low back pain [4], univariate screening was initially used as a selection process to reduce the number of candidate variables, then continuous scale variables were dichotomized using the results of ROC analysis and their sensitivity, specificity and positive likelihood ratios were calculated from contingency tables for descriptive purposes, prior to the remaining candidate variables being entered into a logistic regression model. In other examples, contingency tables are used when identifying the number of items that need to be positive before a person is classified as 'rule positive', or in measuring prediction rule performance [3]. Even when a prediction rule is created using some form of sum score from a multivariable model such as linear regression, simple dichotomization of 'over or under' a threshold indicator and 'with or without' the outcome of interest is often used in the process of rule calibration or for describing model performance. Similarly, recursive partitioning approaches to studying diagnostic pathways, such as Classification and Regression Trees, are based on contingency tables and provide predicted probabilities of a diagnosis [5]. So the use of statistical estimates based on contingency tables commonly occurs at some stage in the creation of prediction rules, regardless of the overall method pathway used.

However, there is evidence that estimates based on contingency table statistics are highly variable *across* samples, due to variations in prevalence (selection bias) and disease severity (spectrum bias) [6-8]. These estimates can also be highly variable *within* samples, due the presence of other clinical characteristics that may reflect the existence of subgroups in the sample [9, 10]. While the influence of these attributes (selection bias, spectrum bias and the presence of clinical subgroups) on the variability in estimates based on contingency table statistics has been investigated [6-10], variability in estimates due to sample size has not been adequately researched.

Currently, the a priori estimation of adequate sample size is difficult in studies designed to derive clinical prediction rules, as (i) the performance characteristics of the rule cannot be known a priori, and (ii) the prevalence and severity of a particular health condition in a particular clinical setting may not be known. Sample sizes for studies that have derived musculoskeletal prediction rules have varied greatly, from 54 [11] to 8,924 [12], and are often less than 100 [4, 11, 13, 14].

Therefore, the aims of this study were to (i) quantify variability in the estimates of clinical prediction rule performance (sensitivity, specificity, positive and negative likelihood ratios, post-test probabilities, odds ratios and risk/prevalence ratios) that typically result from contingency tables of dichotomised predictors and outcomes, and (ii) discuss the implications of the results for sample sizes decisions in future studies.

2. Methods

2.1. Method summary

Three pre-existing sets of Danish cohort data were analyzed. The first dataset was of 4,062 patients with spine pain from which the diagnostic accuracy of a screening test for generalized hypermobility was assessed. The second dataset of 7,457 patients provided data for evaluating the association between fear of movement at a baseline consultation and scores for low back pain-related activity limitation 6-months later. The third dataset was 8,674 people in a twin registry that enabled assessment of the cross-sectional association between male sex and grip strength. Repeated sampling of various sample sizes was performed on each dataset and the variability in estimates of prediction rule performance at each sample size was calculated. From each dataset, we modelled single item predictors rather than multi-item prediction rules, but the statistical implications are identical, as scoring positive or negative on a multi-item prediction rule results in a dichotomous predictor variable.

2.2. Datasets

The first dataset consisted of 4,062 people with chronic spine pain that was assembled for a study of the diagnostic accuracy of elbow extension as a screening test for systemic hypermobility [15]. Briefly, from the records of a consecutive cohort of 17,117 back pain patients presenting to the Back Centre of Funen - a public

hospital department - from 1999 to 2008, all patients were identified who had been tested using the Beighton criteria for systemic hypermobility. The Beighton assessment includes nine physical tests for systemic joint hypermobility and these were tested in that clinical setting only with people suspected by their clinician of having hypermobility. The Albert et al. 2010 study used the cut point of 4 or more positive tests, as recommended by Grahame et al. [16], which has been shown to have good reproducibility (Kappa 0.80) when differentiating people diagnosed as having generalized joint hypermobility from those without this condition [17]. Using that criterion standard for systemic hypermobility, the accuracy of the individual Beighton test items for predicting the total Beighton sum scores was calculated to determine the most accurate single-item screening test. Beighton test items of the side of handedness (dominant side) were found to be more accurate than tests of either the right or left side, and extension of the dominant elbow >10 degrees was the most accurate single-item screening test with an overall accuracy of 93.9%. Therefore, in the current study, the effect of sample size on the diagnostic accuracy of >10 degrees extension in the dominant elbow (yes/no) as a screening test for systemic hypermobility (yes/no) was investigated.

The second dataset was of 7,457 chronic low back pain patients from the SpineData Registry at the Spine Centre of Southern Denmark [18]. The longitudinal association between fear of movement at a baseline consultation and high pain-related activity limitation 6-months later was assessed. The SpineData registry is a consecutive cohort of all consenting patients presenting to a regional spine center - a public hospital department. For the current study, all low back pain patients who had completed both the baseline and 6-month self-reported questionnaires were selected. Fear of movement at the baseline consultation was measured using two screening questions from the physical activity subscale of the Fear Avoidance Beliefs Questionnaire [19] that have been shown to have an overall accuracy of 93.2% in this setting, relative to the full subscale score [20]. People were categorized as positive on the dichotomized fear of movement variable if their average combined score on these two screening questions was equal to or above the validated cut point of 7.0 (0-10 scale) [20]. Pain-related activity limitation at 6-months follow-up was measured using the 23-item version of the Roland Morris Disability Questionnaire (RMDQ) [21, 22], with sum scores expressed as a proportional score [23]. The 6-month RMDQ scores were dichotomized using the mean of the sample (47.8) as the cut point, with scores above that threshold being classified as high activity limitation. Therefore, in the current study, the effect of sample size on the predictive accuracy of baseline fear of movement (yes/no) as a predictor of high pain-related activity limitation at 6-months follow-up (yes/no) was investigated.

The third dataset was of data from 8,674 people from a grip strength cohort within the Danish Twin Registry [24]. Briefly, this cohort included people from three national population-based surveys: the Longitudinal Study of Middle-Aged Twins, the Longitudinal Study of Aging Danish Twins, and the Danish 1905 Birth Cohort Study. In those studies, grip strength was recorded as the maximum grip of six attempts, three from each hand, measured using a Smedley dynamometer (TTM; Tokyo, Japan). The width of the dynamometer handle was adjusted to fit the participant's hand size,

and, during the measurement, the elbow was in 90 degrees of flexion, with the upper arm tight against the chest wall. Data were excluded from participants with less than three attempts or a difference of 20 kg or more between two attempts. On average, males have higher grip strength [24] than females. Therefore, in the current study, the cross-sectional association between male sex (yes/no) and grip strength (yes/no) above the sample mean (30.4 kg) was investigated [24]. Additional details of all the cohorts can be seen in Table 1.

Table 1: Sample characteristics

Elbow hyper-extension/systemic hypermobility	
Cohort sample size (n)	4,062
Mean (SD) age (years)	42.3 (13.9)
Sex (male)	43.0% (n=1,747)
Dominant arm (right)	90.4% (n=3,672)
Dominant elbow >10 degrees hyperextension	13.9% (n=565)
Generalised hypermobility (Beighton score of 4 or more)	14.6% (n=595)
Baseline fear of movement/6-months high activity limitation	
Cohort sample size (n)	7,457
Mean (SD) age (years)	55.8 (14.9)
Sex (male)	44.2% (n=3,296)
Median (IQR) episode duration (months)	10.1 (IQR 4.0 to 36.9)
Mean (SD) low back pain intensity - baseline (0 to 10 scale)	5.8 (2.4)
Mean (SD) activity limitation - baseline (0 to 100 scale)	62.1 (23.1)
Mean (SD) fear of movement - baseline (0 to 10 scale)	4.1 (3.1)
Baseline fear of movement score above 7.0 (0 to 10 scale)	24.0% (n=1,747)
Mean (SD) Low back pain intensity - 6-months follow-up (0 to 10 scale)	4.3 (2.7)
Mean (SD) Activity limitation - 6-months follow-up (0 to 100 scale)	47.8 (28.2)
Activity limitation above sample mean (47.8) at 6-months follow-up (0 to 100 scale)	48.5% (n=3,619)
Male sex/grip strength	
Cohort sample size (n)	8,674
Mean (SD) age (years)	71.9 (15.6)
Sex (male)	44.9% (n=3,895)
Mean (SD) grip strength (kg)	30.4 (13.6) full range (2 to 74)
Grip strength above sample mean (30.4kg)	43.4% (n=3,763)

SD= standard deviation, IQR=interquartile range, kg=kilogram

2.3. Statistics

In all datasets, 2 x 2 contingency table statistics were calculated for randomly selected samples of 100, 200, 400, 600, 800 and 1000 in the hypermobility dataset, samples of 100, 200, 400, 600, 800, 1000, 1200, 1400, 1600 and 1800 in the fear of

movement dataset, and samples of 100, 200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800 and 2000 in the grip strength dataset. Arbitrarily, the maximum size limit for these random samples was set at 25% of the total cohort size, to reduce the number of times that different samples contained some of the same individuals and thereby ensure variation across samples of the same size. The samples were randomly selected without replacement, which ensures that *within each sample*, the same person cannot be selected twice. However, repeating this random selection over and over means that the same person may have be selected *across multiple samples.*

For each sample size, random selection and the calculation of contingency table statistics were replicated 100 times. This simulated performing the study 100 times, and enabled an estimate of the variability attributable to sample-specific characteristics for each sample size. In addition, the same contingency table statistics were calculated using the complete cohort data, to represent our best estimate of the 'real' population parameter. Graphs were constructed to display the full range and median of point estimates for each contingency table statistic across sample sizes in each cohort. An overview of the study procedures is shown in Figure 1. The replications were limited to 100 times because as the number of repetitions increases, the frequency of more extreme estimates also increases.

Insert Figure 1 about here

2.3.2 Contingency table statistics

Eight statistical parameters were calculated from each randomly selected sample: sensitivity, specificity, positive and negative likelihood ratios, post-test probabilities for both a positive and negative test result, odds ratios and risk/prevalence ratios. These are briefly described here[20]:

- Sensitivity is the proportion of true positives that are correctly identified by the test and therefore, a sensitivity of 90% indicates that one in ten people with the outcome of interest are missed by the test (in this context, a prediction rule).
- Specificity is the proportion of true negatives that are correctly identified by the test and therefore, a specificity of 90% indicates that one in ten people who have do not have the outcome of interest are incorrectly classified by the test.
- Positive and negative likelihood ratios are measures that can be used with an estimate of the pre-test probability of having the outcome of interest (in this context, such as having a high activity limitation at 6-months follow-up) to calculate the post-test probability of that state. A likelihood ratio greater than 1 indicates that a positive test result is associated with the presence of the outcome, whereas a likelihood ratio less than 1 indicates that the positive test result is associated with the absence of the outcome [25]. The further likelihood ratios are from 1, the stronger the evidence for the presence or absence of the outcome.
- Post-test probabilities for a positive and negative test result are the probabilities of the outcome, depending on the test result.

- Odds ratios are the odds that a person with a positive test result will have the outcome, divided by the odds of having the outcome if the test result was negative.
- The risk ratio (also known as the relative risk) is the probability of a future event in a person with a positive test result, divided by the probability of the same event in a person with a negative test result. Whereas risk ratio is used in longitudinal data, when the same formula is used in cross-sectional data, the resultant parameter is called a prevalence ratio (because risk of an outcome is a longitudinal concept).

2.3.2 Exploratory statistics

To determine whether the variability in contingency table statistical estimates differed by sample size, pairwise comparisons of sequential sample sizes were performed using the STATA robvar command, which reports Levene's robust test statistic for the equality of variances between the samples.

As contingency table statistics partly reflect the prevalence of the condition in the sample, to understand whether the variability in estimates for these statistical parameters was related to variability in prevalence, we used the event-per-variable method to create new samples, at each sample size, in which the 'event' prevalence was fixed to that of the background prevalence in the whole cohort. In the event-per-variable method, the number of events (E) required at each sample size is determined by E= n x rate, where 'rate' is the proportion of events in the 'population' or whole data set[26]. For example, if the population event rate is 0.21, then 84 events would be needed for a sample containing 400 people (E= 400 x 0.21 = 84). As 2 x 2 contingency tables contain one test variable, these samples were extracted using a fixed prevalence determined by E x 1. Therefore, when using this method, all samples at all sample sizes had the same prevalence as the whole cohort from which they were drawn. We then performed pairwise comparisons (robvar), for each parameter at each sample size, to determine whether the variability in contingency table statistical estimates was different between samples with a fixed prevalence and those with a varying prevalence. Where there was a difference, we compared the SD of variance to identify the direction of the difference.

All statistical analyses were performed using STATA version 13.1 (StataCorp, College Station, Texas, USA) and Excel 2011 version 14.5 (Microsoft Corp, Redmond, Washington, USA). The flow chart and graphs were constructed using InDesign CS6 version 8.0 (Adobe Systems, San Jose, California, USA). We considered a result to be statistically significant if p= 0.05 or less.

2.4 Ethics

Under Danish law, the secondary analysis of such de-identified data does not require separate ethics approval (The Act on Processing of Personal Data, December 2012, Section 5.2; Act on Research Ethics Review of Health Research Projects, October 2013, Section 14.2).

3. Results

The results, graphed as the full range and median of point estimates for each contingency table statistic, across sample sizes, are shown in Figure 2 (sensitivity and specificity), Figure 3 (odds/risk/prevalence ratios) and in the online Appendix (likelihood ratios and post-test probabilities). The results were similar across all datasets. Some statistical parameters, such as likelihood/odds/risk/prevalence ratios generally showed much more sample-specific variability than other parameters, such as sensitivity, specificity and post-test probabilities.

Insert Figures 2 to 3 about here

Predictably, wide variability was seen in smaller samples, especially n=100 and n=200, and this diminished as sample sizes increased. In some instances, the size of the variability was so large as to indicate that results at that sample size were highly imprecise. For example, from samples of 100 people, estimates of the post-test probability after a positive test for hypermobility ranged from 8% to 95% (Figure 4). Similarly, from samples of 100 people, estimates of the odds ratio for grip strength ranged from 7.2 to 200.6 (Figure 5).

Across all the measures of prediction rule performance that were tested, the variability of estimates was significantly larger in the n=100 sample size than in the n=200 samples and in the n=200 than in the n=400 samples (Table 2). For most performance measures, the variability of estimates was also significantly larger in the n=400 sample size than in the n=600 samples. For some measures in some samples, this trend continued, but it was patchy and inconsistent. So the broad observation is that across cohorts, variability mostly stabilized when sample sizes were between 400 and 600.

Our results also contain two counter-intuitive findings in the hypermobility data (two out of 216 comparisons), where there was greater variability for the positive likelihood ratios and odds ratios at the n=200 sample size than at n=100 (sdtest $p>0.001$), which is opposite to the pattern we usually observed. We do not have an explanation as to why this occurred and it was not seen on the other six statistical parameters calculated for this dataset. Increasing the replications to 1,000 and 10,000 times did not affect these findings. However, in the samples extracted using the fixed prevalence method, this anomalous result only persisted for the positive likelihood ratios (one out of 216 comparisons).

The results in Table 2 show that across the cohorts there was consistently greater variability in the estimates of prevalence in the n=100 sample size than in the n=200 samples and in the n=200 than in the n=400 samples. Again, in some of the cohorts this trend continued, but it was patchy and inconsistent.

The results from the pairwise comparisons identifying whether the size of the variability in contingency table statistical estimates was different between samples with a fixed prevalence and those with a varying prevalence are shown in Table 3. In 14.8% of these 216 pairwise comparisons, there was a difference in the variability and almost always (31 of 32 occasions) that difference was less in the samples with a fixed prevalence. This indicates that variability was either the same or less when prevalence was fixed.

The frequency of differences in the size of the variability in contingency table statistical estimates between samples with a fixed prevalence and those with a varying prevalence varied across datasets (hypermobility 12.5%, fear 21.3%, grip strength 10.2%). It also varied across the statistical parameters. It was most common in estimates of post-test probabilities, which may reflect a compounding effect because post-test probabilities are the product of two other estimates (pre-test odds and a likelihood ratio) each of which have their own variability.

Table 2: Pairwise comparisons of whether the variability was different between one sample size and the next largest sample size.

Pairwise comparisons of variability in specified sample sizes		100 vs 200	200 vs 400	400 vs 600	600 vs 800	800 vs 1000	1000 vs 1200	1200 vs 1400	1400 vs 1600	1600 vs 1800	1800 vs 2000
Sensitivity[#]	Hypermobility	**<0.01**	0.06	**<0.01**	0.60	0.47					
	Fear	**<0.01**	**<0.01**	0.40	**0.05**	**0.01**	0.33	0.27	0.29	0.50	
	Grip strength	**<0.01**	**<0.01**	0.74	0.47	0.15	0.06	0.25	0.30	0.28	0.73
Specificity	Hypermobility	0.10	**<0.01**	**<0.01**	0.26	0.20					
	Fear	**0.03**	**<0.01**	**0.04**	**0.01**	0.67	0.15	0.44	0.34	0.54	
	Grip strength	**<0.01**	**<0.01**	**<0.01**	0.12	0.08	0.24	0.93	0.21	0.76	0.55
Positive Likelihood Ratio	Hypermobility	**0.04**	**<0.01**	**<0.01**	0.20	**0.02**					
	Fear	**<0.01**	**<0.01**	**0.03**	**0.04**	0.27	0.21	0.44	0.18	0.67	
	Grip strength	**<0.01**	**<0.01**	**<0.01**	0.12	0.15	0.21	0.32	0.34	0.47	0.64
Negative Likelihood Ratio	Hypermobility	**<0.01**	0.06	**<0.01**	0.59	**0.04**					
	Fear	**<0.01**	**<0.01**	**0.05**	**0.05**	0.08	0.27	0.32	0.30	0.95	
	Grip strength	**<0.01**	**<0.01**	0.38	0.40	0.38	**0.04**	0.25	0.42	0.33	0.53
Post-test probability (+ve test)	Hypermobility	**<0.01**	**<0.01**	**<0.01**	**0.01**	0.26					
	Fear	**<0.01**	**0.02**	**0.02**	0.17	**0.04**	**0.02**	0.26	0.60	0.37	
	Grip strength	**<0.01**	**<0.01**	**0.01**	**<0.01**	0.56	**0.03**	0.84	0.61	0.97	0.73
Post-test probability (-ve test)	Hypermobility	**<0.01**	**0.01**	**<0.01**	0.11	0.45					
	Fear	**<0.01**	**<0.01**	0.11	0.20	0.18	0.07	0.10	0.80	0.35	
	Grip strength	**<0.01**	**<0.01**	0.24	0.48	0.19	0.34	0.06	0.20	0.34	0.06
Odds ratio	Hypermobility	0.07	**<0.01**	**<0.01**	0.45	**0.01**					
	Fear	**<0.01**	**<0.01**	0.29	0.08	0.07	0.06	0.31	0.16	0.73	
	Grip strength	**<0.01**	**<0.01**	**<0.01**	0.06	0.97	0.06	0.11	0.86	0.79	0.06
Risk/ prevalence ratio	Hypermobility	0.14	**<0.01**	**0.02**	0.30	0.09					
	Fear	**<0.01**	**<0.01**	**0.03**	0.07	0.09	0.50	0.27	0.18	0.68	
	Grip strength	**<0.01**	**<0.01**	**0.02**	0.40	0.27	0.38	**0.04**	0.28	0.60	0.06
Prevalence	Hypermobility	**<0.01**	**<0.01**	**0.31**	**<0.01**	0.47					
	Fear	**<0.01**	**<0.01**	0.31	0.12	0.06	**0.01**	0.47	0.68	**0.04**	
	Grip strength	**<0.01**	**<0.01**	0.06	**0.01**	0.50	**0.01**	0.36	0.44	0.35	0.30

[#]All parameters were tested using a 2-tailed *robvar* test for the variance being different between 100 samples drawn at each of two samples sizes. +ve test = result with a positive test result and -ve test = negative test result. vs = a statistical comparison of the variability at one sample size *versus* the variability at the next largest sample size.

Table 3: Pairwise comparisons of whether the variability was different between the samples with a fixed prevalence (the whole cohort prevalence) and those with a variable prevalence.

Pairwise comparisons of variability at each sample size		100	200	400	600	800	1000	1200	1400	1600	1800	2000
Sensitivity[#]	Hypermobility	=	=	=	=	=	=					
	Fear	=	=	=	=	=	=	=	=	=	=	
	Grip strength	=	=	=	=	=	=	=	=	=	=	=
Specificity	Hypermobility	=	=	=	=	=	=					
	Fear	=	=	=	=	=	=	=	<	=	=	
	Grip strength	=	=	<	=	=	=	=	=	=	=	=
Positive Likelihood Ratio	Hypermobility	<	=	=	=	=	=					
	Fear	=	=	=	=	=	=	=	=	=	=	
	Grip strength	=	=	<	=	=	=	=	=	=	=	=
Negative Likelihood Ratio	Hypermobility	=	=	=	=	=	=					
	Fear	=	=	=	=	=	=	=	=	=	=	
	Grip strength	=	=	=	=	=	=	=	=	=	=	=
Post-test probability (+ve test)	Hypermobility	<	=	<	=	=	=					
	Fear	<	<	<	<	<	<	=	<	<	<	
	Grip strength	=	=	<	<	=	=	=	=	<	<	<
Post-test probability (-ve test)	Hypermobility	=	=	=	=	=	=					
	Fear	=	<	<	<	<	<	<	=	<	=	
	Grip strength	=	=	=	=	=	=	=	=	<	=	=
Odds ratio	Hypermobility	>	=	<	=	=	=					
	Fear	=	=	=	=	=	=	=	=	=	=	
	Grip strength	=	=	=	=	=	=	=	=	=	=	=
Risk/ prevalence ratio	Hypermobility	=	=	=	=	<	=					
	Fear	=	=	=	=	=	=	=	=	=	=	
	Grip strength	=	=	=	=	=	=	=	=	<	=	=

[#]All parameters were tested using a 2-tailed *robvar* test ($p<0.05$). '=' indicates the variability was the same. '<' indicates the variability was less in samples with a fixed prevalence, and '>' indicates the reverse.

4. Discussion

4.1. Summary of main findings

We found evidence of very wide and statistically significant variability in estimates derived from contingency tables when calculated in sample sizes of 100 or 200 people, and that typically this variability stabilised in samples of 400 to 600 people. Our findings suggest that, as a broad rule of thumb in the musculoskeletal area, sample sizes in studies that derive prediction rules using contingency tables or measure their performance using contingency tables should include 400 or more participants.

When represented diagrammatically, the relationship between variability and sample size might have been expected to display a cone shape, where increasingly larger sample sizes resulted in greater precision that was observed by increasingly smaller variability. However, some results suggested a more trumpet shape, where the variability plateaued out after a sample size of approximately 600. Some variability in estimates persisted, even in samples of more than 1,000 people, although the consequences of this residual variability for clinical decision-making in some settings, such as the conservative care of musculoskeletal conditions, are likely to be negligible. This visual pattern was reinforced by the results of statistical comparisons.

We found that some statistical parameters showed much more sample-specific variability than others. Nonetheless, all of these statistical parameters showed sample-specific variability of a size that is likely to be clinically important in some settings. One reason for this variability could be due to the varying prevalence of the health condition of interest in the randomly selected samples. For example, the prevalence of high activity limitation in the SpineData cohort ranged from 29% to 56% in the samples of n=100. However, variability in prevalence is likely to only explain a small component in the variability in these statistical parameters. That is because in only approximately 15% of occasions was there less variability in the samples with a fixed prevalence than in the samples with a varying prevalence.

4.2. Strengths and weaknesses

Using data from large cohorts allowed us to use multiple random samples of data of up to 1,000 people or more as a method for identifying sample sizes at which variability of estimates stabilized. It also allowed repeated random sampling, to simulate the same study being performed with different samples from the same source population. In addition, different types of associations were investigated, between physical impairments, a psychological characteristic and a demographic attribute. We also explored the influence of prevalence. A further strength is that we used real data as compared to simulated data, as the true variance in clinical and general population data might be difficult to model in simulated data.

A potential criticism of this study could be that only one of the three cohorts included longitudinal data. However, this would not have affected our findings, as all

2x2 contingency tables simply reflect the relationship between two variables. In addition, some clinical prediction rules, such as the Ottawa Knee Rules, are about cross-sectional relationships [27].

Another potential hesitation about the results might be that, as the method of randomly extracting samples without replacement has the potential to result in different samples that contain some of the same people, then maybe the reduced variability of estimates in larger sample sizes observed in this study was the result of an increased probability of selecting the same people. For two reasons, we do not believe this was a likely influence on our results. The first reason is that we originally analysed the results using the same method, except that we used only four mutually exclusive samples at each sample size. We subsequently switched the analysis to 100 non-mutually exclusive samples at each sample size due to a concern about how representative four samples would have been. Importantly, the results using both approaches showed exactly the same trend of the variability stabilising in samples of 400 to 600 people. The second reason is that a sample size of 400 represents from 5% to 10% of the total sample in the three cohorts we used and a sample of 600 represents from 7% to 15%. Given the differences between these cohort proportions at each sample size, it is highly unlikely that the reduced variability of estimates in samples of 400 to 600 people was the result of the same probability of selecting the same people within each cohort.

4.3. Findings relative to previous studies

We are not aware of other studies that have investigated sample sizes requirements for studies that derive clinical prediction rules. There are a priori sample size estimation procedures for studies of diagnostic test performance, based on the desired confidence interval for likelihood ratios [28] and predictive values [29]. However, these procedures require a priori knowledge of test performance, such as sensitivity and specificity, which in the case of studies that derive clinical prediction rules, cannot be known prior to the commencement of the study. There are also various formulae for calculating sample sizes when using logistic regression for dichotomous outcomes [30], linear regression for continuous outcomes [31] or proportional hazard models for time-to-event outcomes [32, 33]. However, as these focus on sample sizes required for a given number of predictor variables in a multivariable prediction model, these are not applicable for the subsequent phase of deriving prediction rules using contingency tables or measuring their performance using contingency tables. Our findings augment event-per-variable recommendations for binary predictors when constructing Cox regression models[34] and logistic regression models[35], by providing recommendations for sample sizes that minimise variability in outcome estimates from contingency tables.

4.4. Implications of findings

The main implication of these findings is that they provide a 'rule of thumb' estimate for researchers when planning studies that derive prediction rules using 2 x 2 contingency tables or measure their performance using contingency tables. Another implication for the field of subgrouping and clinical prediction rules is that inadequate sample sizes may be a reason why the performance of some prediction

rules has been difficult to replicate, especially when sample sizes used were less than 100.

This knowledge of the influence of sample size on contingency table statistics may assist researchers to adequately power studies, as existing sample size methods for studies of clinical tests require a priori knowledge of test performance that cannot be known prior to the derivation of a prediction rule. It also flags that researchers need to be aware of the extent to which differences in prevalence can, in some instances, affect contingency table-based estimates and therefore, where population prevalence estimates are available, it would be a prudent to select samples using the event-per-variable method.

We have shown that differences in prevalence can affect contingency table-based estimates, with the variability in contingency table statistics being lower in approximately 15% of cases when the prevalence was fixed. However, fixing the prevalence did not change the pattern we observed that samples of 400 to 600 were required before sample-specific variability stabilised.

The extent to which the accuracy of prediction rule performance measures can be generalised across samples with quite different background prevalences was not directly addressed in the current study. One approach to investigating that question would be to use real samples that have quite different background prevalences. Another option might be to artificially manipulate the background prevalence in simulated data (for example using the event-per-variable method to double the prevalence) but the extent to which this would mimic the results of real data is unknown.

5. Conclusions

Increasingly in musculoskeletal care, clinical prediction rules are being created using contingency tables to make decisions about the content of the rule, or estimate prediction rule performance using contingency table statistics. Our findings suggest that, as a broad 'rule of thumb' sample sizes in such studies should be n=400 or more in order to reduce excessive sample-specific variability.

Conflict of interest

The authors declare there to be no conflicts of interest.

Authors' contributions

The concept of the paper originated from PK, who also wrote the first draft of the manuscript. All authors were involved in the design of the study, the drafting and revision of the manuscript, and gave final approval of the manuscript.

Acknowledgements

PK was partially funded by the Danish Fund for Chiropractic Research and Post-graduate Education. No funding source played any role in the scientific conduct of the study.

Figure legends:

Figure 1. Study flow

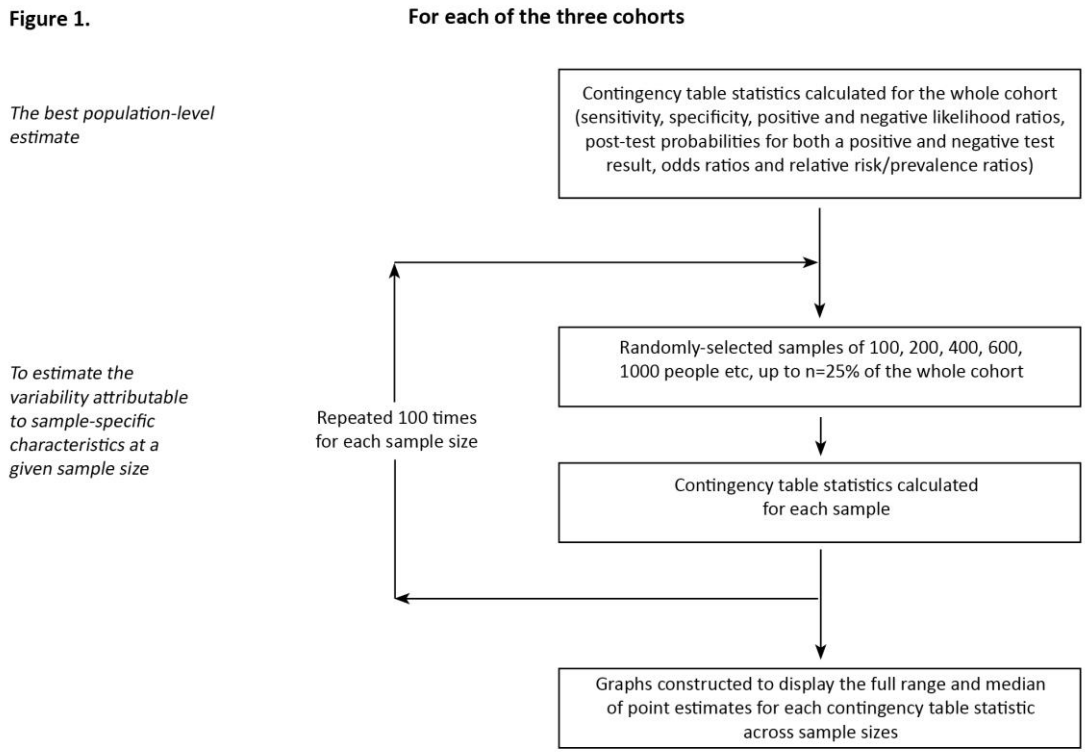


Figure 2. Median and range of *sensitivity and specificity* across the different sample sizes for each of the datasets

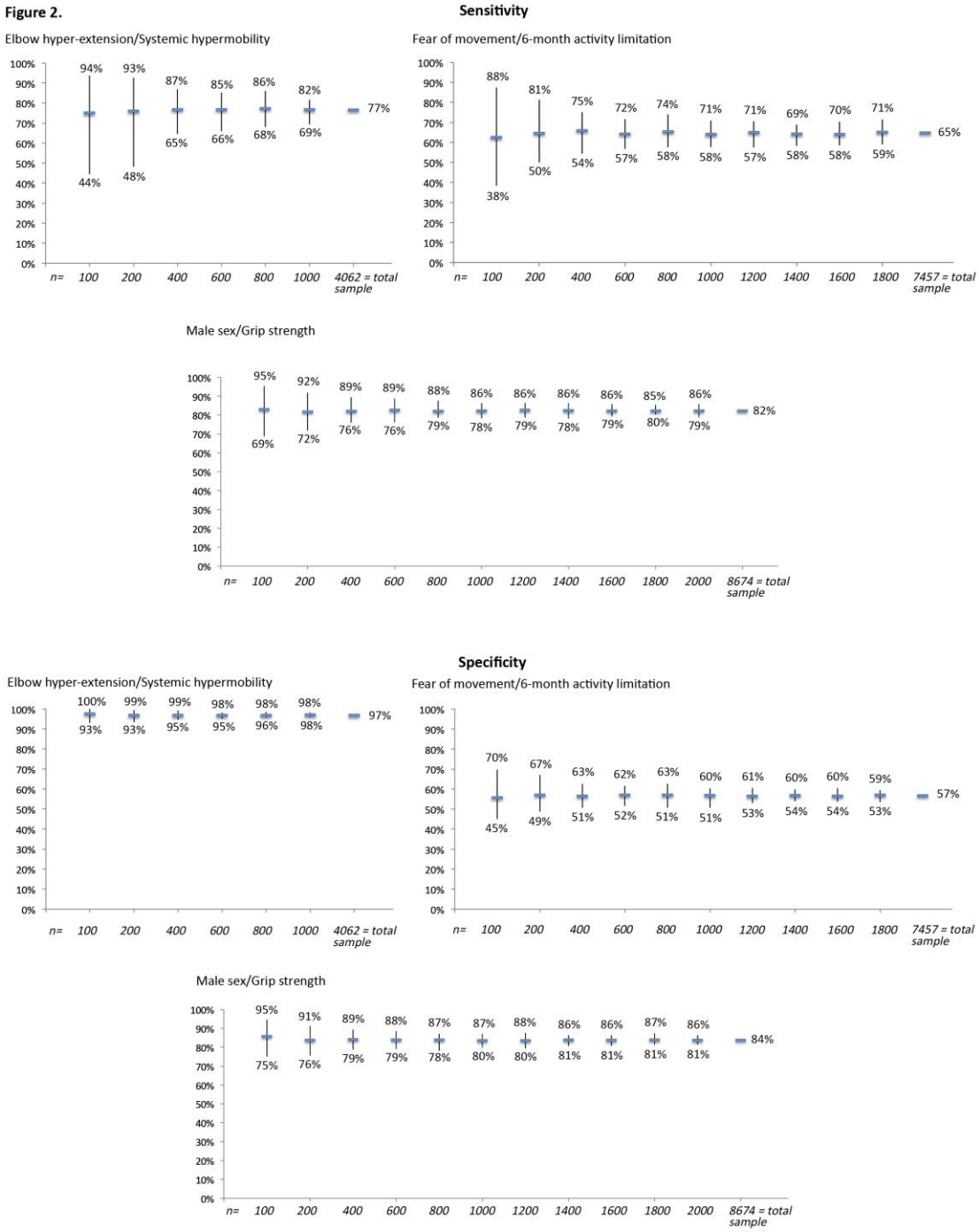
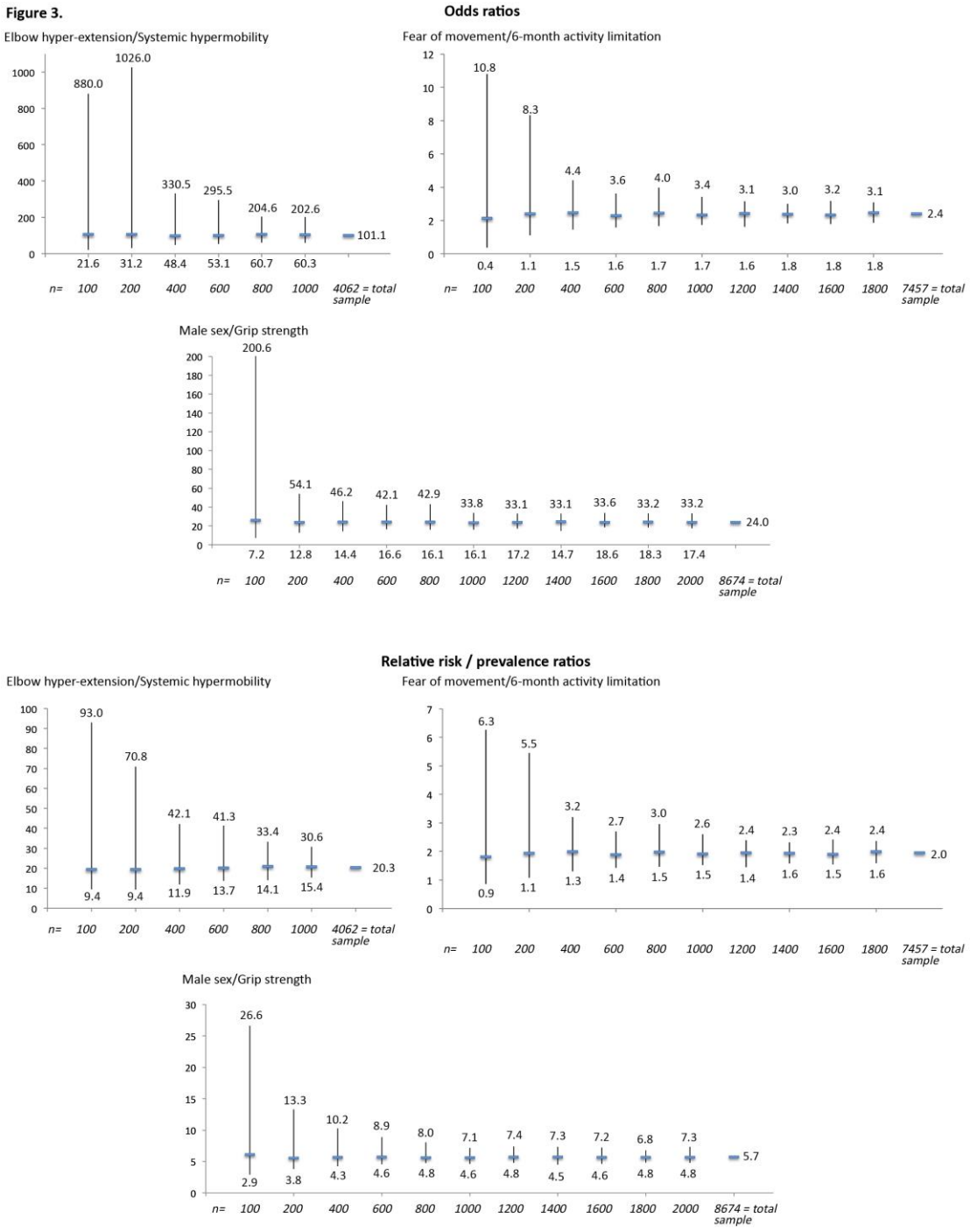


Figure 3: Median and range of *odds ratios and risk/prevalence ratios* across the different sample sizes for each of the datasets



References

[1] Keogh C, Wallace E, O'Brien KK, Galvin R, Smith SM, Lewis C, et al. Developing an international register of clinical prediction rules for use in primary care: a descriptive analysis. Ann Fam Med. 2014;12:359-66.

[2] McGinn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG, Richardson WS. Users' guides to the medical literature: XXII: how to use articles about clinical decision rules. JAMA. 2000;284:79-84.

[3] Childs JD, Cleland JA. Development and application of clinical prediction rules to improve decision making in physical therapist practice. Phys Ther. 2006;86:122-31.

[4] Flynn T, Fritz JW, Whitman M, Wainner RS, Magel J, Rendeiro D, et al. A clinical prediction rule for classifying patients with low back pain who demonstrate short-term improvement with spinal manipulation. Spine. 2002;27:2835-43.

[5] Breiman L, Friedman J, Olshen R, Stone C. Classification and Regression Trees. Belmont, CA, USA: CRC Press; 1984.

[6] Begg CB. Biases in the assessment of diagnostic tests. Stat Med. 1987;6:411-23.

[7] Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999;282:1061-6.

[8] Brenner H, Gefeller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Stat Med. 1997;16:981-91.

[9] Moons KG, van Es GA, Deckers JW, Habbema JD, Grobbee DE. Limitations of sensitivity, specificity, likelihood ratio, and Baye's Theorem in assessing diagnostic probabilities: a clinical example. Epidemiology. 1997;8:12-7.

[10] Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. Annals Int Med. 2002;137:598-602.

[11] Hicks GE, Fritz JM, Delitto A, McGill SM. Preliminary development of a clinical prediction rule for determining which patients with low back pain will respond to a stabilization exercise program. Arch Phys Med & Rehab. 2005;86:1753-62.

[12] Stiell IG, Wells GA, Vandemheen KL, Clement CM, Lesiuk H, De Maio VJ, et al. The Canadian C-spine rule for radiography in alert and stable trauma patients. JAMA. 2001;286:1841-8.

[13] Wainner RS, Fritz JM, Irrgang JJ, Delitto A, Allison S, Boninger ML. Development of a clinical prediction rule for the diagnosis of carpal tunnel syndrome. Arch Phys Med & Rehab. 2005;86:609-18.

[14] Wainner RS, Fritz J, Irrgang JJ, Boninger ML, Delitto A, Allison S. Reliability and diagnostic accuracy of the clinical examination and patient self-report measures for cervical radiculopathy. Spine. 2003;28:52-62.

[15] Albert HB, Kent P, Jensen J, Dragsbæk L. Screening for generalised hypermobility in patients with low back pain. J Bone Joint Surg Br 2010;92-B.

[16] Grahame R, Bird HA, Child A. The revised (Brighton 1998) criteria for the diagnosis of benign joint hypermobility syndrome (BJHS). J Rheumatol. 2000;27:1777-9.

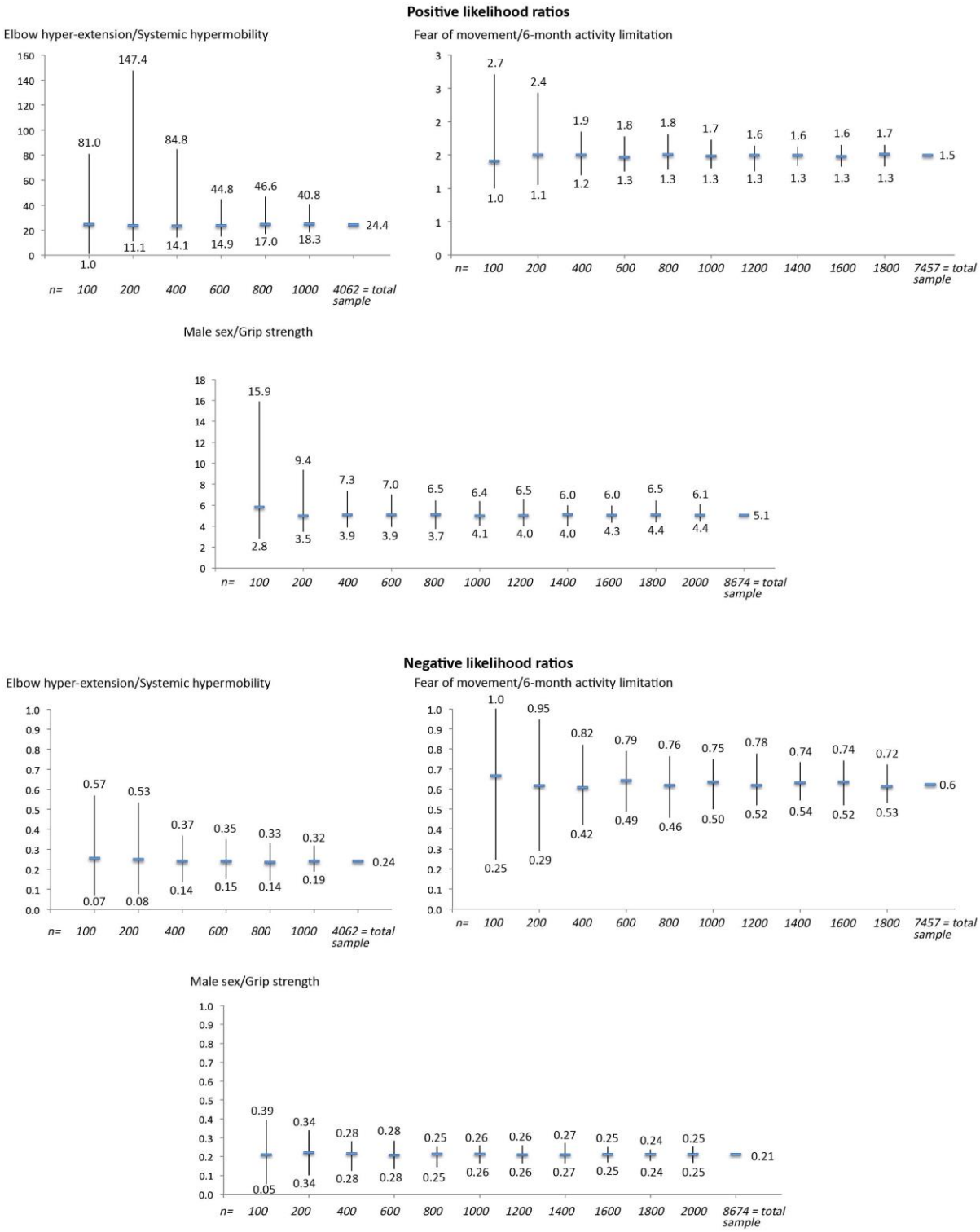
[17] Juul-Kristensen B, Rogind H, Jensen DV, Remvig L. Inter-examiner reproducibility of tests and criteria for generalized joint hypermobility and benign joint hypermobility syndrome. Rheumatology (Oxford). 2007;46:1835-41.

[18] Kent P, Kongsted A, Jensen TS, Albert HB, Schiøttz-Christensen B, C. M. SpineData – a Danish clinical registry of people with chronic back pain. Clin Epi 2015;369-80. doi.org/10.2147/CLEP.S83830.
[19] Waddell G, Newton M, Henderson I, Somerville D, Main CJ. A Fear-Avoidance Beliefs Questionnaire (FABQ) and the role of fear-avoidance beliefs in chronic low back pain and disability. Pain. 1993;52:157-68.
[20] Kent P, Mirkhil S, Keating J, Buchbinder R, Manniche C, Albert HB. The concurrent validity of brief screening questions for anxiety, depression, social isolation, catastrophization, and fear of movement in people with low back pain. Clin J Pain. 2014;30:479-89.
[21] Albert HB, Jensen AM, Dahl D, Rasmussen MN. Criteria validation of the Roland Morris questionnaire. A Danish translation of the international scale for the assessment of functional level in patients with low back pain and sciatica. Ugeskr Laeger. 2003;165:1875-80.
[22] Lauridsen HH, Hartvigsen J, Manniche C, Korsholm L, Grunnet-Nilsson N. Responsiveness and minimal clinically important difference for pain and disability instruments in low back pain patients. BMC Musculoskelet Disord. 2006;7:doi:10.1186/471-2474-7-82.
[23] Kent P, Lauridsen HH. Managing missing scores on the Roland Morris Disability Questionnaire. Spine. 2011;36:1878-84.
[24] Frederiksen H, Hjelmborg J, Mortensen J, McGue M, Vaupel JW, Christensen K. Age trajectories of grip strength: cross-sectional and longitudinal data among 8,342 Danes aged 46 to 102. Ann Epidemiol. 2006;16:554-62.
[25] Deeks JJ, Altman DG. Diagnostic tests 4: Likelihood ratios. BMJ. 2004;329:168–9.
[26] Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med. 2016;35:214-26.
[27] Stiell IG, Greenberg GH, Wells GA, McDowell I, Cwinn AA, Smith NA, et al. Prospective validation of a decision rule for use of radiography in acute knee injury. JAMA. 1996;275:611-15.
[28] Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. J Clin Epidemiol. 1991;44:763-70.
[29] Steinberg DM, Fine J, Chappell R. Sample size for positive and negative predictive value in diagnostic research using case-control designs. Biostatistics. 2009;10:94-105.
[30] Demidenko E. Sample size determination for logistic regression revisited. Stat Med. 2007;26:3385-97.
[31] Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. Stat Med. 1998;17:1623-34.
[32] Schmoor C, Sauerbrei W, Schumacher M. Sample size considerations for the evaluation of prognostic factors in survival analysis. Stat Med. 2000;19:441-52.
[33] Jinks RC, Royston P, Parmar MK. Discrimination-based sample size calculations for multivariable prognostic models for time-to-event data. BMC Med Res Methodol. 2015;15:82 doi: 10.1186/s12874-015-0078-y.
[34] Ogundimu EO, Altman DG, Collins GS. Adequate sample size for developing prediction models is not simply related to events per variable. J Clin Epidemiol. 2016.

[35] Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49:1373-9.

Appendix.

Appendix Figure 1: Median and range of *positive and negative likelihood ratios* across the different sample sizes for each of the datasets



Appendix Figure 2: Median and range of *post-test probabilities* across the different sample sizes for each of the datasets