The Andersen aerobic fitness test: New peak oxygen consumption prediction equations in 10 and 16-year-olds

Eivind Aadland PhD[1], Lars Bo Andersen PhD[1], Øystein Lerum MSc[1], Geir Kåre Resaland PhD[1]

[1]*Western Norway University of Applied Sciences, Faculty of Teacher Education and Sports, Campus Sogndal, Norway.*

Running title: The Andersen test for children

Corresponding author:

Eivind Aadland

Faculty of Teacher Education and Sport, Western Norway University of Applied Sciences, Campus Sogndal, Box 133, 6851 Sogndal, Norway. Phone: +47 5767 6086; Email: eivind.aadland@hvl.no

Word count: Main text 4433; Abstract 232

Abstract

Introduction: Measurement of aerobic fitness by determining peak oxygen consumption (VO_{2peak}) is often not feasible in children and adolescents, thus field-tests such as the Andersen test are required in many settings, for example in most school-based studies. The present study provides cross-validated prediction equations for VO_{2peak} based on the Andersen test in 10 and 16-year-old children. **Methods:** We included 235 children (n = 113 10-year-olds and 122 16-year-olds) who performed the Andersen test and a progressive treadmill test to exhaustion to determine VO_{2peak}. Joint and sex-specific prediction equations were derived and tested in 20 random samples. Performance in terms of systematic (bias) and random error (limits of agreement) was evaluated by means of Bland Altman plots. **Results:** Bias varied from -4.28 to 5.25 ml/kg/min across testing datasets, sex and the two age groups. Sex-specific equations (mean bias -0.42–0.16 ml/kg/min) performed somewhat better than joint equations (-1.07–0.84 ml/kg/min). Limits of agreement were substantial across all datasets, sex and both age groups, but were slightly lower in 16-year-olds (5.84–13.29 ml/kg/min) compared to 10-year-olds (9.60–15.15 ml/kg/min). **Conclusions:** We suggest the presented equations can be used to predict VO_{2peak} from the Andersen test performance in children and adolescents on a group level. Although the Andersen test appears to be a good measure of aerobic fitness, researchers should interpret cross-sectional individual level predictions of VO_{2peak} with caution due to large random measurement errors.

Keywords: cardiorespiratory fitness; measurement; agreement; prediction, field-test

Introduction

High aerobic fitness is consistently associated with a favorable metabolic risk profile in children [1 2], and can together with fatness be applied to identify children with clustering of cardiovascular disease risk factors [3]. Maximal or peak oxygen consumption (VO_{2peak}) measured to voluntary exhaustion is regarded the gold standard for determining aerobic fitness, but is often not feasible for testing large samples of children. Therefore, a range of maximal and submaximal measures exists [4], of which two highly used tests in children is the 20 m multistage shuttle run test (MSRT) [5 6] and the Andersen test [7 8]. Compared to the MSRT, the Andersen test may be particularly suitable for children, as it relates closer to children's natural intermittent running pattern, and as it does not exclude and stigmatize children having poor fitness. Regarding measurement properties, both tests seem to perform well in terms of reliability [7-9], but have certain limitations when it comes to prediction of VO_{2peak} [8 10 11].

We have previously evaluated the measurement properties of the Andersen test in 10-year-old children, concluding an overall good performance in this group [8]. Yet, the study had two limitations. First, we only included 10-year-old children, meaning that the equation suggested for VO_{2peak} might not be valid in other age groups. Increased work economy as children age is a clear candidate to affect this relationship [12], despite no interaction with age was found by Andersen et al [7]. Second, as we found a small learning effect and reduced random error from test 1 to test 2 (mean bias ± limits of agreement [LoA] 27 ± 125 m, r = 0.82), but similar performance for test 2 and test 3 (3.9 ± 89 m, r = 0.92), we argued that future studies should undertake two tests and use the better performance for analysis. Thus, the prediction equation for VO_{2peak} was suggested on this basis. However, in large population-based studies using the Andersen test [13-15] it might not be feasible to conduct two tests. Because the test is subject to both systematic and random error, applying the previously suggested equation established for the better of two tests, will likely lead to biased prediction estimates when applied to the first (one and only) test. For the better measure, systematic error will lower the intercept (as a better Andersen performance is related to the same VO_{2peak}), whereas less random

error will cause a steeper slope (and a lower intercept), because of less inflation of the regression coefficient due to regression dilution bias [16]. Such an effect has previously been shown for the MSRT [17].

Therefore, the aim of the present study was to evaluate the performance of the previously suggested prediction equations for VO_{2peak} based on the Andersen test [7 8] in 16-year-old children whose VO_{2peak} were determined from a maximally graded treadmill protocol. We hypothesized that 1) new equations would be needed to fit the joint sample of 10 and 16-year-old children, but that joint equations could be established across the age groups, as previously suggested [7], and that 2) equations based on one test would provide lower slopes and higher intercepts than when using the best of several tests.

Materials and methods

Participants

We included two convenience samples for the current analyses; one in 10-year old children [8] and one in 16-year old children. Both samples were recruited from the general population of schoolchildren in the rural county of Sogn og Fjordane, in the western part of Norway.

10-year old children: We invited all 121 children attending fifth grade at one school over two consecutive school years (2012-2014) to participate in the study. A total of 118 children (67 boys and 51 girls; 58 during 2012–2013 and 60 during 2013–2014) were included in the study.

16-year old children: We included children at two schools during 2014-2015 recruited to a five-year follow-up of the Sogndal School Intervention Study [18]. Of the 247 children originally included in the intervention study, 241 provided consent for the follow-up study.

The relevant school authorities were invited to take part in the studies, and provided their consent, before children and their parents were given thorough oral and written information regarding the study protocol. Each child orally agreed to participate in the study, and written informed consent was obtained from each child's parent(s)/guardian(s) prior to the child's inclusion in the study. The studies met the standards of the Declaration of Helsinki and was approved by the Regional Committee for Medical Research Ethics (REC West) in Norway (reference numbers: 2012/1089 and 2004/14).

Study protocol

As previously described, the 10-year-old children performed three Andersen tests (weeks 1, 3, and 4), and performed one incremental treadmill test to exhaustion (week 2) to measure their VO_{2peak} within three weeks [8]. The 16-year-olds performed one Andersen test (week 1) and one VO_{2peak} test (week 2). For the purpose of the present analyses, we applied the first and the best of three Andersen tests in 10-year-old children to explore the differences between these two approaches in this age group only. For establishing equations in the joint sample of 10-year-olds and 16-year-olds, we used the 10-year-olds' first test.

The Andersen test was performed according to standard procedures [7]. Two parallel lines 20 m apart were marked in a gym hall with a wooden floor. We informed the children about the procedures and performed a collective five-minute warm-up before the test. The test has a total duration of 10 minutes, where children run from one end line to another in a to-and-fro movement intermittently, with 15-second work periods and 15-second breaks signaled by the test leaders blowing a whistle. When the children finished one 15-second period of work, they were instructed to stop as fast as possible and to take one to three steps back, depending on how fast they were able to stop. Each time the children turned around at an end line, they had to touch with one finger the floor behind

the end line. The goal was to cover the longest possible distance during the 10-minute run. The distance covered (number of laps performed) was recorded by adult test assistants who counted for one or two children each. We tested 15–20 children per test. The gym hall was 18 meter wide, giving each child a lane of about 1 m.

VO_{2peak} was measured to exhaustion using an incremental treadmill test. The treadmill's inclination (Woodway PPS 55, Woodway GmbH, Weil am Rhine, Germany) was constant at 5.3% during the test. Children started to walk at 5 km/h for 5 minutes. Thereafter the speed increased by 1 km/h each minute until the children were exhausted. Oxygen consumption was measured using the Moxus Modular Metabolic System (AEI Technologies Inc., Pittsburgh, USA). A two-point gas calibration according to known concentrations and calibration according to atmospheric pressure were performed each test day. Volume calibration of the breathing valve (Hans Rudolph model 2700, Hans Rudolph Inc., Shawnee, Kansas, USA) was performed between each test using a 3-l syringe (Series 5530, Hans Rudolph, Kansas, USA). The oxygen analyzer has shown to be reliable and valid compared to the Douglas-bag technique [19]. The child and parent(s)/guardian(s) were informed of test procedures before testing, and the child's parent(s)/guardian(s) were allowed and encouraged to observe the testing.

After each test, test leader and associates discussed several subjective criteria to verify a near maximal performance: hyperpnoea, unsteady running pattern, and verbal and body language clearly indicating that the child wanted to stop testing despite repeated strong verbal encouragement. Additionally, the peak respiratory exchange ratio (RER_{peak}) and heart rate (HR_{peak}) (Polar S610i HR monitor, Polar Electro OY, Kempele, Finland) were noted.

The VO_{2peak} is presented as absolute (l/min) and relative values (ml/kg/min), each of which is defined as the highest set of two successive 30-second measurements. Height and body weight were measured without shoes and socks before the children started the VO_{2peak} test. Height was measured to the nearest 0.1 cm using a wall-mounted stadiometer. Body weight was measured to the nearest

0.1 kg (subtracting 0.2 kg for light clothes) using an electronic scale (Seca 770, SECA GmbH, Hamburg, Germany). Body weight was used as a continuous variable in the statistical analyses. To report descriptive statistics, children were also categorized as normal weight, overweight, or obese according to the criteria set by Cole et al. [17].

Statistical analyses

The anthropometric subject characteristics and data on VO_{2peak} and the Andersen test are presented as the mean values and standard deviations (SD). Before evaluating performance of prediction equations in 16-year-old children, we reanalyzed a sample of 10-year-olds having three valid Andersen tests and a valid VO_{2peak} test (n = 100). To test for the impact of establishing (using the first vs. the best of three Andersen tests) and predicting VO_{2peak} (using the first vs. the best of three Andersen tests) applying equations based on different Andersen tests, we evaluated the performance of these approaches by determining the bias towards measured VO_{2peak}.

Performance of prediction equations for VO_{2peak} using the Andersen test was assessed using linear regression and Bland Altman plots with LoA in three steps. 1) We applied the suggested equations from Andersen et al. [7] (VO_{2peak} = 18.38 + 0.033*Andersen distance – 5.92*gender [boys = 0; girls = 1]) and Aadland et al. [8] (VO_{2peak} = 23.262 + 0.050*Andersen distance –3.858*gender – 0.376*body weight [boys = 0, girls = 1]) to predict VO_{2peak} in the 16-year-olds. 2) To develop new equations to predict VO_{2peak} from the Andersen test, we randomly split our total joint sample of 10-year-olds and 16-year-olds in two to perform a cross-validation of our new equations in an independent dataset. The random draw was stratified for age and sex, leaving 119 children in the training dataset (n = 37 15-year old boys, n = 25 15 year old girls, n = 33 10-year old boys, and n = 24 10-year old girls) from which the equations were developed, whereas the other half served as the testing dataset (n = 116). We repeated this process 20 times to evaluate stability and variability, meaning that all cross-

validation results are based on the performance over 20 different training and test datasets. We developed and tested six different equations. As in the previous study in 10-year olds, we included the Andersen test, body mass and sex in the equation (VO_{2peak} = a + b*Andersen distance + c*sex + d*body weight, [boys = 0; girls = 1]) [8]. As body mass might not be available in some settings, but is highly correlated with age, we also tested a model including age instead of body mass (VO_{2peak} = a + b*Andersen distance + c*sex + d*age, [boys = 0; girls = 1]). Prior to splitting the sample, we tested whether a sex-specific association was indicated for these models by adding the interaction term Andersen test*sex to the models described above. As expected, a sex-specific association with the Andersen test was indicated for both models (Andersen test*sex p = .025 in the model for body mass and p = .090 in the model for age). Thus, we cross-validated models for body mass and age for boys and girls in joint (joint equations) and in separate groups (sex-specific equations). The predicted and measured VO_{2peak} were then compared in the testing datasets by calculating the mean as well as minimum-maximum bias (i.e., systematic error), LoA (i.e., random error, LoA = SD of the differences*1.96), explained variance (R^2), and standard error of the estimate (SEE) over the 20 cross-validation models [20]. 3) Finally, we calculated new equations based on the whole sample (n = 235). All equations are reported as regression coefficients, with its R^2 and SEE.

All analyses were performed using IBM SPSS v. 23 (IBM Corporation, Software Group, Somers, New York, USA). A p-value < .05 indicated statistically significant findings.

Results

Children's characteristics

Numbers of children were relatively similar among 10-year-olds (n = 113) and 16-year-olds (n = 122), but there was an overweight of boys (n = 139) compared to girls (n = 96) (Table 1). Children were mainly of Caucasian origin. All variables differed between the 10 and 16-year old boys (p < .001),

whereas all variables but VO_{2peak} (ml/kg/min) ($p = .814$) differed between the 10 and 16-year old girls ($p < .001$). Boys exhibited higher aerobic fitness compared to girls in both age groups, but differences were much greater in 16-year olds ($p < .001$) compared to 10-year olds ($p \leq .048$). The 16-year old girls were somewhat heavier than the 16-year old boys ($p = .029$); boys and girls were otherwise similar within each age group. Beside a somewhat lower RER_{peak} in 10-year old boys compared to 16-year old boys ($p = .003$) and 10-year old girls ($p = .035$), the effort on the VO_{2peak} test was similar across age and sex groups.

The numbers of children that provided data for the analyses is shown in figure 1. In 10-year-olds some children could not perform the Andersen tests for being sick or out of school. Valid tests were obtained in 113 (test 1), 112 (test 2), and 112 (test 3) children. All the 113 children providing valid data on directly measured VO_{2peak} on the graded treadmill protocol (two children did not perform the test, one child was excluded for not performing a maximal test, and two children were excluded due to technical errors) provided valid data on the first Andersen test, whereas 100 children provided valid data on all Andersen tests. Thus, 100 children were included in the analysis of performance of different equations in the 10-year-olds only, whereas 113 children were included in the main analyses.

In the 16-year-olds we did not manage to perform the Andersen test in one class (n = 63 children) for scheduling problems. Moreover, 14 children were injured or sick at the time of testing and did not perform the test, whereas 32 children did not want to perform the test for unknown reasons. Of those performing the test (n = 132), 126 children provided valid data. Regarding VO_{2peak}, 14 children were injured or sick at the time of testing and did not perform the test, whereas 47 children did not want to perform the test for unknown reasons. We obtained valid data in all 180 children who performed the test, of whom 122 children also had a valid Andersen test and were included in the analyses

Insert table 1 about here

Performance of equations using different Andersen tests in the 10-year-old children

In the sample of 10-year-olds, the prediction equations derived when using the first Andersen test and the best of three Andersen tests differed in their intercepts and slopes for the Andersen test. Equations derived from and applied to the same test (i.e., derived from test 1 and applied to test 1; derived from the best of three tests and applied to the best of three tests) resulted in estimates of VO_{2peak} similar to the measured values. On the contrary, the equation derived from the first test but applied to the best of three tests overestimated VO_{2peak}, and the equation derived from the best of three tests but applied to the first test underestimated VO_{2peak}.

Insert table 2 about here

Performance of previously suggested equations in 16-year-olds

The bivariate relationships between the Andersen tests and VO_{2peak} were r = 0.67 in the overall sample, and r = 0.68, r = 0.51, r = 0.59, and r = 0.74 in 10-year old boys, 10-year old girls, 16-year old boys, and 16-year old girls, respectively.

The equations to predict VO_{2peak} suggested by Andersen et al. [7] and Aadland et al [8] was clearly inadequate to predict VO_{2peak} in the 16-year olds (Supplemental Figure 1). Both equations appeared to have an overall acceptable model fit (Andersen equation: R^2 = 0.70, SEE = 5.04, LoA = 10.02 ml/kg/min; Aadland equation: R^2 = 0.61, SEE = 5.77, LoA = 12.23 ml/kg/min). Yet, both equations led to a clear underestimation of VO_{2peak} (Andersen equation: mean (95% CI) bias = -5.15 (-6.06–-4.23) ml/kg/min in the total sample, -6.03 (-7.32–-4.73) ml/kg/min in boys, and -3.84 (-5.01–-2.67) ml/kg/min in girls; Aadland equation: mean (95% CI) bias = -4.15 (-5.27–-3.03) ml/kg/min in the total

sample, -5.43 (-6.92–-3.94) ml/kg/min in boys, and -2.25 (-3.84–-0.66) ml/kg/min in girls). Thus, a new equation was required in the 16-year olds.

Development of new equations

To develop new equations to predict VO_{2peak} from the Andersen test, we initially split our sample in two groups (a training dataset including 119 children and a testing dataset including 116 children) to perform 20 cross-validations of our equations in independent datasets. Model fit in the training datasets for equations including body mass were (mean (minimum-maximum)) R^2 = 0.65 (0.56–0.72) and SEE = 5.49 (4.92–5.90) ml/kg/min in the overall sample, R^2 = 0.55 (0.45–0.63) and SEE = 5.78 (5.02–6.45) ml/kg/min in boys, and R^2 = 0.47 (0.35–0.58) and SEE = 4.88 (4.33–5.55) ml/kg/min in girls. Corresponding numbers for equations including age was R^2 = 0.61 (0.54–0.68) and SEE = 5.79 (5.37–6.18) ml/kg/min in the overall sample, R^2 = 0.51 (0.44–0.61) and SEE = 6.07 (5.20–6.61) ml/kg/min in boys, and R^2 = 0.38 (0.25–0.52) and SEE = 5.30 (4.78–6.08) ml/kg/min in girls. The bivariate correlation between body mass and age was r = 0.80.

Cross-validation performance

Mean values were close to identical between predicted and measured VO_{2peak} for all equations, but with substantial variation between test datasets (Table 3, Figure 2). Mean systematic bias were somewhat larger in joint equations (mean -1.07–0.84 ml/kg/min) than in sex-specific equations (mean -0.42–0.16 ml/kg/min). There were otherwise no specific pattern of bias between age or sex groups. Equations using body mass and age performed very similarly. Yet, bias differed approximately 0–2 ml/kg/min between children performing below (generally overestimated VO_{2peak})

versus above (generally underestimated VO_{2peak}) median Andersen performance (Supplemental Table 1).

Random error (i.e., LoA, R^2 and SEE) was somewhat lower for equations using body mass compared to age (Table 3, Figure 3). Moreover, random error was lower in 16-year-olds than in 10-year-olds, and lower in girls compared to boys, especially in 16-year-olds. For sex-specific equations using body mass and age, respectively, LoA amounted to 22 and 24%, 24 and 26 %, 17 and 17 %, and 15 and 18 % of the mean VO_{2peak} for 10-year-old boys, 10-year-old girls, 16-year-old boys, and 16-year-old girls.

Insert table 3 about here

As bias was less in sex-specific equations than for the equations constructed in the dataset including both boys and girls due to the interaction with sex, we established new sex-specific equations for the whole sample of boys (n = 139) and girls (n = 96) separately (Table 4).

Insert table 4 about here

Discussion

We found that previously suggested equations for prediction of VO_{2peak} based on the Andersen test provided biased estimates in the 16-year-old children. Moreover, VO_{2peak} should be predicted on the same assumptions that were the basis for the development of the equations to avoid biased estimates, meaning that one should be aware whether one or the best of several tests were used for equation development. Therefore, we suggest two new sex-specific equations to predict VO_{2peak} from one Andersen test based on a joint sample of 10-year-olds and 16-year-olds. Importantly, this age-span cover the period before and after the most pronounced maturational developments, lending

credit to the usefulness of the suggested equations in the vast majority of schoolchildren. Although the equations can be used to predict VO_{2peak} for groups of children, our findings show great errors in predictions for individual children, meaning that predicting VO_{2peak} on an individual level should be performed with caution.

Similar to the finding that the original Andersen equation [7] provided biased estimates of VO_{2peak} in 10-year-old children [8], we found that both previous equations [7 8] provided biased estimates of VO_{2peak} in 16-year old children. This finding also extends to external validations of different equations for the MSRT in children and youth [10 11 21 22] as well as in adults [23 24]. Given the inherent variation among often small and homogeneous samples from which equations are derived and tested, measurement error in the determination of VO_{2peak} [25 26], and variation in execution of protocols, biased predictions are expected. However, the amount of bias in these studies (± approximately 0–8 ml/kg/min) is rather similar to the bias previously found among test centers in a multi-site study [27] and biases found in the 20 cross-validations performed in the current study (± approximately 0–5 ml/kg/min), relying on more similar methodology. These findings clearly show that prediction of VO_{2peak} should be performed with caution. Yet, the final equations suggested are based on twice the sample size as the cross-validated equations, thus we would expect improved performance in future settings.

Our findings also reveal that the precision in determining the Andersen test performance and thus measurement error introduce a systematic bias. As previously shown for three repeated MSRT tests [17], using the better of three Andersen tests compared to the first test improved the relationship between the Andersen test and VO_{2peak} in the 10-year-old children (i.e., the slope for the Andersen test steepened). This imply that the prediction equation improves as noise in the independent (x) variable(s) and thus regression dilution bias decrease [16]. A learning effect, as previously shown in this sample of 10-year-olds [8], will further influence the equation (i.e., the intercept). Still, we show in the present study that the “optimal” equation is not necessarily the equation with the best external validity. Contrary to measures of for example muscle strength, motor skills, anthropometry or blood

pressure, where one might easily perform several measurements and apply the average or the best result for analysis, conducting several measurements of aerobic fitness is often not feasible. Our findings demonstrate that the prediction equation should be developed from and applied to the same test situation, that is, using the first (one and only) test, or the best of two or more tests, to avoid systematic under or overestimation. Importantly though, an equation based on a poor test (i.e., where the slope is attenuated towards the null) will cause a systematic underestimation of VO_{2peak} in the most fit and an overestimation of VO_{2peak} in the least fit, as indicated in the present study for the Andersen test and previous studies using the MSRT [10 11]. To this end, we also compared the bias in estimates of VO_{2peak} between the children exhibiting an Andersen test performance below vs. above the median level. This analysis showed the expected trends, but indicated minor biases in both groups (up to ± approximately 1 ml/kg/min for sex-specific equations).

While Leger et al [6] in their original MSRT-equation included age, later equations for the MSRT have included sex, in addition to age and/or a measure of body fat [28-30], with no obvious differences in performance [10 11]. Andersen et al [7] included only the Andersen distance and sex in their original equation, whereas we added body mass in the previously suggested equation in 10-year-olds [8]. In the present study, we found that equations including body mass and age (separately) performed similarly with regard to bias, indicating they can be used interchangeably. However, equations using body mass provided marginally lower LoA across all sex and age groups. The strong association between body mass and age in children and youth (r = 0.80 in the present study) underpins the similar performance. Arguably, though, consistent with previous conclusions with regard to both the Andersen test [8] and the MSRT [10 11], considering the large random error, neither of the equations are suited to predict VO_{2peak} on an individual level, in which case errors of ± 10–15 ml/kg/min should be expected. Thus, we suggest predicting VO_{2peak} on an individual level from indirect tests is a praxis to avoid. Yet, because there are currently no available reference material for the Andersen test, predicting VO_{2peak} could be relevant for clinical purposes. Because the Andersen test is relatively easy to perform in large samples and together with fatness provide a good measure of metabolic health

status [3], it could be a suitable tool for health screening of schoolchildren. However, in addition to the challenge related to interpretation of the data for clinical use, such use need to be considered carefully related to the school's purpose and mandate, how testing possibly could be included in the school curriculum, and procedures for follow-up and intervention of children at risk. Hence, although the Andersen test is a candidate measure for health screening of children, these issues should be thoroughly considered before possibly implemented in the school system.

Perspective

Each researcher must ultimately decide whether a measurement tool is reliable and valid, given the purpose of the study [31]. Given the high reliability of the field tests discussed herein [7-9], and their obvious face validity given their strain induced upon the cardiorespiratory and muscular systems, we argue these tests are well suited to assess aerobic fitness in a range of settings, despite evidence of moderate associations with VO_{2peak}. Actually, most evidence linking aerobic fitness to health are based on performance measures, as exemplified by the most influential study cohort in terms of establishing aerobic fitness as a strong predictor of longevity, the Cooper Center Longitudinal Study [32-34], which relies on performance on the Balke protocol [35]. When compared to each other, directly measured VO_{2peak} and various performance measures have been shown to be of similar predictive value regarding mortality [36 37] and sport performance [38 39]. Moreover, when compared in the present sample of 10-year-olds, the Andersen test shows the strongest relationship with metabolic health [40]. Thus, we do not regard the present study an investigation of validity, but rather to what extent both measures can be used, and results expressed, in a common metric for aerobic fitness.

Strengths and limitations

The present study has two main strengths. First, we included a relatively large sample of 10-year-old and 16-year-old children (total n = 235 children), thus, overcoming the limitation of the previous

investigation including 10-year-olds only [8]. By doing so, we included children on both shoulders of puberty. Second, we performed a thorough cross-validation deriving and applying 20 different equations to independent datasets. This approach gives a solid picture of the stability and variability of our prediction equations when applied to new samples. The variability of results for both systematic and random error indicate a certain variation over datasets that is not captured by performing one cross-validation only. Yet, a limitation is that our test datasets was not fully independent, as the children composing both the training and test datasets came from the same sample and performed the tests under the same conditions. Thus, the equations suggested might perform worse in other contexts, calling for further external validation studies in new settings and with new samples of children. Nevertheless, considering the double sample size for the final equations compared to the training equations, we believe future performance will be within the variability of estimates shown herein.

Conclusions

We conclude that previously suggested equations for prediction of VO_{2peak} based on the Andersen test [7 8] was insufficient when applied to 16-year-old children. We therefore suggest new sex-specific equations to predict VO_{2peak} based on a joint sample of 10-year-olds and 16-year-olds. This age-span cover the vast majority of middle-school children and high-school adolescents, lending credit to the usefulness of the suggested equations in many pediatric study settings. Yet, our findings add to the literature showing that prediction of VO_{2peak} should be given considerable thought and be performed with caution. Although the Andersen test appears to be a good measure of aerobic fitness, researchers should interpret cross-sectional individual level predictions of VO_{2peak} carefully do to large random measurement errors.

Acknowledgements

We thank all children, teachers and principals at the two participating schools for their excellent cooperation during the data collection. We also thank students at the Western Norway University of Applied Sciences (previously Sogn og Fjordane University College) for their assistance during the test sessions.

References

1. Andersen LB, Sardinha LB, Froberg K, et al. Fitness, fatness and clustering of cardiovascular risk factors in children from Denmark, Estonia and Portugal: the European Youth Heart Study. *International Journal Of Pediatric Obesity* 2008;3 Suppl 1:58-66. doi: 10.1080/17477160801896366

2. Anderssen SA, Cooper AR, Riddoch C, et al. Low cardiorespiratory fitness is a strong predictor for clustering of cardiovascular disease risk factors in children independent of country, age and sex. *European Journal of Cardiovascular Prevention & Rehabilitation* 2007;14(4):526-31. doi: 10.1097/HJR.0b013e328011efc1

3. Lerum Ø, Aadland E, Andersen LB, et al. Validity of noninvasive composite scores to assess cardiovascular risk in 10-year-old children. *Scandinavian Journal Of Medicine & Science In Sports* 2017. doi: 10.1111/sms.12826

4. Jorgensen T, Andersen LB, Froberg K, et al. Position statement: Testing physical condition in a population - how good are the methods? *European Journal of Sport Science* 2009;9(5):257-67. doi: 10.1080/17461390902862664

5. Leger LA, Lambert J. A maximal multistage 20-m shuttle run test to predict VO_{2max}. *European Journal of Applied Physiology and Occupational Physiology* 1982;49(1):1-12. doi: 10.1007/bf00428958

6. Leger LA, Mercier D, Gadoury C, et al. The multistage 20 metre shuttle run test for aerobic fitness. *Journal of Sports Sciences* 1988;6(2):93-101. doi: 10.1080/02640418808729800

7. Andersen LB, Andersen TE, Andersen E, et al. An intermittent running test to estimate maximal oxygen uptake: the Andersen test. *The Journal Of Sports Medicine And Physical Fitness* 2008;48(4):434-37.

8. Aadland E, Terum T, Mamen A, et al. The Andersen aerobic fitness test: reliability and validity in 10-year-old children. *Plos One* 2014;9(10):e110492-e92. doi: 10.1371/journal.pone.0110492

9. Artero EG, Espana-Romero V, Castro-Pinero J, et al. Reliability of field-based fitness tests in youth. *International Journal of Sports Medicine* 2011;32(3):159-69. doi: 10.1055/s-0030-1268488

10. Melo X, Santa-Clara H, Almeida JP, et al. Comparing several equations that predict peak VO2 using the 20-m multistage-shuttle run-test in 8-10-year-old children. *European Journal of Applied Physiology* 2011;111(5):839-49. doi: 10.1007/s00421-010-1708-z

11. Batista MB, Cyrino ES, Arruda M, et al. Validity of equations for estimating VO_{2peak} from the 20-m shuttle run test in adolescents aged 11-13 years. *Journal of Strength and Conditioning Research* 2013;27(10):2774-81. doi: 10.1519/JSC.0b013e3182815724

12. Lacour JR, Bourdin M. Factors affecting the energy cost of level running at submaximal speed. *European Journal of Applied Physiology* 2015;115(4):651-73. doi: 10.1007/s00421-015-3115-y

13. Resaland GK, Moe VF, Aadland E, et al. Active Smarter Kids (ASK): Rationale and design of a cluster-randomized controlled trial investigating the effects of daily physical activity on children's academic performance and risk factors for non-communicable diseases. *BMC Public Health* 2015;15:709-09. doi: 10.1186/s12889-015-2049-y

14. Wedderkopp N, Jespersen E, Franz C, et al. Study protocol. The Childhood Health, Activity, and Motor Performance School Study Denmark (The CHAMPS-study DK). *BMC Pediatrics* 2012;12 doi: 10.1186/1471-2431-12-128

15. Have M, Nielsen JH, Gejl AK, et al. Rationale and design of a randomized controlled trial examining the effect of classroom-based physical activity on math achievement. *BMC Public Health* 2016;16 doi: 10.1186/s12889-016-2971-7

16. Hutcheon JA, Chiolero A, Hanley JA. Random measurement error and regression dilution bias. *British Medical Journal* 2010;340 doi: 10.1136/bmj.c2289

17. McVeigh SK, Payne AC, Scott S. The reliability and validity of the 20-meter shuttle test as a predictor of peak oxygen uptake in Edinburgh school children, age 13 to 14 years. *Pediatric Exercise Science* 1995;7(1):69-79.

18. Resaland GK, Andersen LB, Mamen A, et al. Effects of a 2-year school-based daily physical activity intervention on cardiorespiratory fitness: the Sogndal school-intervention study. *Scandinavian Journal of Medicine and Science in Sports* 2011;21(2):302-9. doi: 10.1111/j.1600-0838.2009.01028.x

19. Medbø J, Mamen A, Beltrami F. Examination of the Moxus Modular Metabolic System by the Douglas-bag technique. *Applied Physiology, Nutrition & Metabolism* 2012;37(5):860-71.

20. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1(8476):307-10.

21. Ernesto C, Martins da Silva F, Pereira LA, et al. Cross validation of different equations to predict aerobic fitness by the shuttle run 20 meters test in brazilian students. *Journal of Exercise Physiology Online* 2015;18(1):46-55.

22. Burns RD, Hannon JC, Brusseau TA, et al. Cross-validation of aerobic capacity prediction models in adolescents. *Pediatric Exercise Science* 2015;27(3):404-11. doi: 10.1123/pes.2014-0175

23. Aandstad A, Holme I, Berntsen S, et al. Validity and reliability of the 20 meter shuttle run test in military personnel. *Military Medicine* 2011;176(5):513-18.

24. Green MS, Esco MR, Martin TD, et al. Crossvalidation of two 20-m shuttle-run tests for predicting VO_{2max} in female collegiate soccer players. *Journal of Strength & Conditioning Research* 2013;27(6):1520-28.

25. Hodges LD, Brodie DA, Bromley PD. Validity and reliability of selected commercially available metabolic analyzer systems. *Scandinavian Journal of Medicine & Science in Sports* 2005;15(5):271-79. doi: 10.1111/j.1600-0838.2005.00477.x

26. Hopkins WG, Schabort EJ, Hawley JA. Reliability of power in physical performance tests. *Sports Medicine* 2001;31(3):211-34.

27. Aadland E, Solbraa AK, Resaland GK, et al. Reference values for and cross-validation of time to exhaustion on a modified Balke protocol in Norwegian men and women. *Scandinavian Journal Of Medicine & Science In Sports* 2016 doi: 10.1111/sms.12750

28. Matsuzaka A, Takahashi Y, Yamazoe M, et al. Validity of the multistage 20-m shuttle-run test for Japanese children, adolescents, and adults. *Pediatric Exercise Science* 2004;16(2):113-25.

29. Barnett A, Chan LYS, Bruce IC. A preliminary study of the 20-m multistage shuttle run as a predictor of peak VO_2 in Hong Kong Chinese students. *Pediatric Exercise Science* 1993;5(1):42-50.

30. Mahar MT, Welk GJ, Rowe DA, et al. Development and validation of a regression model to estimate VO_{2peak} from pacer 20-m shuttle run performance. *Journal of Physical Activity & Health* 2006;3:S34-S46.

31. Hopkins WG. Measures of reliability in sports medicine and science. *Sports Medicine* 2000;30(1):1-15.

32. Blair SN, Kampert JB, Kohl HW, et al. Influences of cardiorespiratory fitness and other precursors on cardiovascular disease and all-cause mortality in men and women. *JAMA* 1996;276(3):205-10. doi: 10.1001/jama.276.3.205

33. Blair SN, Kohl HW, 3rd, Paffenbarger RS, Jr., et al. Physical fitness and all-cause mortality. A prospective study of healthy men and women. *JAMA* 1989;262(17):2395-401.

34. Wei M, Kampert JB, Barlow CE, et al. Relationship between low cardiorespiratory fitness and mortality in normal-weight, overweight, and obese men. *JAMA* 1999;282(16):1547.

35. Balke B, Ware RW. An experimental study of physical fitness of Air Force personnel. *United States Armed Forces medical journal* 1959;10(6):675-88.

36. Kavanagh T, Mertens DJ, Hamm LF, et al. Prediction of long-term prognosis in 12 169 men referred for cardiac rehabilitation. *Circulation* 2002;106(6):666-71. doi: 10.1161/01.cir.0000024413.15949.ed

37. Goel K, Thomas RJ, Squires RW, et al. Combined effect of cardiorespiratory fitness and adiposity on mortality in patients with coronary artery disease. *American Heart Journal* 2011;161(3):590-97. doi: 10.1016/j.ahj.2010.12.012

38. Hawley JA, Noakes TD. Peak power output predicts maximal oxygen uptake and performance time in trained cyclists. *European Journal of Applied Physiology* 1992;65(1):79-83.

39. Noakes TD, Myburgh KH, Schall R. Peak treadmill running velocity during the VO_2 max test predicts running performance. *Journal of Sports Sciences* 1990;8(1):35-45. doi: 10.1080/02640419008732129

40. Aadland E, Kvalheim OM, Rajalahti T, et al. Aerobic fitness and metabolic health in children: A clinical validation of directly measured maximal oxygen consumption versus performance measures as markers of health. *Preventive Medicine Reports* 2017 doi: 10.1016/j.pmedr.2017.05.001

Figure legends

Figure 1*. Flow chart showing the numbers of children providing data for the study. na = not available as 16-year-olds only performed one Andersen test.*

Figure 2*. Cross-validation predicted - measured mean ± 1.96 SD bias in the 20 testing datasets.*

Figure 3. *Bland Altman plot for predicted versus measured VO_{2peak} for equations including body mass in the 20 testing datasets (mean values). The full line indicate the bias; the dotted lines indicate limits of agreement.*

Supplemental Figure 1*. Bland Altman plots for the Andersen et al, 2008 and Aadland et al, 2014 equations in the 16-year-old children. The full line indicate the bias; the dotted lines are Limits of Agreement.*

***Table 1**. Children's characteristics (mean (SD)).*

	Overall	**Boys - 10 yr**	**Girls - 10 yr**	**Boys - 16 yr**	**Girls - 16 yr**
Number (%)	235	66 (28)	47 (20)	73 (31)	49 (21)
Age (years)	13.1 (2.8)	10.3 (0.3)	10.3 (0.3)	15.8 (0.3)	15.8 (0.3)
Height (cm)	158.3 (16.1)	143.8 (6.2)	142.6 (5.2)	176.3 (7.1)	166.1 (5.5)
Body mass (kg)	51.0 (15.8)	38.1 (9.5)	37.7 (6.9)	64.8 (9.8)	60.6 (11.0)
Body mass index (kg/m^2)	19.9 (3.6)	18.3 (3.6)	18.5 (2.7)	20.8 (3.0)	21.9 (3.8)
Overweight and obese (%)	17.8	19.7	21.3	13.7	18.3
Andersen test (m)	1011 (171)	912 (131)	872 (81.2)	1181 (112)	1025 (136)
VO_{2peak} (l/min)	2.86 (1.00)	2.10 (0.35)	1.88 (0.29)	4.07 (0.59)	3.00 (0.42)
VO_{2peak} (ml/kg/min)	56.1 (9.3)	56.5 (9.1)	50.5 (7.5)	63.4 (6.8)	50.1 (6.0)
HR_{peak} (beats/min)	201 (8)	201 (10)	201 (7)	201 (9)	201 (6)
RER_{peak} (ratio)	1.09 (0.07)	1.06 (0.08)	1.09 (0.06)	1.10 (0.07)	1.11 (0.07)

VO_{2peak} = peak oxygen consumption; HR_{peak} = peak heart rate; RER_{peak} = peak respiratory exchange ratio

***Table 2**. Biases when estimating equations and predicting VO_{2peak} from the first Andersen test versus the best of three Andersen tests in 10-year old children.* **Equations to predict VO_{2peak}**	**Mean (95% CI) bias predicted - measured VO_{2peak} (ml/kg/min)**	
	First test	**Best test**
First test: VO_{2peak} = 43.047 + 0.034*Andersen - 0.448*Body mass - 4.743*Sex (R^2 = 0.56, SEE = 6.05)	0.38 (-0.80–1.57)	2.17 (1.04–3.30)
Best test: VO_{2peak} = 21.843 + 0.050*Andersen - 0.356*Body mass - 3.321*Sex (R^2 = 0.62, SEE = 5.63)	-2.43 (-3.65—1.20)	0.20 (-0.90–1.30)

VO_{2peak} = peak oxygen consumption; CI = confidence interval

***Table 3**. Cross-validation performance of prediction equations. Values are means (minimum-maximum) across 20 different testing datasets.*

Equation	Body mass		Age	
	Joint	Sex-specific	Joint	Sex-specific
	Bias			
Total	-0.15 (-1.15–1.42)	-0.12 (-1.23–1.16)	-0.12 (-1.12–1.19)	0.01 (-1.24–1.42)
Boys, 10 yr	0.72 (-1.78–2.79)	0.16 (-2.81–2.85)	0.35 (-2.42–3.55)	0.06 (-2.46–3.04)
Girls, 10 yr	-1.00 (-4.28–2.98)	0.02 (-3.53–3.84)	-0.82 (-2.91–2.87)	-0.16 (-2.55–3.55)
Boys, 16 yr	-1.07 (-2.57–1.18)	-0.42 (-3.17–1.34)	-0.61 (-2.65–1.51)	0.16 (-2.23–5.25)
Girls, 16 yr	0.84 (-2.66–3.16)	-0.21 (-3.06–1.66)	0.64 (-3.14–2.79)	-0.12 (-3.83–2.02)
	Limits of Agreement			
Total	11.11 (10.29–12.29)	11.12 (10.35–12.43)	11.90 (11.11–12.90)	12.08 (11.07–14.17)
Boys, 10 yr	12.13 (10.68–13.62)	12.33 (10.66–13.92)	13.47 (11.49–15.07)	13.51 (11.45–15.15)
Girls, 10 yr	11.97 (9.60–13.31)	12.00 (9.70–13.90)	12.88 (11.58–14.50)	13.09 (11.43–14.92)
Boys, 16 yr	10.73 (7.96–12.47)	10.82 (7.96–13.29)	10.97 (8.53–12.37)	11.04 (8.55–12.58)
Girls, 16 yr	7.71 (5.84–9.41)	7.64 (6.45–10.23)	8.91 (6.59–10.92)	9.18 (7.08–12.39)
	Explained variance			
Total	0.65 (0.58–0.72)	0.65 (0.59–0.71)	0.60 (0.51–0.65)	0.59 (0.52–0.64)
Boys, 10 yr	0.56 (0.41–0.67)	0.54 (0.41–0.65)	0.45 (0.34–0.58)	0.45 (0.34–0.58)
Girls, 10 yr	0.40 (0.11–0.59)	0.41 (0.14–0.65)	0.30 (0.07–0.50)	0.30 (0.08–0.51)
Boys, 16 yr	0.43 (0.30–0.56)	0.42 (0.30–0.56)	0.39 (0.29–0.52)	0.39 (0.29–0.51)
Girls, 16 yr	0.65 (0.53–0.80)	0.66 (0.54–0.76)	0.54 (0.41–0.77)	0.54 (0.41–0.77)
	Standard error of the estimate			
Total	5.63 (5.22–6.07)	5.65 (5.24–6.06)	6.03 (5.70–6.48)	6.09 (5.68–6.44)
Boys, 10 yr	6.15 (5.27–7.05)	6.24 (5.42–7.10)	6.85 (5.94–7.77)	6.85 (5.93–7.77)
Girls, 10 yr	5.99 (4.87–6.90)	5.93 (4.70–6.78)	6.48 (5.87–7.29)	6.47 (5.90–7.31)
Boys, 16 yr	5.41 (4.08–6.42)	5.42 (4.08–6.39)	5.59 (4.41–6.39)	5.60 (4.42–6.42)
Girls, 16 yr	3.62 (3.03–4.14)	3.59 (3.06–4.16)	4.14 (3.36–5.00)	4.15 (3.36–5.01)

***Table 4**. Final sex-specific prediction equations based on the Andersen test and body mass or age*

	Equations	R^2	SEE
	Andersen test and body mass		
Boys	27.1689 + 0.0397*Andersen - 0.1698*body mass	0.56	5.82
Girls	32.5793 + 0.0309*Andersen - 0.2351*body mass	0.47	4.97
	Andersen test and age		
Boys	26.4523 + 0.0427*Andersen - 0.8553*age	0.51	6.14
Girls	29.8705 + 0.0363*Andersen - 1.0730*age	0.36	5.45

R^2 = explained variance; SEE = standard error of the estimate

***Supplemental Table 1**. Cross-validation bias of prediction equations in children performing below versus above median on the Andersen test. Values are means (SD) across 20 different testing datasets.*

Equation	**Body mass**		**Age**	
Andersen test performance	**< median**	**> median**	**< median**	**> median**
	Sex-specific equations			
Total	0.36 (5.66)	-0.56 (5.51)	0.65 (6.26)	-0.61 (5.64)
Boys, 10 yr	1.03 (6.02)	-0.59 (6.56)	1.17 (6.75)	-0.96 (6.81)
Girls, 10 yr	0.98 (6.16)	-0.97 (6.24)	0.77 (7.00)	-1.05 (6.42)
Boys, 16 yr	-0.25 (6.05)	-0.48 (5.09)	0.36 (6.19)	-0.05 (5.13)
Girls, 16 yr	-0.22 (4.00)	-0.23 (3.84)	0.26 (5.16)	-0.55 (3.76)
	Joint equations			
Total	0.31 (5.69)	-0.54 (5.53)	0.45 (6.22)	-0.67 (5.64)
Boys, 10 yr	1.71 (5.84)	-0.14 (6.50)	1.51 (6.71)	-0.74 (6.84)
Girls, 10 yr	-0.27 (6.34)	-1.71 (6.07)	-0.07 (6.95)	-1.53 (6.28)
Boys, 16 yr	-0.73 (5.95)	-1.27 (5.07)	-0.34 (6.05)	-0.87 (5.13)
Girls, 16 yr	0.45 (4.19)	1.15 (3.80)	0.71 (5.10)	0.52 (3.77)
	Aerobic fitness level			
Andersen test performance	**< median**	**> median**	**< median**	**> median**
	Andersen test		**VO_{2peak}**	
Total	925 (157)	1099 (138)	51.97 (8.38)	60.38 (8.23)
Boys, 10 yr	809 (104)	1015 (44)	51.16 (7.12)	61.94 (7.62)
Girls, 10 yr	816 (67)	929 (48)	47.48 (6.96)	53.58 (6.91)
Boys, 16 yr	1099 (89)	1266 (53)	59.71 (6.11)	67.17 (5.35)
Girls, 16 yr	925 (117)	1129 (49)	45.88 (4.85)	54.58 (3.33)