Using relative improvement over chance (RIOC) to examine agreement between tests: Three case examples using studies of developmental coordination disorder (DCD) in children

John Cairney [a,b,c,d,*], David L. Streiner [b,c,d]

[a] Departments of Family Medicine, Clinical Epidemiology & Biostatistics and Kinesiology; Offord Centre for Child Studies and the CanChild Centre for Childhood Disability Research, McMaster University, Hamilton, ON, Canada
[b] Health Systems Research and Consulting Unit, Centre for Addiction & Mental Health, Toronto, ON, Canada
[c] Department of Psychiatry and Behavioural Neuroscience, McMaster University, Hamilton, ON, Canada
[d] Department of Psychiatry, University of Toronto, Toronto, ON, Canada

ABSTRACT

Although statistics such as kappa and phi are commonly used to assess agreement between tests, in situations where the base rate of a disorder in a population is low or high, these statistics tend to underestimate actual agreement. This can occur even if the tests are good and the classification of subjects is adequate. Relative improvement over chance or RIOC is a statistic that can correct for this bias; however, it is not yet commonly used in the health sciences or disability fields. In this paper, we introduce RIOC and demonstrate its application using the results from 3 previously published studies, all of which assessed the agreement between tests commonly used to identify children with a neurodevelopmental disorder known as developmental coordination disorder (or DCD). The results illustrate the differences between kappa and RIOC under conditions where the distribution of scores in a 2×2 table is unbalanced. The introduction of this statistic in the area of developmental disabilities research is encouraged.

© 2010 Published by Elsevier Ltd.

1. Introduction

While statistics such as kappa and phi have many desirable properties for examining agreement between raters or scales (e.g., easily interpretable; compares, simultaneously, individuals who are correctly and incorrectly allocated), they do not correct for what has been described as the problem of maximum ceilings in 2×2 tables (Copas & Loeber, 1990). The source of the problem is the discrepancy between the base rate (i.e., the proportion of individuals who have a specific condition/disorder), and the selection ratio (i.e., the proportion of individuals identified as having a specific condition/disorder based on the results of a test). The mismatch between base rate and selection ratio results in significant numbers of subjects who cannot be correctly predicted in the analysis; the greater the discrepancy, the greater the number of subjects that cannot be correctly predicted (Copas & Loeber, 1990; Loeber & Dishion, 1983). The problem places significant limits on indices of agreement. Practically, this problem is commonly found in situations where the base rate of a disorder in the sample is low (e.g., population-based studies)

---

* Corresponding author at: Department of Family Medicine, 175 Longwood Drive, Suite 210A, Hamilton, ON L8P 0A1, Canada.
Tel.: +1 905 521 2100x28506; fax: +1 905 905 527 4440.
E-mail address: cairnej@mcmaster.ca (J. Cairney).

or high (e.g., clinic samples). Under such conditions, even if the tests are good and classification based on test results is correct, measures such as kappa and phi will show poor agreement between the test or raters.

Relative improvement over chance (RIOC) corrects for the mismatch between base rate and selection ratio (Copas & Loeber, 1990; Loeber & Dishion, 1983), so it is a particularly useful measure of predictive efficiency in precisely those circumstances where the base rate of condition is low or high.

Despite the advantages of using RIOC, its application in the health sciences literature is surprisingly limited. In the area of developmental disabilities, the use of RIOC is especially pertinent given the relatively low prevalence of many disorders of interest (e.g., autism, pervasive developmental delay, and intellectual disability). In this paper, we use one such developmental disorder, developmental coordination disorder (DCD), to illustrate the potential application of RIOC. In particular, we explore the use of RIOC under conditions where low prevalence is apt to be a particularly vexing problem for assessing agreement between tests – screening for DCD in the general population, and in clinical settings where the presentation of problems is non-specific.

DCD is a neurodevelopmental disorder that affects approximately 5% to 6% of all school-aged children (Gibbs, Appleton, & Appleton, 2007; Kadesjo & Gillberg, 1999). Children with DCD are significantly behind their peers with regard to gross and/or fine motor skill development (Visser, 2003). These problems notably impair functioning in everyday activities and negatively affect school achievement, particularly in areas such as handwriting (Missiuna, Moll, Law, King, & King, 2006) and physical education (Cairney, Hay, Mandigo, Wade, Faught, & Flouris, 2007a; Cairney, Veldhuizen, Kurdyak, Missiuna, Hay, & Faught, 2007b). A diagnosis of DCD is made after ruling out other chronic neurological conditions (e.g., cerebral palsy) as a cause of motor impairment, and when the intellectual abilities of the child are taken into consideration (American Psychiatric Association, 2000). As in many areas of developmental disabilities, there is no gold standard for assessing DCD, so not surprisingly, a significant number of studies have explored the concurrent validity of different tests of motor proficiency and impaired coordination for identifying children with this condition. Because the prevalence of DCD is low, studies examining the association between tests to identify children with the disorder will potentially suffer from the problems of low base rate identified previously. Yet, we could locate only one study (Spironello, Hay, Missiuna, Faught, & Cairney, 2010) that uses RIOC as an alternative to statistics such as kappa, when assessing agreement between tests.

In order to introduce the RIOC statistic into the research literature in this field, we will re-examine previously published data in the area, examining the difference between RIOC and measures such as kappa for significantly unbalanced marginal totals in the analysis of 2×2 tables.

2. Methods

In this study, we use data derived from three previously published studies (Cairney et al., 2007a, 2007b; Green et al., 2005; Schoemaker, Flapper, Verheij, Wilson, Reinders-Messelink, & de Kloet, 2006). With regard to the last two studies, data were extracted from the tables provided in the publications and re-analyzed for the purposes of this investigation. The first study, described in detail elsewhere (Schoemaker et al., 2006), collected data from 322 Flemish children to examine agreement between the developmental coordination disorder questionnaire (DCD-Q; Wilson, Kaplan, Crawford, Campbell, & Dewey, 2000), a pen and paper test where parents are asked to rate their children's abilities in relation to a variety of everyday, motor-based tasks, against the Movement Assessment Battery for Children (M-ABC; Henderson & Sugden, 2007), a standardized test to assess coordination impairment administered directly to children. The study drew a school-based sample from a non-treated (general) population of children. The second study, also described in detail elsewhere (Green et al., 2005), compared the DCD-Q to an assessment performed by occupational therapists to diagnose motor coordination difficulties. The sample for this study was derived from a list of children waiting to receive occupational therapy in the United Kingdom.

For the third study, access to the primary data was obtained to conduct a new analysis involving both the calculation of kappa and RIOC for comparative purposes. The study, described in detail elsewhere (Cairney et al., 2007a, 2007b), used a large sample of children ($n = 540$) from Ontario, Canada to examine the agreement between children's self-perceptions of adequacy in and predilection for physical activity (CSAPPA; Hay, 1992) against another commonly used motor test to assess DCD, the short form of the Bruininks–Oseretsky Test of Motor Proficiency (BOTMP-SF; Bruininks, 1978). This study used a receiver-operator characteristics (ROC) analysis to examine the association between the CSAPPA and diagnosis of probable DCD based on the BOTMP-SF. Based on the results of the third study, we derived an optimal cut-point (balance between false positives and false negatives) on the CSAPPA to compare case agreement between this instrument and the diagnostic measure based on a 2×2 table analysis.

These studies were selected because they represent tools that are commonly used in the DCD literature to screen for the condition – CSAPPA and DCD-Q – and the studies were conducted by recognized experts in the field of DCD. Moreover, because DCD is a low prevalence condition in the population and because these tools are used to screen for the condition, it is likely that these studies would have significantly unbalanced 2×2 tables, and would therefore serve as useful exemplars to demonstrate the application of RIOC. Two of the articles were found by searching Medline using the key search terms DCD, DCD-Q and screening. The article which uses the CSAPPA was conducted by one of the authors (JC). The results of the predictive efficacy of the instruments in question have been reported for all of these studies before; however, the RIOC was not reported in any of these publications. It is important to note, however, that other studies comparing these tools to criterion measures are available in the published literature (e.g., Faught et al., 2008; Loh, Piek, & Barrett, 2009; Schoemaker,