An evidence-based approach to the creation of normative data: base rates of impaired scores within a brief neuropsychological battery argue for age corrections, but against corrections for medical conditions

Megan E. O’Connell, Holly Tuokko, Stacey Voll, Martine Simard, Lauren E. Griffith, Vanessa Taler, Christina Wolfson, Susan Kirkland & Parminder Raina

To cite this article: Megan E. O’Connell, Holly Tuokko, Stacey Voll, Martine Simard, Lauren E. Griffith, Vanessa Taler, Christina Wolfson, Susan Kirkland & Parminder Raina (2017) An evidence-based approach to the creation of normative data: base rates of impaired scores within a brief neuropsychological battery argue for age corrections, but against corrections for medical conditions, The Clinical Neuropsychologist, 31:6-7, 1188-1203, DOI: 10.1080/13854046.2017.1349931

To link to this article: https://doi.org/10.1080/13854046.2017.1349931

Published online: 05 Jul 2017.
An evidence-based approach to the creation of normative data: base rates of impaired scores within a brief neuropsychological battery argue for age corrections, but against corrections for medical conditions

Megan E. O’Connell\textsuperscript{a}, Holly Tuokko\textsuperscript{b}, Stacey Voll\textsuperscript{b}, Martine Simard\textsuperscript{c}, Lauren E. Griffith\textsuperscript{d}, Vanessa Talere\textsuperscript{e}, Christina Wolfson\textsuperscript{f}, Susan Kirkland\textsuperscript{g} and Parminder Raina\textsuperscript{h}

\textsuperscript{a}Department of Psychology, University of Saskatchewan, Saskatoon, Canada; \textsuperscript{b}Institute on Aging & Lifelong Health, University of Victoria, Victoria, Canada; \textsuperscript{c}School of Psychology, Laval University and Centre de recherche de l’Institut universitaire en santé mentale de Québec, Québec City, Canada; \textsuperscript{d}Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Canada; \textsuperscript{e}School of Psychology, University of Ottawa & Bruyère Research Institute, Ottawa, Canada; \textsuperscript{f}Department of Epidemiology and Biostatistics and Occupational Health, McGill University, Montreal, Canada; \textsuperscript{g}Department of Community Health and Epidemiology, Dalhousie University, Halifax, Canada; \textsuperscript{h}Faculty of Health Sciences, Department of Health Research Methods, Evidence and Impact, McMaster Institute for Research on Aging & Labarge Centre for Mobility in Aging, McMaster University, Hamilton, Canada

\textbf{ABSTRACT}

\textbf{Objective:} We detail a new approach to the creation of normative data for neuropsychological tests. The traditional approach to normative data creation is to make demographic adjustments based on observations of correlations between single neuropsychological tests and selected demographic variables. We argue, however, that this does not describe the implications for clinical practice, such as increased likelihood of misclassification of cognitive impairment, nor does it elucidate the impact on decision-making with a neuropsychological battery. \textbf{Method:} We propose base rate analyses; specifically, differential base rates of impaired scores between theoretical and actual base rates as the basis for decisions to create demographic adjustments within normative data. Differential base rates empirically describe the potential clinical implications of failing to create an appropriate normative group. We demonstrate this approach with data from a short telephone-administered neuropsychological battery given to a large, neurologically healthy sample aged 45–85 years old. We explored whether adjustments for age and medical conditions were warranted based on differential base rates of spuriously impaired scores. \textbf{Conclusions:} Theoretical base rates underestimated the frequency of impaired scores in older adults and overestimated the frequency of impaired scores in younger adults, providing an evidence base for the creation of age-corrected normative data. In contrast, the number of medical conditions (numerous cardiovascular, hormonal, and metabolic conditions) was not related to differential base rates of impaired scores. Despite a

\textbf{ARTICLE HISTORY}

Received 14 January 2017
Accepted 27 June 2017

\textbf{KEYWORDS}

Normative data; base rates; demographic adjustments; medical conditions; CLSA

\textbf{CONTACT} Megan E. O’Connell \textsuperscript{a} megan.oconnell@usask.ca

This article was originally published with errors. This version has been corrected. Please see Corrigendum (https://doi.org/10.1080/13854046.2017.1373979).

© 2017 Informa UK Limited, trading as Taylor & Francis Group
small correlation between number of medical conditions and each neuropsychological variable, normative adjustments for number of medical conditions does not appear warranted. Implications for creation of normative data are discussed.

Comparison of an individual’s performance with an appropriate normative group is critical to the interpretation of neuropsychological test performance. The salient questions, however, are what is an appropriate normative group, and should these normative data be corrected for demographic variables? Creation of normative data with demographic corrections lack an evidence base: the decision to create age corrections, for example, is typically based on associations between age and cognitive performance. Normative corrections for demographic variables have consequences and should be evidence based. Corrections for age reduce the dementia sensitivity of cognitive tests because age is a risk factor for dementia (O’Connell & Tuokko, 2010). Reduced sensitivity has implications for clinically useful predictive values, depending on base rates of conditions within different clinical contexts. A similar argument can be made for the use of corrections based on medical corrections: single and comorbid medical conditions are increasingly described as risk factors for dementia (e.g. Kelaiditi et al., 2013; Xu, Xiao, Rahardjo, & Hogervorst, 2015; Vassilaki et al., 2015). Therefore, decisions to create normative corrections for medical conditions such as suggested by Bergman and Almkvist (2015), could have negative consequences for the diagnostic utility of cognitive tests. In this paper, we provide illustrative examples of a novel approach to decision-making about whether to create normative data that are corrected for demographic and medical variables.

Foremost, our approach is focused on clinical relevance – what are the implications of failing to create demographic/medical corrections in normative data for classification of cognitive impairment? We argue that decisions regarding an appropriate normative group should be based on evidence that failing to create an appropriate normative group would result in cognitively normal persons being incorrectly identified as cognitively impaired (i.e. spurious impairment, false positive errors). Further, we sought to explore these clinical implications within the context of a neuropsychological battery – were base rates of impaired scores in a cognitively healthy sample different from theoretical base rates for different subgroups based on demographic/medical variables? We propose baserate analyses – specifically, differential base rates of impaired scores – as the basis for decisions to create demographic adjustments within normative data. Differential base rates refer to differences between theoretical base rates, which are calculated from intercorrelations of cognitive test scores, and actual base rates of impaired scores seen in a battery of co-administered tests in a cognitively healthy sample. If differential base rates exist for demographic variables, such as age, the age-based differences in base rates empirically demonstrates the potential clinical implications of failing to create an appropriate normative group based on age; failing to account for age would result in incorrect interpretation of impaired scores across a neuropsychological battery.

Demographic variables, such as age, are associated with cognitive test performance (e.g. Bank, Yochim, MacNeill, & Lichtenberg, 2000; Marcopulos, McLain, & Giuliano, 1997; Nabors, Vangel, Lichtenberg, & Walsh, 1997; Salthouse, 2012; Youngjohn, Larrabee, & Crook, 1993). Consequently, many normative sources provide demographic adjustments via normative tables or regression formulas (see resources such as Mitrushina, Boone, Razani, & D’Elia, 2005; Strauss, Sherman, & Spreen, 2006). Is merely finding an association between a demographic
variable and a score from a single neuropsychological test (hereafter referred to as the ‘descriptive approach’) sufficient evidence for the creation of demographically stratified normative data? We argue that to address the magnitude of demographic/neuropsychological test score association required for neuropsychological test correction, the descriptive approach to normative data creation may not suffice. Moreover, a descriptive approach to the creation of demographically stratified normative data does not explore the impact of demographic adjustments on neuropsychological interpretation. On occasion, single raw test scores are revealed to have superior predictive value relative to demographically corrected scores (e.g. O’Connell, Tuokko, Graves, & Kadlec, 2004; Silverberg & Millis, 2009). When, then, are demographically corrected scores preferred?

The impact of uncorrected and corrected neuropsychological test scores has been explored for diagnostic accuracy. Within the context of dementia, high rates of misclassifications of impairment have been demonstrated for persons with advanced age (e.g. Adams, Boake, & Crain, 1982; Ainslie & Murden, 1993; Belle et al., 1996; Bornstein, 1986; Jagger, Clarke, & Anderson, 1992; Tombaugh, McDowell, Kristjansson, & Hubley, 1996). Misclassification of cognitive impairment is a cogent argument for the creation of demographically corrected normative data. Age-corrected test scores increased diagnostic specificity, but sensitivity was decreased (e.g. O’Connell & Tuokko, 2010). In O’Connell and Tuokko’s (2010) Monte Carlo simulation study supported by clinical data, the reduction in sensitivity for age-corrected normative data was dependent on the magnitude of association between age and dementia: modeling age to become a stronger risk factor for dementia further reduced the sensitivity of the age-corrected scores.

Although useful theoretically, accuracy statistics (e.g. sensitivity/specificity and area under the receiver operating characteristic, or ROC, curve) are limited in their utility across disorders, depend on base rates of conditions for predictive values, and, more problematically, are limited by a general lack of reporting of accuracy statistics in neuropsychology. Accuracy statistics are useful in choosing tests to incorporate into a battery, but they are based on interpretation at a single-test level, which has limited utility in the context of neuropsychological assessment. Most neuropsychologists use a battery approach to assessment (Rabin, Barr, & Burton, 2005). Baserate analysis is another approach to evidence-based interpretation, which, we argue, has broader interpretative appeal because it provides information inherent in the neuropsychological battery itself. Furthermore, baserate analyses are not limited in the same way as accuracy statistics: base rates are not dependent on how the test scores will be used, whereas evidence for diagnostic accuracy is limited to the data on diagnosis for the condition of interest, such as dementia.

Informal baserate analyses have always been integral to practice in clinical neuropsychology. For example, attempts to account for spuriously impaired neuropsychological scores (i.e. false positives) is typically performed with two methods: (1) use of redundancy in the choice of tests for a neuropsychological battery, with evidence for impairment weighted as strongest when seen across tests measuring the same domain, and (2) use of the convention from statistical interpretation that, if the cutoff for impairment were set at the 5th percentile, every test had a 5% chance (i.e. probability of a Type I error = .05) to be recorded as impaired even in the absence of impairment, and the probability of spurious impairment increased as the number of tests increased (Binder, Iverson, & Brooks, 2009; Decker, Schneider, & Hale, 2012; Schretlen, Testa, Winicki, Pearlson, & Gordon, 2008). Formal Bonferroni corrections for multiple testing (i.e. significance cutoff set at α/number of tests) were shown to be overly
cautious (Huizenga, Smeding, Grasman, & Schmand, 2007; Ingraham & Aiken, 1996), and estimating base rates a priori given $\alpha$ and the number of tests (Ingraham & Aiken, 1996) was found to be inaccurate (Crawford, Garthwaite, & Gault, 2007; Decker et al., 2012; Schretlen et al., 2008).

Moreover, spurious impairment may not simply follow the pattern of increased base rates of impaired scores with an increase in the number of tests, at least for tests of suboptimal effort (i.e. tests of performance validity; Davis & Millis, 2014). It has been postulated that this finding with tests of suboptimal effort may be related to the atypical characteristic of these tests, such as highly skewed distributions (Larrabee, 2014), which along with kurtosis impact theoretical estimates of base rates (Decker et al., 2012). Despite some unknowns in baserate corrections, ignoring the ‘high’ frequency of spuriously impaired scores results in misdiagnosis (Brooks, Iverson, Holdnack, & Feldman, 2008), overestimation of cognitive impairment (Binder et al., 2009; Gisslen, Price, & Nilsson, 2011), and overestimation of suboptimal performance (Odland, Lammy, Martin, Grote, & Mittenberg, 2015).

More recently, authors have used Monte Carlo simulations (the approach was first proposed by Crawford et al., 2007 and has been used frequently and supported; e.g. Decker et al., 2012; Schretlen et al., 2008) and careful descriptions of impairment in normative neuropsychological batteries (Brooks, Holdnack, & Iverson, 2011; Brooks & Iverson, 2010; Crawford, Garthwaite, Longman, & Batty, 2012; Mistridis et al., 2015). The theoretical and actual baserate data provide higher empirical quality relative to the conventional approach to the probability of Type I error, and provide clinically useful evidence of the base rates of spuriously impaired scores. These data have repeatedly demonstrated the higher frequency of spuriously impaired scores when a more lenient criterion for cutoff for impairment is used (Binder et al., 2009; Brooks et al., 2011; Brooks & Iverson, 2010; Brooks et al., 2008; Mistridis et al., 2015; Schretlen et al., 2008). Base rates of impaired scores can improve interpretation of a neuropsychological battery when considering single test interpretation within a battery (Brooks et al., 2011; Brooks & Iverson, 2010; Crawford et al., 2007, 2012; Huizenga, van Rentergem, Grasman, Muslimovic, & Schmand, 2016; Mistridis et al., 2015; Schretlen et al., 2008).

We argue that baserate analyses are clinically useful as the empirical basis for creation of normative data. Schretlen and colleagues (2008) detailed theoretical and actual base rates of impaired scores across a variety of neuropsychological batteries (based on $T$-scores < 40, <35, and < 30). They found that both theoretical (Monte Carlo) and actual base rates depended on demographic variables and estimated premorbid IQ. Importantly, they found that correcting the single tests within the battery for these demographic variables and IQ resulted in no association between demographic variables and IQ with base rates of impaired scores across the battery (i.e. spuriously impaired scores remained, but these were no longer associated with individual difference variables). Schretlen and colleagues (2008) postulated that the lack of association between demographic variables and IQ with base rates of impaired scores was stronger evidence that the impaired scores were due to chance, and accounting for these chance events could, therefore, be diagnostically useful.

Brooks and Iverson (2010) detail a disconnect between theoretical base rates from Monte Carlo estimations (Crawford et al., 2007) and actual base rates of low scores (on the Neuropsychological Assessment Battery, NAB, and the Wechsler Adult Memory Scale – III; low scores as one or more low index score below the 10th percentile) for those with low IQ. In this cognitively healthy sample described by Brooke and Iverson (2010), it could be argued that IQ reflects, at least to some degree, premorbid cognitive status. Brooks and Iverson
found that theoretical base rates of low scores were overestimated for those with Wechsler Adult Intelligence Scale – III IQ scores in the average to superior range, but underestimated for those with low average to unusually low IQ scores. In their study, the theoretical base rate underestimates of low scores was marked: 67% of those with unusually low IQs would have been misclassified as impaired on the NAB, if the theoretical base rates alone were used for interpretation (i.e. 96% had one or more low scores, but theoretically 29% were estimated to have one or more low scores). We interpret this finding as supporting modification of the criterion for impairment on the NAB based on IQ (in addition to the demographic variable of age which is already adjusted for in the current normative data). In other words, because the estimated base rates differed from the actual base rates for IQ, we argue this is evidence that NAB normative data should be corrected for the moderator variable of IQ (or premorbid cognitive status) to minimize misclassification rates.

We believe that differential base rates of impaired scores between theoretical and actual base rates is a strong empirical argument for creation of an appropriate normative group (i.e. normative data stratified by important variables such as age). A descriptive approach (i.e. single test correlation with demographic or other individual difference variable) to normative data creation does not speak to the implications for clinical practice, such as likelihood of misclassification, nor does it elucidate the impact on decision-making for a neuropsychological battery. Differential base rates, however, empirically detail the potential clinical implications of failing to create an appropriate normative group. In many ways, this is similar to calls for patterns of analysis across batteries based on normative data (Schretlen, Munro, Anthony, & Pearlson, 2003), and a recently published multivariate normative approach (Huizenga et al., 2016).

As we prepare to create normative data for the brief Canadian Longitudinal Study on Aging (CLSA) telephone administered neuropsychological battery (Tuokko, Griffith, Simard, & Taler, 2017), we decided to explore the of use an evidence-based approach to decisions for demographic corrections and used demographic corrections for age as one illustrative example. As was discussed previously, age is a common demographic correction within normative data due to high rates of misclassification for those with advanced age (e.g. Adams et al., 1982; Ainslie & Murden, 1993; Belle et al., 1996; Bornstein, 1986; Jagger et al., 1992; Tombaugh et al., 1996). Age corrections reduce dementia diagnostic sensitivity because age is a risk factor for dementia, but within a battery approach this has fewer clinical implications because use of a battery increases sensitivity (O’Connell & Tuokko, 2010). Consequently, exploration of the impact of age on differential base rates of impaired scores within a battery approach is important.

In addition to exploring the evidence base for creating normative corrections for age, we explored the possibility of normative corrections for medical conditions. Medical conditions, like age, might be risk factors for dementia (e.g. Kelaiditi et al., 2013; Xu et al., 2015; Vassilaki et al., 2015). Moreover, like age, medical conditions (such as hypertension, diabetes, chronic obstructive pulmonary disease, renal dysfunction, hypo- and hyper-thyroidism) appear to be associated with cognitive test performance (Ahles, Root, & Ryan, 2012; Akintola et al., 2015; Anand, Johansen, & Kurella Tamura, 2014; Andreotti, Root, Ahles, McEwen, & Compas, 2015; Chen et al., 2016; Hogervorst, Huppert, Matthews, & Brayne, 2008; Hogervorst et al., 2008; Jutkowitz et al., 2016; Kalmijn et al., 2000; Piotrowicz et al., 2016; Schou, Ostergaard, Rasmussen, Rydahl-Hansen, & Phanareth, 2012; van Osch, Hogervorst, Combrinck, & Smith, 2004; Yaffe et al., 2010, 2014). Multiple co-existing medical conditions are also associated
with reduced cognition (Viscogliosi, Chiriac, Andreozzi, & Ettorre, 2016; Downer, Raji, & Markides, 2016; Fabbri et al., 2016; Aarts et al., 2011).

In our illustrative examples, we expected different rates of impaired scores between theoretical and actual base rates for different age groups and for groups based on number of medical conditions. Support for these hypotheses will provide an evidence base that we need to create age-adjusted normative data and normative data based on number of concomitant medical conditions.

**Method**

**Participants**

The present study uses data collected from ‘Tracking’ participants [version 3.0] during the baseline wave of the CLSA (Raina et al., 2009). The CLSA recruited a large national sample of 51,425 men and women between the ages of 45 and 85 across 10 Canadian provinces; assessments were conducted in either English or French. The ‘Tracking’ participants comprised a nationally representative sample of 21,241 individuals who provided information via a 60-min telephone interview concerning personal demographics, social, physical/clinical, and psychological functioning, economic status, and health services that are relevant to health and aging. The overall purpose of the CLSA is to examine the complex interplay among determinants of health transitions and trajectories and the dynamic processes of adult development and aging.

For the purpose of this paper, inclusion/exclusion criteria were as follows:

1. No self-reported neurological disease: participants with self-reported cerebrovascular accident (CVA or stroke) \(n = 390\); transient ischemic attack (TIA or mini-stroke) \(n = 748\); memory problems \(n = 449\); dementia or Alzheimer’s disease \(n = 43\); parkinsonism or Parkinson’s disease \(n = 78\); multiple sclerosis (MS) \(n = 141\); epilepsy \(n = 166\) were excluded.

2. English-only: cognitive measures that were both administered and responded to in English were retained. French-speaking participants \(n = 2056\) or participants who switched between languages were excluded \(n = 824\) were excluded.

3. Completion of all four cognitive measures administered: Rey Auditory Verbal Learning Test (Rey, 1964), trial 1 recall (RAVLT-trial 1) and a 5-min delayed recall (RAVLT-delayed), the Mental Alternation Test (MAT; Teng, 1994) and Animal Fluency (AF; Read, 1987); \(n = 3987\) were excluded due to incomplete measures.

This resulted in the sample for this study: self-reported neurologically normal, English-speaking, participants who completed all four cognitive measures administered \(n = 13,149\).

**Measures**

Measures were selected from CLSA ‘Tracking’ data [version 3.0] for use in the paper. The administration and scoring for measures of cognition (i.e. RAVLT Trial 1 and RAVLT 5-min delayed recall, MAT and AF) have been described elsewhere (Tuokko et al., 2017). While these measures have been used in other research, the manner in which the measures were administered and scored within the CLSA ‘Tracking’ differed from most other studies: they were
administered over the telephone and in a specific sequence embedded within a broader set of questions concerning health and social functioning. In addition, for animal fluency (AF) two scoring methods (i.e. AF1, a stricter, and AF2, a more lenient scoring method) were provided.

Participants reported the presence/absence of specific chronic medical conditions (i.e. conditions that are expected to last or have lasted longer than 6 months, with the exception of myocardial infarctions) when asked ‘Has a doctor ever told you that you have …?’ The following self-reported medical conditions were selected because the extant literature suggests possible impact on cognitive functioning: High blood pressure (HBP)/hypertension ($n = 4751$); diabetes/borderline diabetes/blood sugar too high ($n = 2126$); cancer ($n = 2034$); under-active thyroid gland/hypothyroidism/myxedema ($n = 1469$); over-active thyroid gland/hyperthyroidism/Grave’s disease ($n = 264$); chronic obstructive pulmonary disease (COPD)/emphysema/chronic bronchitis ($n = 820$); and kidney disease/failure ($n = 318$). A variable for ‘cardiac chronic conditions’ ($n = 1510$) was created where response to any of multiple cardiac-related conditions were tallied only once: (i) heart disease/congestive heart failure (CHF) ($n = 1105$), (ii) myocardial infarction (MI)/heart attack/acute myocardial infarction (AMI) ($n = 564$), or (iii) angina/chest pain due to heart disease ($n = 713$). In doing so, we were able to avoid the complications that comes with self-reported heart conditions – multiple submissions for one true variable.

The summation of these self-reported medical conditions was a simple tally.

**Statistical procedure**

All statistical procedures were based on the method of least-squares method (i.e. parametric); consequently, assumptions of normality and homoscedasticity of the residuals were critical for all analyses. Theoretical baserate analyses and associations between cognitive tests with age and number of medical conditions were performed with Pearson $r$, and associations between cognitive test scores with presence vs. absence of specific medical conditions were performed with a point-biserial correlation ($r_{bp}$, appropriate for one dichotomous and one continuous variable).

In very large sample sizes such as the one in CLSA even small effects might be statistically significant; a main criticism of the null hypothesis significance test approach to statistics (Nickerson, 2000). Despite the lack of extreme skew or kurtosis in our data, (see final two columns in Table 1) and our analysis of the histograms tending to normality, our tests of normality/homogeneity of variance were statistically significant due to the large degrees of freedom. Fortunately, the very large sample size affords leniency under the Central Limit Theorem, such that even if the distributions were not normal, we would not violate assumptions of least-squares statistics (see Lumley, Diehr, Emerson, & Chen, 2002 for a discussion of the unique challenges of statistics with large data-sets).

With respect to homogeneity of variance (with 1 df), which is the underlying assumption for the $r_{bp}$ reported for each cognitive test with each medical condition (Table 2), homogeneity of variance (Levene’s test with alpha .05) tended to be violated for RAVLT delayed, for HBP, and for angina. Violations, particularly for tests with fewer participants and asymmetrical cell sizes, as seen here with hyperthyroidism (Table 2), limit the usefulness of the $r_{pb}$ It is important to stress, however, that individual medical conditions were not used for any
subsequent analyses; rather, the tallies were used where the sample sizes are large and therefore robust.

Theoretical base rates were obtained using the method detailed by Crawford and colleagues (2007), and is available online: https://homepages.abdn.ac.uk/j.crawford/pages/
The Monte Carlo program estimates the frequency of low scores based on inter-test Pearson $r$ correlations, which were obtained from the above described sample and including all five scorings (because AF has two possible scoring procedures, AF1-stricter scoring and AF2-lenient scoring, whereas other tests have only one scoring procedure) (see Table 2).

Actual base rates were obtained from the distribution of raw scores. The overall English-speaking, neurologically healthy sample was used to create the $M$ and $SD$ for each cognitive test, and $z$-score distributions, population parameters, were used. We consider this $z$-score approach to be appropriate given the large sample size, consistent with cautions for procedures in small samples outlined by Crawford and Garthwaite (2002).

For theoretical and actual base rates, scores corresponding to the 5th percentile were compared. For simplicity, we have provided the base rates for the 5th percentile, which is most likely to be used clinically: rarely do clinicians need to use the extremely conservative 1st percentile cutoff in evidence-based practice using base rates analysis to interpret results from a battery. The other cutoffs for impairment provided by Crawford et al.'s (2007) formula (i.e. 15th and 25th percentile) are higher than would be recommended for clinical use.

For specific comparisons between theoretical and actual base rates, the above procedure was repeated, but restricted to specific sub-samples: youngest versus oldest age groups; or groups based on participants with zero, one, or two or more medical conditions. For example, the intercorrelations between tests were performed for the young sample only, to obtain theoretical base rates of impaired scores. Then, for example, actual base rates of impaired scores were explored for this young sample. All analyses were conducted with the operational definition of impaired scores as being below the 5th percentile. Percentages of the sample demonstrating no impaired scores versus one or more impaired scores are presented for age groups (Table 3) and for medical condition groups (0, 1, or 2 + medical conditions) (Table 4).

Chi-square analyses, with percentages of impaired scores (see Tables 3 and 4), were used to determine if the theoretical and actual base rates differed significantly across age ($df = 1$) and across medical condition groups ($df = 2$).

### Table 3. Theoretical and actual base rates (%) of impaired scores, at the 5th percentile, for the youngest and oldest age groups.

<table>
<thead>
<tr>
<th>Number of impaired scores</th>
<th>Youngest (45–54) ($n = 3813$)</th>
<th>Oldest (75–85) ($n = 2389$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Actual</td>
</tr>
<tr>
<td>0</td>
<td>75.1</td>
<td>90.9</td>
</tr>
<tr>
<td>1 or more impaired score</td>
<td>24.9</td>
<td>9.1</td>
</tr>
</tbody>
</table>

*actual base rates of impaired scores exceed theoretical base rates of impaired scores.

### Table 4. Theoretical and actual base rates (%) of impaired scores, at the 5th percentile, for 3 groups based on sums of self-reported medical conditions of interest.

<table>
<thead>
<tr>
<th>Number of Impaired Scores</th>
<th>0 medical conditions ($n = 5348$)</th>
<th>1 medical condition ($n = 4071$)</th>
<th>2 or more medical conditions ($n = 3730$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Actual</td>
<td>Estimate</td>
</tr>
<tr>
<td>0</td>
<td>75.0</td>
<td>88.7</td>
<td>74.9</td>
</tr>
<tr>
<td>1 or more impaired score</td>
<td>25.0</td>
<td>11.3</td>
<td>25.1</td>
</tr>
</tbody>
</table>
Results

Theoretical base rates of impaired scores (reported in percentages in Table 3) were similar for younger and older adults ($\chi$ probability = .98), but actual base rates differed for younger and older adults ($\chi$ probability < .01). As can be seen in Table 3, the theoretical base rates overestimated actual base rates of impaired scores for younger adults, and the theoretical base rates underestimated actual base rates of impaired scores for older adults. These data suggest that small to moderate correlations between age and each neuropsychological test score (see Table 2) have implications for misclassification. This finding constitutes evidence that the cognitive test scores should be corrected for age.

When considering medical conditions, however, theoretical ($\chi$ probability = 1.00) and actual ($\chi$ probability = .16) base rates of impaired scores were similar (see Table 4). Across all groups, irrespective of number of medical conditions, the theoretical base rates overestimate actual base rates. This is an important finding, given the associations between the number of medical conditions and each cognitive variable: clearly the small correlations seen in Table 2 do not have clinical importance. That is, people with multiple medical conditions are no more likely to be misclassified as having impaired cognitive performance on this battery than people who do not have medical conditions.

Discussion

We proposed that demographic adjustments within normative data should only occur when unadjusted scores would misclassify cognitive impairment, which has implications for clinical diagnosis. Most neuropsychologists interpret impairment within the context of a battery approach (Rabin et al., 2005); therefore, evidence for misclassification of cognitive impairment should be at the level of a battery, versus focusing on interpretation for individual neuropsychological tests. Differences between theoretical and actual base rates of spuriously impaired scores on the Neuropsychological Assessment Battery (NAB) have been demonstrated based on IQ (Brooks & Iverson, 2010). Schretlen and colleagues (2008) demonstrated that base rates of impaired scores were associated with demographic variables and estimated IQ. Base rates of spuriously impaired scores were no longer associated with demographic variables and estimated IQ once the individual tests in the battery were adjusted for these individual differences (Schretlen et al., 2008). We argue that differential base rates between theoretical and actual base rates indicate the clinical implications of failing to create an appropriate normative group. Unsurprisingly, we found that age was associated with differential base rates of impaired scores. Consequently, we recommend these normative data be adjusted for age. In contrast, number of medical conditions was not associated with differential base rates of impaired scores and, therefore, we do not recommend that normative adjustments be created for medical conditions.

This study not only provides a new framework for the creation of normative data, but also highlights the limitations of simply using correlations between cognitive test performance and individual difference variables as the basis for normative adjustments. We demonstrated that the magnitude of these associations is important to consider. The magnitude of association between the single test scores and age, for example, was small to modest, ranging from .24 to .33 (using effect size descriptors from Cohen, 1988). Nevertheless, there were clear implications when considering performance across the battery of tests: younger
cognitively healthy adults were less likely to be classified as impaired, and those with advanced age were more likely to be classified as impaired. The small magnitude of association between the number of medical conditions and neuropsychological test performance (ranging from .14 to .18) did not have implications for misclassification. That is, people with multiple medical conditions were no more likely to demonstrate impaired cognitive performance than people without medical conditions.

Although Cohen (1988) provided effect size (ES) descriptors, he urged caution: ‘… the author proposes, as a convention, ES values to serve as operational definitions of the qualitative “small,” “medium,” and “large.” This is an operation fraught with many dangers: The definitions are arbitrary…’ (p. 12). Cohen urged researchers to establish magnitude of effects that were important within their respective fields. In neuropsychology, importance can be defined as clinical significance. The current findings suggest that the use of the descriptor ‘small’ effect size (associated with correlation coefficients between .10 and .30; Cohen, 1988) requires fine tuning with respect to clinical significance. The threshold for clinical significance, at least in terms of likelihood of misclassification, appears to fall somewhere within the range of what is considered a small to moderate association. Monte Carlo simulations would be an ideal method for empirically determining the threshold of single test associations with individual difference variables required for clinical significance when interpreting a battery of tests. Determination of clinically important differences using Monte Carlo simulations would also prove useful in other areas of neuropsychology, most notably when we hope for a lack of clinical significance, such as during tests for psychometric equivalence (see inferential confidence interval approach detailed by Ross, Furr, Carter, & Weinberg, 2006).

An additional contribution of this study is the exploration of the clinical significance of the association between medical conditions and cognition. Clearly, the cumulative data (the body of literature is immense and we focused primarily on recent findings) suggest medical conditions are an important source of variance in cognition (e.g. Aarts et al., 2011; Downer et al., 2016; Fabbri et al., 2016; Jutkowitz et al., 2016; Piotrowicz et al., 2016; Viscogliosi et al., 2016) and may be a critical risk factor for mild cognitive impairment (Vassilaki et al., 2015) and dementia (Kelaiditi et al., 2013; Vassilaki et al., 2015; Xu et al., 2015). These data, however, are largely epidemiological in nature. Large scale epidemiological studies highlight the limitations of methods for statistical significance (we reference a classic ‘The earth is round ($p < .05$’; Cohen, 1994): statistical significance for trivial effects will be demonstrated in large samples, but does not elucidate clinical significance. Small associations underlie robust findings in large studies (in smaller ones they are likely to be obscured by a higher probability of Type II error), but the current data suggest that small (statistically significant) findings may not always be clinically significant. In this study, the accumulation of multiple medical conditions was not associated with increased likelihood of misclassification, suggesting that clinically accounting for medical conditions during neuropsychological interpretation is not recommended.

These data are limited by the self-report nature of the medical conditions explored, including those medical conditions that were used as exclusionary factors, which could influence the status of this sample as neurologically healthy. No external criterion was used to determine whether this sample was indeed neurologically, and by extension, cognitively healthy, and we were limited by exclusion for self-reported conditions that could impact brain health. It remains possible that our sample, particularly the older adult group, contained persons who have undiagnosed neurodegenerative or other brain disease. In addition, our data are
limited by the nature of the cognitive tests used: this was a very brief neuropsychological battery that was administered by telephone. It is unknown how these results would have differed with a more comprehensive neuropsychological battery, particularly with a battery including measures such as learning over trials or measures of executive function that might be sensitive to age-associated diseases. Although the value in the current manuscript is its argument for a novel approach to the creation of normative data that is less dependent on the particulars of the current sample, replication within other batteries is needed.

In conclusion, this approach to an empirical basis for normative data creation focuses on implications for interpretation at the level of the neuropsychological battery rather than merely focusing on single tests. This is more in keeping with clinical practice in neuropsychology: interpretation of cognitive profiles occurs in the context of a battery. This focus on the battery is similar to the multivariate interpretive approach of Huizenga and colleagues (2016), who propose creating normative corrections using resampling methods. The approach supported in the current paper, however, is to first explore differential base rates of impaired scores between theoretical and actual base rates, and then to use this information to guide decisions to create demographic adjustments within normative data. Differential base rates empirically detail the potential clinical implications of failing to create an appropriate normative group, thus maximizing clinical relevance by avoiding corrections for variables, such as number of medical conditions, which have no clinical significance. After an appropriate normative group has been created, clinical interpretation using Crawford and colleagues’ (2007) theoretical base rates is suggested, particularly if a clinician were to choose only some rather than all of the measures detailed in actual base rates.

Acknowledgements

This research was made possible using the data/samples collected by the CLSA [Data Set Baseline Tracking version 3.0]. Funding for the CLSA is provided by the Government of Canada through the Canadian Institutes of Health Research (CIHR) [grant number LSA 94473] and the Canada Foundation for Innovation. The CLSA is led by Drs. Parminder Raina, Christina Wolfson and Susan Kirkland. Support was provided for the fifth author (LEG) by a Canadian Institutes of Health Research New Investigator award and the McLaughlin Foundation Professorship in Population and Public Health.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The preparation of this manuscript was supported by funding provided by the Alzheimer Society of Canada/Alzheimer Société de Canada and the Pacific Alzheimer Research Foundation [grant number 17–29].

References


Crawford, J. R., Garthwaite, P. H., & Gault, C. B. (2007). Estimating the percentage of the population with abnormally low scores (or abnormally large score differences) on standardized neuropsychological


accuracy to identify neurocognitive impairment with the Consortium to Establish a Registry for Alzheimer’s Disease—Neuropsychological Assessment Battery (CERAD-NAB). *European Archives of Psychiatry and Clinical Neuroscience*, 265, 407–417. doi:10.1007/s00406-014-0571-z


Teng, E. L. (1994). *The mental alternation test (MAT)*. Los Angeles: Department of Neurology, University of Southern California School of Medicine.


