Title: Should cognitive screening tests be corrected for age and education? Insights from a causal perspective

Authors: Marco Piccininni, Jessica L. Rohmann, Maximilian Wechsung, Giancarlo Logroscino, and Tobias Kurth

Correspondence Address: Correspondence to Marco Piccininni, Institute of Public Health, Charité – Universitätsmedizin Berlin, Chariteplatz 1, Berlin, Germany 10117 (e-mail: marco.piccininni@charite.de)

Affiliations: Institute of Public Health, Charité – Universitätsmedizin Berlin, Berlin, Germany (Marco Piccininni, Jessica L. Rohmann, and Tobias Kurth); Center for Stroke Research Berlin, Charité – Universitätsmedizin Berlin, Berlin, Germany (Marco Piccininni, Jessica L. Rohmann); Institute of Biometry and Clinical Epidemiology, Charité – Universitätsmedizin Berlin, Germany (Maximilian Wechsung); Department of Basic Medical Sciences, Neuroscience and Sense Organs, University of Bari Aldo Moro, Italy (Giancarlo Logroscino); and Center for Neurodegenerative Diseases and the Aging Brain, Department of Clinical Research in Neurology, University of Bari Aldo Moro, Pia Fondazione Cardinale G Panico, Tricase, Italy (Giancarlo Logroscino).

Funding: The authors received no specific funding for this work.

Data Availability Statement: All data used in the section “Case study using OASIS-1 data” were provided by OASIS: Cross-Sectional: Principal Investigators: D. Marcus, R. Buckner, J. Csernansky J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382. The data are open access and can be accessed via https://www.oasis-brains.org/. Documents needed to understand the data collection process are also accessible via the aforementioned link.

Thanks: Data were provided by OASIS: Cross-Sectional: Principal Investigators: D. Marcus, R. Buckner, J. Csernansky J. Morris; P50 AG05681, P01 AG03991, P01 AG026276, R01 AG021910, P20 MH071616, U24 RR021382. G.L. acknowledges funding from Regione Puglia (D.G.R. n. 2117–21.11.2018) and “Tecnomed Puglia per la Medicina di Precisione” (Tecnomed Puglia for Precision Medicine). We thank Simen Gaure for providing the wrapper for the hcubature function.

Conference presentation: Presented as a poster at the 2022 Society for Epidemiologic Research annual meeting, Chicago, Illinois, June 15, 2022

Preprint Information: N/A

© The Author(s) 2022. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
Abstract
Cognitive screening tests like the Mini-Mental State Examination are widely used in clinical routine to predict cognitive impairment. The raw test scores are often corrected for age and education, although documented poorer discrimination performance of corrected scores has challenged this practice. Nonetheless, test correction persists perhaps due to the seemingly counterintuitive nature of the underlying problem. We use a causal framework to inform the longstanding debate from a more intuitive angle. We illustrate and quantify the consequences of applying the age-education correction of cognitive tests on discrimination performance. In an effort to bridge theory and practical implementation, we computed differences in discrimination performance under plausible causal scenarios using OASIS-1 data. When age and education are causal risk factors for cognitive impairment and independently also affect the test score, we show that correcting test scores for age and education removes meaningful information, thereby diminishing discrimination performance.

1. Introduction
Cognitive screening tests are tools used to screen for dementia and cognitive impairment [1,2]. These tests are validated by assessing their ability to predict clinical diagnoses, such as dementia and mild cognitive impairment, characterized by a decline in cognitive performance [1,2,3,4,5,6].

An individual’s raw test score is the numerical result of the test, such as the number of correct answers or errors, time to complete the test, or performance rating [7,5]. On their own, these raw scores are generally considered to have no inherent meaning and only become interpretable when compared with pre-existing norms or standards from individuals with similar demographic characteristics (e.g., age, sex, ethnicity, and/or education level) [7,5]. This comparison with the corresponding demographic-specific norm is conducted by transforming the raw score into a corrected score. In psychology, this practice is referred to as standardization, adjustment or correction.
The correction of raw scores is often employed in cognitive screening [8,9,10,11,1], despite having repeatedly encountered sharp criticism pertaining to its poor discrimination performance. First theoretical concerns were raised together with initial empirical observations [12,13,14,15,16,17,18,19], and found additional support in simulation studies [13]. However, the aforementioned criticism did not noticeably shift the paradigm. One reason might be the lack of an explicit causal framework, which, in our opinion, is necessary for an intuitive, complete understanding of this problem. Therefore, we aim to take a causal approach to address this long-standing debate regarding prediction performance. We provide an explanation for the counterintuitive results in the literature surrounding the problem of the age-education correction in cognitive screening. Specifically, we detail the explicit relationship between the correction of these tests and discrimination performance operationalized as Area Under the ROC Curve (AUC).

2. Theoretical background and main hypothesis

1. What is test correction?

Test correction is a common practice that relies on normative data. For practical reasons, correction factors or norms are established using the raw scores of a normative sample, typically composed of so-called ‘healthy’ individuals who are not affected by diseases thought to lead to cognitive or behavioral impairment [5,10]. We consider a normative sample to refer to a group of individuals without the specific condition (i.e., impairment) of interest. Once normative data are available for a given test, each new individual raw score is compared to the distribution of scores of individuals with similar demographic characteristics in the normative sample [5,9]. To make this comparison simpler and more explicit, the raw score is converted into a corrected score. The Z score is the most commonly encountered example of correction in practice [5,20,21,11], likely due to its several convenient statistical properties and straightforward interpretation [5].

The Z score is the distance between an individual’s performance and the average performance of the group of individuals from the normative sample with similar demographics, measured in units of this group’s standard deviation [7,5,22]. The Z score correction relies on the assumption that the distribution of normative data within each demographic stratum is approximately normal [5,20]. Alternatively, model-based corrections can be used to accommodate sparse data but rely on stronger assumptions [5,20,21,11]. Regardless of the correction method, the average score of all non-impaired individuals in one demographic stratum will be the same as the average score of all non-impaired individuals in any other demographic stratum. The correction process transforms the raw test scores into new corrected values by ‘removing’ the components of the raw scores attributable to the demographic variables [22]. This interpretation requires the assumption that the effects of demographic variables are the same among impaired and non-impaired individuals.

Age and education are considered particularly important in cognitive assessment. Almost all tests for cognitive impairment are traditionally corrected for these two variables [22]. This implies that the components of each raw score attributable to age and education are ignored when distinguishing between impaired and non-impaired individuals. Only the comparison to the performance of the non-impaired individuals within the same demographic stratum is considered. This means that an older individual with lower educational attainment having a low raw score may still be classified as “non-impaired or “normal [17]. Such an individual will be classified as “impaired at a lower score than a younger person or one with a higher educational attainment [22].

The concept of age correction is thus intrinsically linked to the life-span development theoretical model [23], in which a certain amount of cognitive decline is considered normal and
the expected performance is a function of age [22]. This stands in contrast with the biological aging model [24], in which any loss of functionality, even if attributable to aging, should be detected and potentially addressed in treatment strategies [22].

2. Why can test correction be problematic?

The argument for accounting for demographic variables in cognitive screening tests is that they are known to impact test performance [5,11,25,1]. Therefore, removing the components of the raw scores explained by these variables will reduce the variance of the distribution of the scores [5,18]. This reduces the “noise and thereby makes the difference in test scores between impaired and non-impaired individuals more pronounced, resulting in increased discrimination performance. Therefore, the use of such correction is often underpinned by observed associations between demographic variables and raw test scores [16].

Despite this seemingly sound rationale, a growing body of evidence challenges the case for correction [13]. Already in 1996, the results from a large survey of older persons showed that the ability of a psychological screening battery to discriminate dementia cases was not improved by correcting for demographic characteristics compared to using the raw scores [15]. Well-known screening tests, such as the Mini-Mental State Examination [16] (MMSE) and the modified 3MS version [12], when corrected, showed equal or worse performance compared with raw scores in screening for dementia and cognitive impairment. Furthermore, the ability of psychological tests to predict future progression to dementia was reduced by correction in longitudinal studies [17,18]. Moreover, using age-corrected test scores is known to lower the sensitivity for detecting Mild Cognitive Impairment due to Alzheimer’s Disease, especially outside the original development datasets [22].

The reason why the age correction reduces the discrimination performance is not due to a flawed rationale behind the correction, but rather that it does not consider a crucial element of the whole picture. Indeed, age and education not only influence raw test scores, but they, themselves, are also causal risk factors for cognitive impairment and dementia. Therefore, our hypothesis is that using the correction, which removes the impact of these variables from the raw scores, likely also removes some valuable predictive information provided by these risk factors, thereby diminishing the test’s ability to discriminate between impaired and non-impaired individuals [12].

The general idea that test correction ignores the relationships between demographic variables and the probability of being cognitively impaired, and that the concept of using age-corrected scores to detect cognitive impairment can be misleading, were first proposed by Berkman [19] and Sliwinski et al. [14,22]. Thereafter, O’Connell and Tuokko illustrated how the correction differentially affects the overlap of the test score distribution of impaired and non-impaired groups if the demographic variables only inject noise or are risk factors [13].

This theory was later corroborated in simulations and using real data [13]. Overall, discrimination performance was found to be reduced when tests were corrected for demographic variables [13]. After applying the correction, a lowered sensitivity was observed, and as a trade-off, a slight increase in specificity [13].

3. Correction from a causal perspective

Despite this empirical evidence and theoretical criticism, the practice of applying the age-education correction in cognitive screening persists. We argue that this is due to a perceived paradox pertaining to the problem at hand. First, theoretically, indifferent situations, the application of the same correction may either be beneficial or detrimental in terms of discrimination. Second, practically, it seems unintuitive that using additional information (i.e., about age and education) to create the corrected score ultimately results in an inferior
discrimination performance. Historically, many counterintuitive situations perceived to be paradoxes only became readily understandable in a causal framework [26]. Although this age-education correction paradox arose in the domain of prediction, we argue it is no exception, as evidenced by the necessary use of causal terms (e.g., "impact", "effect", "attributable", "explained", and "risk factor") to describe the problem. Indeed, recent work has shown that causal knowledge and causal diagrams can be used to inform prediction modeling strategies in biomedical research; specifically, to inform predictor selection, assess transportability, and predict outcomes under possible interventions [27,28,29,30,31]. This work centers on the intersection of causal inference, focused on understanding determinants of and treatments for diseases, and prediction, focused on identifying clinical conditions to optimize care.

In the field of causal inference, Directed Acyclic Graphs (DAGs) are used to visually depict underlying causal structures and inform strategies to address biases [26,32,33]. In these causal diagrams, single nodes are used to represent variables, while arrows represent direct causal paths between two variables. DAGs are used to encode a priori assumptions behind the underlying data generation process, and through a series of graphical rules, provide qualitative information about conditional independencies between the variables. Several helpful introductions to DAGs have been published elsewhere [26,32,33,34].

1. Understanding the age-education test correction using DAGs

We can use a DAG to view the age-education correction problem through a causal lens. Specifically, we can draw a DAG describing the causal relationships between age (A), education level (E), cognitive impairment (D), and raw test score (X). Age and education are known risk factors for cognitive impairment; therefore, we can draw an arrow from age (A) to cognitive impairment (D) and another arrow from education (E) to cognitive impairment (D) (Figure 1). Demographic factors are also known to causally affect cognitive test performance (X) [5,11,1,25]; for example, age and education may affect comprehension, education may affect cognitive reserve, and age affects the occurrence of other pathologies that can influence test performance. Moreover, the cognitive test performance (X) is also impacted by cognitive impairment (D). These relationships are made explicit by the three arrows from A, E, and D into X (Figure 1).

Of course, this is a very parsimonious representation of reality, and the DAG could be expanded to include other causes of cognitive impairment and explicitly depict the measurement error in test performance; however, we believe this simple DAG is sufficient to correctly illustrate the problem at hand, which is also at the core of complex, real-world scenarios. Our DAG is implicitly supported by the intuition of Hessler et al. and Magni et al. (despite some inconsistencies in their reasoning) that age and education can be conceptualized as “confounding variables of the relationship between raw cognitive screening test score and disease status [17,25].

In our DAG (Figure 1), three direct arrows connect node D to the other nodes, A, E, and X. This indicates that age, education level, and raw test score each provide unique information about cognitive impairment that cannot be captured by the other two variables. In more technical terms, all three variables are part of X’s Markov Blanket [35,36,37] and are needed for an optimal prediction in terms of calibration [27,35,37]. Intuitively, it therefore seems wise to consider age, education level, and the raw test score jointly to predict cognitive impairment. However, since the focus of our work is to inform the ongoing debate on the use of the age-education correction in cognitive screening, we compared only two commonly encountered strategies to predict cognitive impairment: (1) using the raw score alone or (2) using the corrected score alone.

There are three distinct paths connecting the cognitive impairment (D) node to the raw test
score (X) node: D \rightarrow X, D \leftarrow A \rightarrow X, D \leftarrow E \rightarrow X. All of these paths are open; therefore, according to causal graph rules, information about cognitive impairment “flows” to the raw test score through all of them [26,33,34]. This means that the association between raw test score and cognitive impairment observed in the data, given correct DAG specification and some other common assumptions [26,33], is the joint result of these three paths. Of these three paths, only one (D \rightarrow X) is causal, while the paths passing through the nodes age or education are non-causal because they each include an arrow pointing from effect to cause.

The correction transforms the raw test scores by removing the components of the raw scores attributable to the demographic variables [22]. Indeed, we can conceptualize the correction (under some assumptions presented explicitly later) as an attempt to remove the information flowing into X from D via the two non-causal paths, specifically through the arrows A \rightarrow X and E \rightarrow X. If we assume the DAG in Figure 1 is correct, then deleting the information delivered by the arrows A \rightarrow X and E \rightarrow X will modify the association between X and D because these two arrows each constitute the final step in a path that delivers information about cognitive impairment to the raw test score. Therefore, the age-education correction implies a loss of useful information assuming this causal structure. In the next section, we show that given the signs of the causal effect sizes in play, this loss of information translates into a lower discrimination performance.

Ultimately, the consequence of this action depends on whether age and education are risk factors for cognitive impairment. If one alternatively assumes that age and education are not causes of cognitive impairment (see modified DAG in Figure 2), neither age nor education are on a path through which information about cognitive impairment flows to raw test score. In this scenario, the arrows A \rightarrow X and E \rightarrow X deliver no information about cognitive impairment (D) to the raw test score (exit(X)), and simply represent injected noise. In such a situation, the correction is actually beneficial, since it removes this noise.

2. The age-education correction in a fully specified causal model

Returning to the DAG depicted in Figure 1, each node represents a variable and the diagram represents an underlying structural model [26,33,34]. Each variable can be written as the output of an assignment function of the variable’s direct causes (parent nodes) as well as some unknown random noise variable [26,33,34]. The assignment formalizes the directionality of the corresponding causal relationships. Noise variables are not conventionally depicted in DAGs and are assumed to be independent of each other.

Suppose our causal model is fully specified and we know both the true assignment functions and the noise probability distributions. For individuals i = 1, ..., n, we model age and education as random variables A_i and E_i with distributions P_A and P_E, respectively. A Bernoulli variable D_i with success probability p_i denotes an individual’s cognitive status, with D_i = 1 denoting cognitive impairment. The success probability is assigned as a function of its parents’ nodes. We define

\[ p_i = \frac{\exp (\gamma_0 + \gamma_1 A_i + \gamma_2 E_i)}{1 + \exp (\gamma_0 + \gamma_1 A_i + \gamma_2 E_i)}. \]

We choose \( \gamma_1 \geq 0 \) and \( \gamma_2 \leq 0 \) to acknowledge that higher age is a known risk factor for cognitive impairment while higher education is known to protect against it.

For an individual’s raw test score, we assume a simple linear assignment function. Let \( \{\varepsilon_i; i = 1, \ldots, n\} \) be independent and normally distributed with mean zero and positive variance \( \sigma^2 \). Then

\[ X_i = \beta_0 + \beta_1 A_i + \beta_2 E_i + \beta_3 D_i + \varepsilon_i \]

where \( \beta_1, \beta_3 \leq 0 \) and \( \beta_2 \geq 0 \). As such, we assume the mean raw test score decreases with
age, increases with education, and decreases with cognitive impairment. These effects have been reported for several cognitive screening tests [13,11,25,1] such as the MMSE. Here, we assume no interaction term in the assignment function. These assumptions are of course idealizing, particularly the normal distribution assumption, common in many model-based methods for test correction [20,21,11], which allows us to conveniently ignore ceiling and floor effects.

The corrected score \((Z_i)\) of an individual with age \(a_i\) and education level \(e_i\) is obtained by subtracting the expected raw score of non-impaired individuals with the same age and education from the individual’s raw score, and then dividing by the corresponding conditional standard deviation,

\[
Z_i = \frac{X_i - (\beta_0 + \beta_1 a_i + \beta_2 e_i)}{\sigma}.
\]

To compare discrimination performance, we used the difference between the AUC of the raw score, \(\text{AUC}(X)\), and the AUC of the corrected score, \(\text{AUC}(Z)\). The choice of this metric is consistent with previously published simulation studies [13]. It is known that the AUC is the probability that a randomly drawn individual from the impaired group has a lower raw score compared with a randomly drawn individual from the non-impaired group. Therefore,

\[
\text{AUC}(X) = \Pr\{X_1 < X_0 | D_1 = i\}
= \Pr\{\beta_1 (A_1 - A_0) + \beta_2 (E_1 - E_0) + \beta_3 < e_0 - e_1 | D_1 = i\}
= \int \Pr[\beta_1 (a_1 - a_0) + \beta_2 (e_1 - e_0) + \beta_3 < e_0 - e_1 | E_1 = e_1, A_1 = a_i, D_1 = i] \nonumber
\]

where conditions of the form \(A_1 = a_i, D_1 = i\) are to be understood as \(A_1 = a_i, A_0 = a_0, D_1 = 1, D_0 = 0\). Since the noise variables are independent standard normal variables, their difference is also normal with mean zero and variance \(2\sigma^2\). Furthermore, they are independent of all the other variables. Hence, the probability inside the integral can be written as \(1 - \Phi_{0.2\sigma^2}(\beta_1 (a_1 - a_0) + \beta_2 (e_1 - e_0) + \beta_3)\). Following the same rationale, we see that:

\[
\text{AUC}(Z) = \Pr\{Z_1 < Z_0 | D_1 = i\}
= \Pr\{X_1 - (\beta_0 + \beta_3 A_1 + \beta_2 E_1) < X_0 - (\beta_0 + \beta_1 A_0 + \beta_2 E_0) | D_1 = i\}
= \Pr[\beta_3 < e_0 - e_1 | D_1 = i]
= 1 - \Phi_{0.2\sigma^2}(\beta_3).
\]

A detailed mathematical derivation of \(\text{AUC}(X)\) is provided in Web Appendix 1. These mathematical expressions corroborate the intuitive reasoning we outlined in Section 3.1. The discrimination performance of the corrected test score (\(\text{AUC}(Z)\)) depends only on the direct causal effect (\(\beta_3\)) of cognitive impairment on the raw test score. This is in line with our previous statement that the corrected score uses only information flowing through the direct arrow \(D \rightarrow X\) as the information about cognitive impairment provided by the paths containing age or education is removed.

By contrast, \(\text{AUC}(X)\) depends additionally on \(\gamma_1, \gamma_2\) and \(\beta_1, \beta_2\), i.e., on the strength of the conditional causal relationships between the demographic factors and cognitive impairment on one hand, and between the demographic factors and the raw test score on the other (see Web Appendix 1). Thus, the choice whether to apply the correction must be informed by knowledge about the underlying causal relationships. Indeed, the difference between the discrimination performance of the two approaches hinges on the magnitude of all these causal effects. Although the derived expressions are too complicated for the analytical computation of \(\text{AUC}(X) - \text{AUC}(Z)\), in Section 4, we present a case study to illustrate its dependence on the coefficients \(\gamma_1, \gamma_2, \beta_1, \beta_2\), and \(\beta_3\) using numerical analysis.
Informally, we can already anticipate the consequence of using the age-education correction on the AUC if age and education affect both cognitive impairment \((\gamma_1 > 0\) and \(\gamma_2 < 0\)) and test score \((\beta_1 < 0\) and \(\beta_2 > 0\)). We saw that \(AUC(Z)\) corresponds to the probability of drawing a value higher than \(\beta_3\) from a normal distribution with mean 0 and variance \(\sigma^2\), depicted as the black area in Figure 3. Similarly, \(AUC(X)\) corresponds to the probability of drawing a value higher than \(\beta_1 (A_1 - A_0) + \beta_2 (E_1 - E_0) + \beta_3\) from the same normal distribution. Since \(A_1\), \(A_0\), \(E_1\), and \(E_0\) are random variables, \(AUC(X)\) actually corresponds to a weighted average of the probabilities of this event for the different realizations of \(A_1\), \(A_0\), \(E_1\), and \(E_0\). For simplicity, however, let’s imagine that \(AUC(X)\) is the area under the normal distribution to the right of the “typical” realization \(\beta_1 (a_1 - a_0) + \beta_2 (e_1 - e_0) + \beta_3\) (grey and black areas in Figure 3). If age and education are risk factors for cognitive impairment, we generally expect that \(a_1 > a_0\) and \(e_1 < e_0\) (i.e., individuals with cognitive impairment have higher age and lower education compared to individuals without cognitive impairment). In this scenario, since \(\beta_1 < 0\) and \(\beta_2 > 0\), the AUC of the raw score will generally be higher than the one of the corrected score (Figure 3). Notably, the same reasoning also applies if the effects of age and education on the test score are different among individuals with and without impairment (see Web Appendix 2).

4. Case study using OASIS-1 data

To compare the discrimination performance of the raw versus the corrected score of a cognitive screening test, we computed \(AUC(X)\) and \(AUC(Z)\) under plausible parameter constellations using data from the Open Access Series of Imaging Studies (OASIS) [38]. The OASIS project makes neuroimaging datasets freely available to the scientific community. We used the data from OASIS-1, a cross-sectional collection of information about brain magnetic resonance imaging of 416 right-handed individuals aged 18 to 96 [39]. Young and middle-aged adults were recruited from the Washington University community, while older individuals were recruited from a longitudinal pool of the Washington University Alzheimer Disease Research Center [39]. We only considered the 198 individuals aged 60 or more who underwent a full clinical assessment. Full details are available elsewhere [39].

We extracted the values of age, education level, Clinical Dementia Rating (CDR) scale and MMSE raw test score for each individual. For the purposes of this example, we converted the categorical education level variable into number of years of education (less than high school graduate: 8 years; high school graduate: 12 years; some college: 14 years; college graduate: 16 years; beyond college: 18 years) [40].

We defined the outcome of interest (cognitive impairment) as having a CDR greater than zero (outcome prevalence: 50.50%). We considered this data suitable to generate parameters for our illustrative purposes because no cognitive test information was used to define cognitive impairment. Instead, it was exclusively based on clinical assessment using the CDR scale [39]. For simplicity, we assumed age to be normally distributed and the number of years of education to follow a \(\chi^2\)-distribution. We estimated the mean age (76.343) and standard deviation (8.087) and the mean number of years of education (13.848) from the OASIS-1 dataset. The linear and logistic regression parameter estimation yielded \(\beta_0 = 26.785, \beta_1 = -0.011\) (effect of age on test score), \(\beta_2 = 0.209\) (effect of education on test score), \(\beta_3 = -4.415\) (effect of cognitive impairment on test score), \(\sigma = 3.028\) and \(\gamma_0 = 0.707, \gamma_1 = 0.011\) (lnOR for conditional effect of age on cognitive impairment), \(\gamma_2 = -0.113\) (lnOR for conditional effect of education on cognitive impairment).

We then estimated \(AUC(X)\) for the MMSE raw test score using the R function \texttt{hcubature()} [41] to solve the quadruple integral (Web Appendix 3), additionally relying on a wrapper accommodating infinite boundaries [42]. \(AUC(Z)\) was computed using the R function
pnorm().
We computed the difference between the AUCs of the raw and corrected test scores for all possible combinations of the following parameter values, including the OASIS-1-derived estimates: \( \gamma_1 = (0,0.011,0.05,0.1) \); \( \gamma_2 = (-0.3,-0.113,0) \); \( a_1 = (-0.1,-0.05,-0.111,0) \); \( \beta_2 = (0,0.209,0.4) \); \( \beta_3 = (-7,-4.415) \). In all scenarios, we held the parameters \( \gamma_0 \) and \( \sigma \) constant at the values estimated from the OASIS-1 dataset. Given correct specifications of the DAG, equations, and noise functions, Web Figure 1 illustrates the difference in discrimination performance of these two approaches under different causal scenarios.

For the scenario in which the OASIS-1-derived parameters were used (black square in Web Figure 1), the AUC of the corrected score was 0.017 lower than the AUC of the raw score, indicating poorer discrimination performance.

If there is no causal effect of age and education on the raw test score (imagine the DAG in Figure 1 without \( A \rightarrow X \) and \( E \rightarrow X \), meaning \( \beta_1,\beta_2 = 0 \), it becomes irrelevant whether the correction is used, because there is no difference in the discrimination performance of the corrected and the raw scores (bottom row in Web Figure 1).

In each outlined quadrant, moving towards the top-right corner, the coloration becomes increasingly red. This means that, holding the other parameters constant, the correction becomes increasingly detrimental as \( \gamma_2 \) increases and \( \beta_1 \) decreases. This corresponds to an increase in the magnitude of the conditional causal effects of age on cognitive impairment (\( A \rightarrow D \)) and age on test performance (\( A \rightarrow X \)). Therefore, the information flowing from \( D \) to \( X \) through the path \( D \leftarrow A \rightarrow X \) also increases (Figure 1), and the correction therefore results in a greater information loss.

The same argument holds for the other non-causal path \( D \leftarrow E \rightarrow X \). Holding all other parameters constant, the correction results in increasingly poor discrimination performance as \( \gamma_2 \) decreases and \( \beta_2 \) increases. This is reflected in Web Figure 1: holding \( \beta_3 \) constant, moving towards the top-right corner, the coloration of the quadrant becomes more red. Under these circumstances, by preventing flow from \( E \) to \( X \), the correction results in a larger loss of information about cognitive impairment delivered via the path \( D \leftarrow E \rightarrow X \).

Conversely, the correction appears beneficial if \( \gamma_1,\gamma_2 = 0 \) and \( \beta_1,\beta_2 \neq 0 \). Indeed, if age and education do not causally affect cognitive impairment (Figure 2), they only inject noise into the raw test score. As always, the correction corresponds to removing information delivered by the arrows \( A \rightarrow X \) and \( E \rightarrow X \); but this time the information is just noise (\( E \) and \( A \) are not on an open path between \( D \) and \( X \)). Therefore, correction results in a reduction of this noise and a higher discrimination performance.

If only one of the demographic variables (i.e., age or education level) is a cause of cognitive impairment, applying the age-education correction may result in either a lower or higher discrimination performance compared to the raw score. This ultimately depends on the trade-off between the amounts of useful information lost and noise removed by correction.

Let us now focus on the relationship between cognitive impairment and raw test score (\( D \rightarrow X \) in Figure 1). Holding all other parameters constant, as the causal effect of cognitive impairment on test score becomes larger in magnitude (lower \( \beta_3 \)), more information flows through the path \( D \rightarrow X \) compared to the other two paths into \( X \). In this case, the decision whether or not to correct becomes less relevant because any potential harms or benefits of the correction are attenuated.

In using OASIS-1 data to estimate some plausible parameters in this illustrative example, we acknowledge that some distributional and functional assumptions are unrealistic (e.g., linearity, no interaction terms, homoscedasticity, normality of errors, and absence of ceiling and floor effects in the test score). Moreover, in the OASIS-1 dataset, \( \beta_1 \) and \( \gamma_1 \) were very small. We emphasize that the use of these parameter estimates was only a starting point and has
allowed us to readily show the impact of correction on discrimination performance under different causal scenarios, bridging the theoretical causal model and a practical implementation. We hope this simplified, didactic exercise serves as a basis for further empirical confirmation.

5. Conclusion

Age-education correction of neuropsychological and cognitive screening test scores is a common practice in the context of neurodegenerative diseases. Corrected scores are anchored in numerous definitions of Mild Cognitive Impairment held in high esteem[43]. Both the original Mild Cognitive Impairment criteria proposed by Petersen [44,45] and the more recent DSM-5 definition of Mild Neurocognitive Disorder [46] rely on the comparison of test scores with appropriate norms (defined by age alone or age and education) to make a diagnosis. Though initial criticisms of test correction have been described in the literature, the problem of poorer discrimination performance has been largely ignored by the clinical and research communities. One explanation may be that in the absence of an explicit causal framework, the problem is far from intuitive.

We approached the problem from a causal perspective, using DAGs to depict underlying data generation processes and graph rules to intuitively show the consequences of correcting cognitive screening tests for age and education. Assuming a fully-specified causal model, we compared the discrimination performance of the raw and corrected test scores in terms of AUC for a plausible range of causal scenarios.

If age and education are causal risk factors for cognitive impairment, then the raw score contains information about cognitive impairment also delivered by age and education via two non-causal paths. As we showed, applying the age-education correction in this causal scenario removes useful information and results in worse discrimination performance compared with using the raw test score given the signs of the causal effect sizes in play.

Since age and education are well-known causal risk factors for cognitive impairment, we argue that the age-education correction is generally hard to justify in real-world scenarios. Thus, we believe correction should not be performed without careful consideration of the underlying causal processes, especially considering the importance of this prediction task in clinical practice. While our focus here on cognitive screening tests, our results are more generally applicable to all situations in which the same underlying causal structure is realistic. We believe our conclusion about discrimination performance holds for any single neuropsychological test used to predict cognitive impairment, given the conceptual similarity [6], although it is unlikely that a single neuropsychological test is used to predict the cognitive status in practice. Furthermore, our work provides insights on how to better predict cognitive impairment when trying to provide clinical guidance, and our results do not apply in the setting of causal research on the determinants of cognitive impairment.

In this work, we focused on discrimination performance and the question of whether the age-education correction should be applied. However, potential implications of this correction for other key prediction metrics (i.e., calibration), predictor selection strategies, and transportability should also be explored. Another important aspect to consider is societal impact. Using neuropsychological test scores without age-education correction results in a higher likelihood of classifying individuals who are older and less educated as cognitively impaired. This may exacerbate or attenuate social inequalities depending on the specific societal context, including the vulnerability of these sociodemographic groups, societal stigma, access to care, and access to financial support. For instance, it has been shown in the context of racial discrimination in workers’ compensation how demographic correction can either amplify or diminish inequalities depending on broader societal conditions [47]. Furthermore, the statistical tool of correction is directly linked to the technique proposed to mitigate
algorithmic bias and achieve so-called counterfactual fairness [48].

We believe our work provides important evidence on how causal knowledge expressed as a parsimonious representation of the underlying data generation process is useful to address relevant prediction tasks regularly encountered in clinical practice.

**Code availability**

The Web Appendix 3 contains the R code for (1) estimating the OASIS-1 derived parameters, (2) building the function to approximate the AUC for raw and corrected test scores under all described scenarios, and (3) approximating the AUC for raw and corrected test scores in the scenario defined by the combination of parameters estimated from OASIS-1 data.

**Author contributions**

M.P. conceptualized the study. Based on M.P.’s idea and inputs, M.W. developed the mathematical proof. M.P. and M.W. discussed the theoretical implications of the proof. M.P. wrote the R code and implemented the numerical analysis. M.P. and J.L.R. created and revised the figures. M.P., J.L.R, and T.K. interpreted and discussed the results. J.L.R. and M.P. wrote the first draft of the manuscript. G.L. provided important contextualization in the field of dementia research and clinical neurology. T.K. supervised the work. All authors provided critical comments and approved the final version of the manuscript.

**References**


[34] Pearl J. Causality: Models, Reasoning and Inference. 2nd ed. Cambridge, UK:
Cambridge University Press; 2009.


Figures

Figure 1. Directed Acyclic Graph depicting assumed causal relationships between age ($A$), education level ($E$), cognitive impairment ($D$), and raw test score ($X$). Figure created using the R package ggdag [49].

Figure 2. Directed Acyclic Graph depicting assumed causal relationships between age ($A$), education level ($E$), cognitive impairment ($D$), and raw test score ($X$) in the scenario in which age and education are assumed not to be risk factors for cognitive impairment. Figure created using the R package ggdag [49].

Figure 3. Plot representing the generic density function $\phi(k)$ of a normal distribution centered at zero. The area below the curve provides a geometrical intuition of the difference between the $\text{AUC}(Z)$ (black area) and $\text{AUC}(X)$ (grey and black areas).