Use and misuse of the calibration slope

Richard Stevens[1], Katrina Poppe[2]

[1]Nuffield Dept Primary Care Health Sciences, University of Oxford, Woodstock Road, Oxford, United Kingdom
[2]Faculty of Medical and Health Sciences, University of Auckland, New Zealand

Background:
The slope of a calibration plot is often referred to as “calibration slope”. Methodology texts emphasize the slope should not be used in isolation but accompanied by other metrics and graphs: poor calibration, by any definition, can occur even when the slope is perfect (equals 1).

Methods and Results:
We review recent usage of the calibration slope. In 33 validation papers (24 external) published 2017-2018, 25 papers identified the slope with calibration, 1 identified calibration slope with discrimination and 7 used the term calibration slope without explicitly interpreting it. In 17 papers (52%) the slope was used as sole measure of calibration. We are currently reviewing papers from 2019 and 2020.

Discussion:
The paper often cited as the origin of the “calibration slope” did not use the term calibration, but “spread”. More recently “spread” has been identified in some papers as an aspect of calibration and in others as an aspect of discrimination, sometimes by the same authors. We resolve this apparent paradox by proposing that calibration and discrimination are not a dichotomy. If we equate the A (calibration-in-the-large), B (calibration slope) and C (discrimination) of Steyerberg and Vergouwe’s ABCD[1] with bias, spread and ordering, then we can see that good calibration-in-the-large equates to low bias; calibration as often defined equates to low bias and adequate spread; good discrimination requires correct spread and correct ordering; and moderate to strong calibration, as defined by Van Calster[2], requires low bias, adequate spread and correct ordering.

Conclusion:
Authors, reviewers and editors have a duty to discourage the perception that calibration is a unidimensional construct quantifiable by a single statistic, the slope.

Keywords
Validaiton; calibration; discrimination; spread; slope.

References

[1] Steyerberg & Vergouwe, European Heart Journal (2014) 35, 1925–1931
[2] Van Calster et al. Journal of Clinical Epidemiology 74 (2016) 167e176