LINK TO ORIGINAL ARTICLE
LINK TO INITIAL CORRESPONDENCE

REPLY

# The STHLM3 prostate cancer diagnostic study: calibration, clarification, and comments

*Sigrid V. Carlsson and Michael W. Kattan*

We would like to thank Martin Eklund, Henrik Grönberg, and Tobias Nordström for their correspondence on our News & Views article (Personalized risk — stratified screening or abandoning it altogether? *Nat. Rev. Clin. Oncol.* **13**, 140–142 (2016))[1]. We are pleased to see the detailed response from these Stockholm 3 (STHLM3)-trial investigators (The STHLM3 prostate cancer diagnostic study: calibration, clarification, and comments. *Nat. Rev. Clin. Oncol.* http://dx.doi.org/10.1038/nrclinonc.2016.80 (2016))[2], which adds clarification on several aspects of their study.

Assessing discrimination of a prediction model is essential; however, we would continue to argue that calibration is even more important than discrimination when assessing the overall performance of a predictive model. As an example, a well-calibrated model for predicting high-grade prostate cancer would tell us that for every 100 patients with an estimated risk of 11%, close to 11 men would actually have high-grade disease (as predicted). Eklund and colleagues[2] claim that "a poorly calibrated model with high discriminatory power is highly useful". This is false: imagine if we took the model and divided all risks predicted by 100, a man with an 11% risk would instead be told his risk was 0.11%. In this case, the model would be poorly calibrated, but the discrimination (area under the curve) would remain unchanged[3]. Importantly, if the decision to biopsy is made at a cutoff of 10%, very different decisions and downstream consequences would result from the use of a model that assigns the man a predicted risk of only 0.11%, rather than 11%. Thus, knowing how close the risk assigned using a prediction model is to a man's true risk (calibration) is more important for the individual man than knowing whether the model distinguishes between men with and those without high-grade disease (discrimination)[3], particularly when the predicted risk is subsequently used for guiding the decision to biopsy, which is contingent upon an accurate prediction of risk. For this reason, we are delighted that the STHLM3 investigators' response to our article included a calibration plot for their predictive model, which shows good calibration — particularly in the risk-range that would form the basis for clinical decision-making.

This consideration brings us to the point that if a predictive model is to be used to inform clinical decisions, the clinical utility of the model needs to be evaluated, in addition to its discrimination and calibration[4]. We therefore disagree with Eklund and colleagues[2] that "demonstrating statistical significance as an independent predictor in a multivariable analysis is sufficient evidence of the value of a biomarker". In the reference cited in support of this statement, Pepe *et al.*[5] indeed also emphasize that "estimation of the increment in prediction performance is more important than testing the null hypothesis of no improvement." Statistical significance does not always imply clinical significance or clinical utility. Because, for instance, the genetic score was added early in the stepwise regression, before the plasma-protein biomarkers, its added clinical value over the plasma-protein markers remains unclear — that is, whether the exclusion of the genetic score would materially affect the number of high-grade cancers identified or biopsies avoided.

We are pleased to learn from Eklund *et al.*[2] that more data is forthcoming on the value of the STHLM3 test as a reflex test. We are hopeful that these thoughts are taken into consideration as the investigators publish in more detail their findings regarding the STHLM3 model.

*Sigrid V. Carlsson is at the Department of Surgery and the Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, 485 Lexington Avenue, New York, New York 10017, USA; and at the Department of Urology, Sahlgrenska Academy at Gothenburg University, Bruna Stråket 11B, 41345 Gothenburg, Sweden.*
*carlssos@mskcc.org*

*Michael W. Kattan is at the Department of Quantitative Health Sciences, Cleveland Clinic Foundation, 9500 Euclid Avenue, Cleveland, Ohio 44195, USA.*
*kattanm@ccf.org*

doi:10.1036/nrclinonc.2016.81
Published online 10 May 2016

1. Carlsson, S. V. & Kattan, M. W. Personalized risk — stratified screening or abandoning it altogether? *Nat. Rev. Clin. Oncol.* **13**, 140–142 (2016).
2. Eklund, M., Grönberg, H. & Nordström, T. The STHLM3 prostate cancer diagnostic study: calibration, clarification, and comments. *Nat. Rev. Clin. Oncol.* http://dx.http://dx.doi.org/10.1038.org/nrclinonc.2016.80 (2016).
3. Carlsson, S., Assel, M. & Vickers, A. Letter to the editor concerning 'Do prostate cancer risk models improve the predictive accuracy of PSA screening? A meta-analysis'. *Ann. Oncol.* **26**, 1031 (2015).
4. Steyerberg, E. W. *et al.* Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* **21**, 128–138 (2010).
5. Pepe, M. S., Kerr, K. F., Longton, G. & Wang, Z. Testing for improvement in prediction model performance. *Stat. Med.* **32**, 1467–1482 (2013).

**Acknowledgements**
The work of S.V.C. is supported by a postdoctoral research grant from AFA Insurance and, in part, by a Cancer Center Support Grant from the National Cancer Institute made to Memorial Sloan Kettering Cancer Center (P30 CA008748).

**Competing interests statement**
The authors declare no competing interests.