LINK TO ORIGINAL ARTICLE
LINK TO AUTHOR'S REPLY

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

Martin Eklund, Henrik Grönberg and Tobias Nordström

In response to the News & Views article by Carlsson and Kattan (Personalized risk — stratified screening or abandoning it altogether? *Nat. Rev. Clin. Oncol.* **13**, 140–142 (2016))[1], we would like to thank the authors for their acknowledgement of, and positive remarks on, the Stockholm 3 (STHLM3) study, in which we were involved[2]. We agree with their view that a blanket rejection of prostate-specific antigen (PSA)-based screening for prostate cancer is ill-advised and would lead to reduced opportunities to prevent death from prostate cancer. Neither do we believe such a rejection to be practically feasible. Clearly, the way forward is to improve our approach to prostate-cancer screening to permit early and accurate diagnosis of disease in men who need treatment, and to avoid overdiagnosis and unnecessary biopsies in those who do not. In light of this fundamentally important aim, we would like to add clarification on a few points raised by Carlsson and Kattan regarding the STHLM3 study.

First, Carlsson and Kattan[1] questioned the applicability of the STHLM3 model in the clinical setting, in which men with elevated serum PSA levels are subject to additional workup before deciding on whether to perform a biopsy. The aim of STHLM3 was to develop a tool to improve high-volume screening in the primary-care setting, building on the findings of the European Randomised Study of Screening for Prostate Cancer (ERSPC)[3]. Thus, the rational decision was to use PSA ≥3 ng/ml as a comparator in the STHLM3 study, in order to infer the same mortality reduction as that observed using this cutoff in the ERSPC. Nevertheless, because further workup in men with elevated levels of PSA is currently common practice, the authors' remark deserves attention, and we will address this issue in a forthcoming publication, in which we compare results from using the STHLM3 model for biopsy recommendations to current clinical practice in Stockholm, Sweden.

Second, we disagree with Carlsson and Kattan[1] regarding the failure of the STHLM3 investigators to address whether the genetic score — based on 232 single-nucleotide polymorphisms associated with prostate cancer — included in the STHLM3 model adds predictive value. Pepe *et al.*[4] have reported that demonstrating statistical significance as an independent predictor in a multivariable analysis is sufficient evidence of the value of a biomarker; such evidence is provided for the genetic score in Table 2 of the STHLM3 study publication by Grönberg *et al.*[2] Testing additionally for an improvement in the area under the curve (AUC) would be redundant and, therefore, unnecessary[4].

Third, Carlsson and Kattan[1] noted that calibration of the STHLM3 model was not reported by Grönberg *et al.*[2] We argue that 'discrimination' (that is, the ability to discriminate between cases and controls) is the most important property of a classification model: a poorly calibrated model with high discriminatory power is highly useful, whereas a well-calibrated model with poor discriminative performance is of limited value. Moreover, poor calibration can always be fixed, provided enough data are available[5]. Having said that, we agree that a well-calibrated predictive model is desirable; FIG. 1 shows the excellent calibration of the STHLM3 model.

Fourth, Carlsson and Kattan[1] point out correctly that the disease prevalence in the overall STHLM3-study population remains unknown, as biopsies were not performed in all participating men. For ethical and practical reasons, performing biopsies in men with low PSA levels was not deemed appropriate, a feature the STHLM3 study shares with virtually all other prostate cancer diagnostic studies. For example, the Prostate Health Index (PHI) and the 4KScore have been validated as reflexive tests in cohorts of men with increased PSA levels (usually defined as a serum PSA concentrations above 2–4 ng/ml)[6–9], making it difficult to infer that reductions in prostate-cancer mortality observed with these tests are equivalent to those associated with PSA screening using 3 ng/ml as a cutoff for biopsy. STHLM3 is, to our knowledge, the only prospective prostate cancer diagnostic study that demonstrates prevented biopsies and decreased overdiagnosis, without decreasing the detection of high-grade tumours.

Finally, we agree with Carlsson and Kattan's[1] view that informing doctors and patients about the individual probability of high-risk prostate cancer on a continuous scale, rather than according to risk group, could be relevant for clinical decision-making. In the ongoing clinical implementation of the STHLM3 model, the individual's risk of having a prostate cancer with a Gleason score ≥7 is reported to the doctor who ordered the test. Many patients (and, indeed, doctors) find it difficult, however, to conceptualize the risks and prefer a clearly stated recommendation on the appropriate course of action.

We hope that these clarifications address the questions posed by Carlsson and Kattan[1] on the performance characteristics of the STHLM3 model.

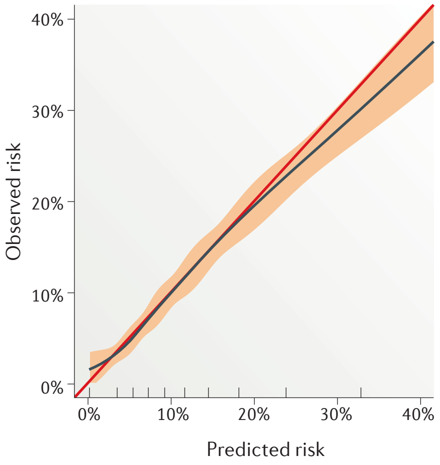

Figure 1 | **Calibration plot of the STHLM3 model for predicting high-risk prostate cancer.** The graph shows the calibration of the model — that is, the agreement between the predicted and observed risk of high-risk prostate cancer (Gleason score ≥7) — based on the results from the 5,344 biopsies performed in the STHLM3 validation cohort. The red line indicates perfect correspondence between predicted and observed risk (perfect calibration) and the black line shows the calibration of the STHLM3 model. The orange shaded area indicates the 95% confidence interval, and the tick lines above the x-axis shows deciles of the risk distribution, each representing one tenth of the population. The graph was produced using the R language and the gbm package[10,11].

Martin Eklund, Henrik Grönberg, and Tobias Nordström are at the Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Nobels väg 12, Stockholm 171 77, Sweden.

Correspondence to M.E.
martin.eklund@ki.se

doi:10.1038/nrclinonc.2016.80
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. Grönberg, H. *et al.* Prostate cancer screening in men aged 50–69 years (STHLM3): a prospective population-based diagnostic study. *Lancet Oncol.* **16**, 1667–1676 (2015).
3. Schröder, F. H. *et al.* Screening and prostate cancer mortality: results of the European Randomised Study of Screening for Prostate Cancer (ERSPC) at 13 years of follow-up. *Lancet* **384**, 2027–2035 (2014).
4. Pepe, M. S., Kerr, K. F., Longton, G. & Wang, Z. Testing for improvement in prediction model performance. *Stat. Med.* **32**, 1467–1482 (2013).
5. Vovk, V., Gammerman, A. & Shafer, G. *Algorithmic learning in a random world.* (Springer, 2005).
6. Punnen, S., Pavan, N. & Parekh, D. J. Finding the wolf in sheep's clothing: the 4Kscore is a novel blood test that can accurately identify the risk of aggressive prostate cancer. *Rev. Urol.* **17**, 3–13 (2015).
7. Bryant, R. J. *et al.* Predicting high-grade cancer at ten-core prostate biopsy using four kallikrein markers measured in blood in the ProtecT study. *J. Natl Cancer Inst.* **107**, djv095 (2015).
8. Loeb, S. The Prostate Health Index selectively identifies clinically significant prostate cancer. *J. Urol.* **193**, 1163–1169 (2015).
9. Catalona, W. J. A multicenter study of [−2]pro-prostate specific antigen combined with prostate specific antigen and free prostate specific antigen for prostate cancer detection in the 2.0 to 10.0 ng/ml prostate specific antigen range. *J. Urol.* **185**, 1650–1655 (2011).
10. R Foundation for Statistical Computing. R: A Language and Environment for Statistical Computing. *R Project* https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf (2015).
11. Ridgeway, G. *et al.* gbm: Generalized Boosted Regression Models. R package version 2.1.1. *R Project* https://cran.r-project.org/web/packages/gbm/index.html (2015).

Acknowledgements

The STHLM3 study was supported by grants from the Strategic Research Programme on Cancer (StratCan) of the Karolinska Institutet; the Linnaeus Centre for Breast and Prostate Cancer (CRisP; grant 108 70867901) of the Karolinska Institutet; the Swedish Research Council (grants K2010–70X-20430–04–109 3 and 2015–03292); the Swedish Cancer Society (grants 11–0287 and 15–649); the Foundation in Memory of Johanna Hagstrand and Sigfrid Linnēr (Stiftelsen Johanna Hagstrand och Sigfrid Linners Minne); and the Swedish Research Council for Health, Working Life and Welfare (Forskningsrådet för hälsa, arbetsliv och välfärd (FORTE); grants 2012–0073 and 2015–00184). The funding sources had no role in the study design; collection, analysis, or interpretation of data; writing of the report; or the decision to submit the report for publication. The researchers were all independent from the funding source. Karolinska Institutet collaborates with Thermo Fisher Scientific in developing the technology for STHLM3.

Competing interests statement

H.G. owns stock in Procant and, via this company, is co-owner of five prostate cancer diagnostic-related patents pending, has patent applications licensed to Thermo Fisher Scientific, and might receive royalties from sales related to these patents. M.E. is named as a co-inventor on four of these five patent applications. M.E. also owns stock in Genetta Soft and EKED Consulting. Karolinska Institutet collaborates with Thermo Fisher Scientific in developing the technology for STHLM3.