Journal of
Clinical Medicine

MDPI

Article

Inter-Rater Agreement in Assessing Risk of Bias in Melanoma Prediction Studies Using the Prediction Model Risk of Bias Assessment Tool (PROBAST): Results from a Controlled Experiment on the Effect of Specific Rater Training

Isabelle Kaiser [1,*], Annette B. Pfahlberg [1], Sonja Mathes [2], Wolfgang Uter [1], Katharina Diehl [1], Theresa Steeb [3], Markus V. Heppt [3,4] and Olaf Gefeller [1]

[1] Department of Medical Informatics, Biometry and Epidemiology, Friedrich Alexander University of Erlangen-Nuremberg, 91054 Erlangen, Germany
[2] Department of Dermatology and Allergy Biederstein, Faculty of Medicine, Technical University of Munich, 80802 Munich, Germany
[3] Department of Dermatology, University Hospital Erlangen, 91054 Erlangen, Germany
[4] Comprehensive Cancer Center Erlangen-European Metropolitan Area of Nuremberg (CCC ER-EMN), 91054 Erlangen, Germany
[*] Correspondence: isabelle.kaiser@fau.de

Abstract: Assessing the risk of bias (ROB) of studies is an important part of the conduct of systematic reviews and meta-analyses in clinical medicine. Among the many existing ROB tools, the Prediction Model Risk of Bias Assessment Tool (PROBAST) is a rather new instrument specifically designed to assess the ROB of prediction studies. In our study we analyzed the inter-rater reliability (IRR) of PROBAST and the effect of specialized training on the IRR. Six raters independently assessed the risk of bias (ROB) of all melanoma risk prediction studies published until 2021 (n = 42) using the PROBAST instrument. The raters evaluated the ROB of the first 20 studies without any guidance other than the published PROBAST literature. The remaining 22 studies were assessed after receiving customized training and guidance. Gwet's AC_1 was used as the primary measure to quantify the pairwise and multi-rater IRR. Depending on the PROBAST domain, results before training showed a slight to moderate IRR (multi-rater AC_1 ranging from 0.071 to 0.535). After training, the multi-rater AC_1 ranged from 0.294 to 0.780 with a significant improvement for the overall ROB rating and two of the four domains. The largest net gain was achieved in the overall ROB rating (difference in multi-rater AC_1: 0.405, 95%-CI 0.149–0.630). In conclusion, without targeted guidance, the IRR of PROBAST is low, questioning its use as an appropriate ROB instrument for prediction studies. Intensive training and guidance manuals with context-specific decision rules are needed to correctly apply and interpret the PROBAST instrument and to ensure consistency of ROB ratings.

Keywords: inter-rater agreement; inter-rater reliability; melanoma; risk of bias; prediction; PROBAST

check for updates

Citation: Kaiser, I.; Pfahlberg, A.B.; Mathes, S.; Uter, W.; Diehl, K.; Steeb, T.; Heppt, M.V.; Gefeller, O. Inter-Rater Agreement in Assessing Risk of Bias in Melanoma Prediction Studies Using the Prediction Model Risk of Bias Assessment Tool (PROBAST): Results from a Controlled Experiment on the Effect of Specific Rater Training. *J. Clin. Med.* **2023**, *12*, 1976. https://doi.org/10.3390/jcm12051976

Academic Editor: Anne Wierinckx

Received: 20 January 2023
Revised: 27 February 2023
Accepted: 28 February 2023
Published: 2 March 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Clinical and epidemiological studies devoted to evaluating prognostic and/or risk factors of a specific disease are prone to many forms of bias [1]. Bias is defined as the presence of systematic error in a study that leads to flawed results and thus impairs the validity of study findings [2,3]. To be able to properly interpret study results and to avoid under- or over-estimation of the parameter of interest, it is essential to assess the risk of bias (ROB) of studies [4]. Especially for the appropriate conduct of systematic reviews and meta-analyses, which have become increasingly important in clinical medicine over the last two decades [5], the assessment of the methodological quality of included studies has become a key element and is part of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guideline [6]. The need for methodological quality assessments has contributed to the development of a large number of ROB instruments over the last two decades [7,8]. Most of the instruments were developed for specific study designs, such as the revised Cochrane Risk-of-Bias tool (ROB2) for randomized controlled trials [9] or ROBIS (Risk Of Bias In Systematic review) for systematic reviews [10]. Another tool specifically designed for prediction studies and published in 2019 is PROBAST (Prediction Model Risk of Bias Assessment Tool) [11,12]. The development of PROBAST was based on a consensus process consisting of a Delphi procedure involving a panel of 38 experts and a refinement through piloting. The final instrument has a domain-based structure and provides criteria for the evaluation of the methodological quality of studies developing, validating, or updating prediction models [11]. The authors of PROBAST defined bias in the context of predictive studies as "shortcomings in study design, conduct, or analysis [that] lead to systematically distorted estimates of model predictive performance" [2]. Although the tool was published only a few years ago, it has already been used extensively [13]. This demonstrates that PROBAST fills an important gap in the repertoire of ROB tools for predictive studies.

Assessing ROB improves transparency about the methodological quality of studies. However, this is only possible if the ROB instruments themselves are valid and reliable. While validity addresses the extent to which the observed results represent the truth, reliability relates to the extent to which results can be reproduced. Low validity and poor reliability of ROB assessment tools, by impairing the quality of systematic evidence synthesis, may ultimately have an impact on decision-making and quality of patient care [14]. One element of reliability is the inter-rater agreement, which refers to the reproducibility or consistency of decisions between two or more raters [4]. ROB instruments often depend on the experience and personal judgment of raters, which can lead to different ROB ratings when multiple raters assess the same study. Thus, to assess and improve consistency in the application of ROB assessment tools, it is necessary to explore the inter-rater reliability (IRR) of ROB instruments. Up to now, only a few ROB tools have undergone extensive IRR or validity testing by independent groups [15–19]. Overall, these studies revealed deficits in the reliability of the tools examined [15]. There is, however, some evidence that intensive, standardized training for raters may significantly improve the reliability of ROB assessments [14,20].

To the best of our knowledge, hitherto no studies examining the effect of specialized training on the reliability of the PROBAST instrument exist. Therefore, our objectives were (i) to investigate the IRR of this instrument and (ii) to explore the effect of intensive rater training and targeted outcome-specific guidance manuals on the IRR in a representative manner, using melanoma prediction studies as an example.

2. Materials and Methods

2.1. Study Selection

We included 42 studies reporting development and validation of models predicting the individual risk of melanoma occurrence. The set of studies to be assessed was based on a recent systematic review of melanoma prediction modeling published in 2020 [21] and a literature update performed in August 2021. The update included the forward snowballing technique, which was applied on [21] and two other previously published systematic reviews on the same topic [22,23], and an electronic literature search in PubMed using the same search string as in [21]. Details on the study selection and eligibility criteria were published previously in a report describing the ROB of melanoma prediction studies based on the consensus ROB rating of the rater team [24].

2.2. ROB Assessment Using PROBAST

Six raters (I.K., S.M., K.D., T.S., M.V.H., O.G.) assessed independently the ROB of each study using the PROBAST instrument provided on the website [12]. The rater panel was multidisciplinary and consisted of epidemiologists (I.K., O.G.), clinical dermatologists (S.M., M.V.H.), and public health experts (K.D., T.S.) at different levels of professional experience with systematic reviews and ROB assessments. Two raters had no previous experience in this area. Although some of the raters had already performed ROB assessments, none had used the PROBAST instrument before.

PROBAST is structured into four domains: (1) The domain "participants" covers possible sources of bias related to the data sources and the participant selection; (2) the domain "predictors" contains bias through selection and assessment of predictors; (3) the domain "outcome" focuses on possible bias because of definition or determination of the outcome; and (4) the domain "analysis" covers bias linked to estimated predictive performance induced by inappropriate analysis methods or omission of statistical considerations. Each domain was rated individually as either low, high, or unclear. The raters were assisted in judging the ROB for each domain by a total of twenty signaling questions that were answered as yes, probably yes, no, probably no, or no information. Based on the ratings in the four domains, an overall ROB was assigned to each study. According to [11], the overall ROB was obtained by taking the lowest rating of any domain-specific ROB. Thus, a study only received a low overall ROB if all four domains were judged as low.

2.3. Rating Process and Training

A timeline of the study is given in Figure 1. Prior to the rating process, an initial meeting was held to discuss the objective and implementation of the ROB assessment. During this meeting, the published PROBAST literature, namely the original PROBAST publication [11] and the explanation and elaboration document [2], was provided to the raters. Thereafter, a random selection of 20 studies was assessed by the raters without any further guidance. After the completion of this part of the rating, two moderated training sessions followed where each PROBAST item was reviewed and its meaning discussed in the group to ensure that all raters interpreted the items in the same way. In addition, disagreements between the raters regarding the ROB ratings of the first twenty studies were discussed in two meetings lasting four hours each to reach consensus decisions. In three cases of sustained disagreement, two independent referees (A.B.P. and W.U.) made the final decisions. A customized guidance manual [24] was developed based on the consensus decisions. It contained decision rules to guide raters in making adjudications for each domain of the PROBAST instrument when specifically applied to melanoma prediction studies, establishing a common standard for the rating process. Afterwards, the ROB of the remaining 22 studies was assessed based on that guidance. Again, six consensus meetings of 1.5 h were held to resolve disagreements regarding the ROB ratings.

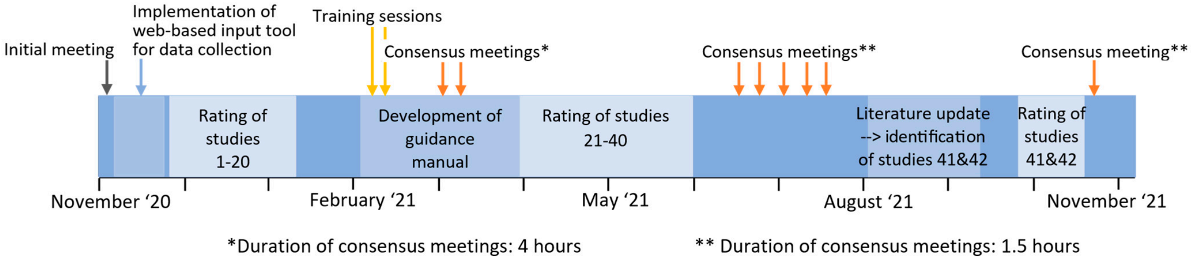

Initial meeting
Implementation of web-based input tool for data collection
Training sessions
Consensus meetings*
Consensus meetings**
Consensus meeting**
Rating of studies 1-20
Development of guidance manual
Rating of studies 21-40
Literature update --> identification of studies 41&42
Rating of studies 41&42
November '20
February '21
May '21
August '21
November '21
*Duration of consensus meetings: 4 hours
** Duration of consensus meetings: 1.5 hours

Figure 1. Timeline of the PROBAST assessment study from the initial meeting until the final consensus meeting showing all steps of the study.

2.4. Statistical Analysis

We determined the IRR before and after training for the domain-specific and overall ROB ratings. We calculated the pairwise agreement and the agreement at the multi-rater level, respectively. Given that six raters participated in the study, there were fifteen possible pairs of raters. To assess the IRR, we used Gwet's AC$_1$ statistic [25] instead of the better-known kappa statistics. A rationale for this decision detailing the difference between Gwet's AC$_1$ and the kappa statistics can be found in Appendix A. We also reported values of Cohen's kappa (κ) [26] and Conger's κ [27] to ensure comparability with other studies.

We interpreted an $AC_1 < 0$ as poor, 0.0 to 0.20 as slight, 0.21 to 0.40 as fair, 0.41 to 0.60 as moderate, 0.61 to 0.80 as substantial, and 0.81 to 1.00 as almost perfect [28]. Additionally, we calculated pairwise raw agreement proportion before and after training for each PROBAST domain and the overall ROB rating. To quantify the training effect at the multi-rater level, we calculated the difference in agreement between AC_1 estimates after training and before training (ΔAC_1). We bootstrapped ΔAC_1 using bias correction and acceleration to obtain 95% confidence intervals (CIs) [29]. Analyses were performed in R version 4.2.1 [30].

3. Results

3.1. Study Characteristics

The PROBAST assessment of the forty-two studies [31–72] resulted in a low overall ROB rating for only one study (2%), while seven studies (17%) received an unclear ROB rating and thirty-four studies (81%) a high ROB rating [24]. The domain "outcome" contributed the highest proportion (n = 37; 88%) of low ROB ratings among all four domains in our investigation. The set of studies before and after training was similar regarding the overall ROB rating. Figure 2 shows the distribution of the overall and domain-specific ROB ratings in the two sets of studies assessed before and after the training. Details of the individual ROB rating results can be found in the Supplementary Table S1.

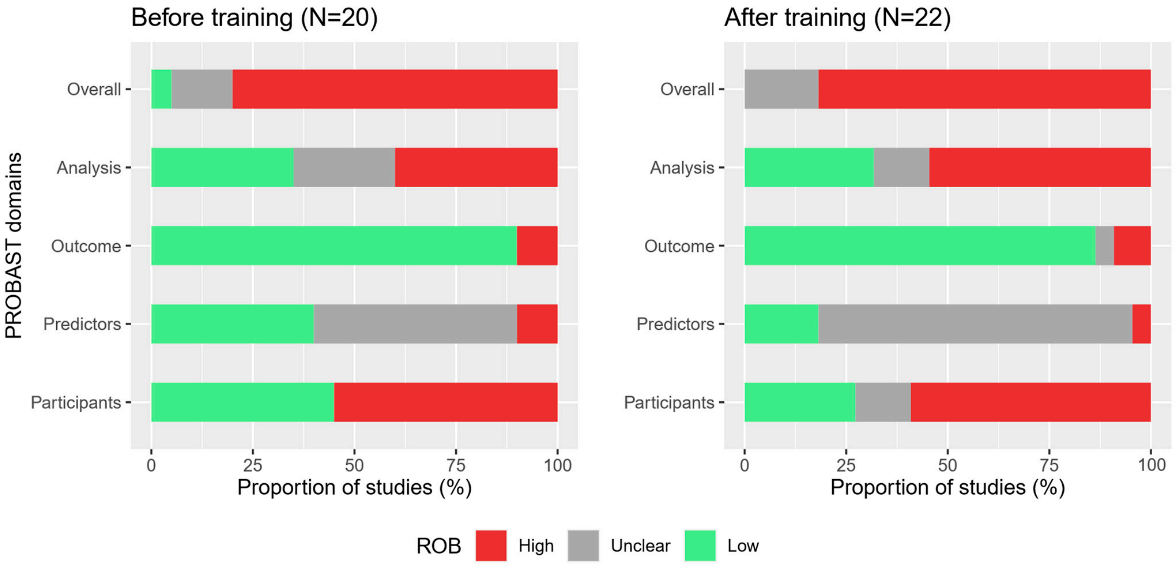
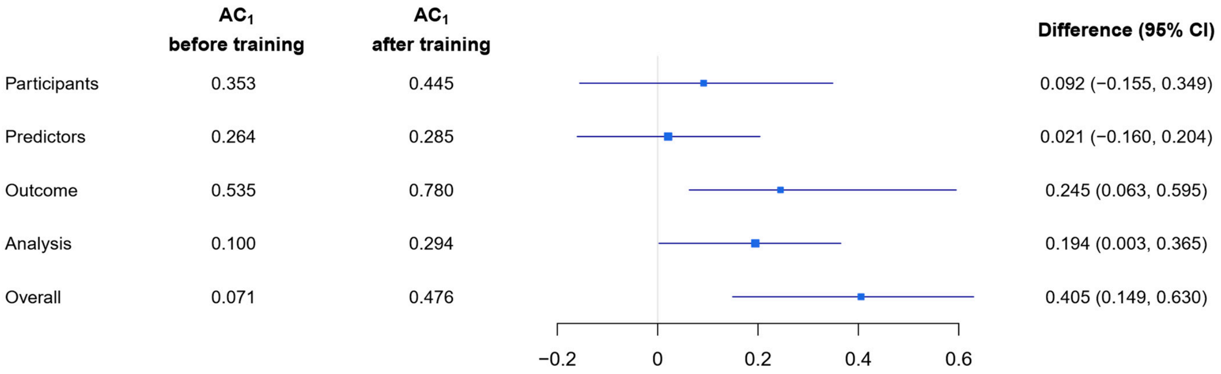


Figure 2. Overall and domain-specific ROB ratings of studies assessed before (n = 20) and after training (n = 22).

3.2. Multi-Rater Agreement

Figure 3 shows the multi-rater agreement before and after training for the four PROBAST domains and the overall ROB rating. Values of AC_1 before training ranged from 0.071 to 0.535. After training, the agreement ranged from 0.294 to 0.780. The highest agreement was observed in the outcome domain. We observed a significant improvement in the agreement after training compared to before training for the overall ROB rating ($\Delta AC_1 = 0.405$; 95%-CI 0.149–0.630) and the domains "outcome" ($\Delta AC_1 = 0.245$, 95%-CI 0.063–0.595) and "analysis" ($\Delta AC_1 = 0.194$; 95%-CI 0.003–0.365). For the domains "participants" and "predictors", the improvement in agreement was negligible. The corresponding estimates of Conger's κ and their difference between before and after training can be found in the Supplementary Figure S1.



	AC_1 before training	AC_1 after training	Difference (95% CI)
Participants	0.353	0.445	0.092 (−0.155, 0.349)
Predictors	0.264	0.285	0.021 (−0.160, 0.204)
Outcome	0.535	0.780	0.245 (0.063, 0.595)
Analysis	0.100	0.294	0.194 (0.003, 0.365)
Overall	0.071	0.476	0.405 (0.149, 0.630)

Figure 3. Multi-rater agreement in terms of AC_1 before and after training for the domain-specific and overall ROB ratings, as well as ΔAC_1 estimates with bootstrapped 95%-CI.

3.3. Pairwise Agreement

The distribution of AC_1 estimates for pairwise agreement before and after training for the domain-specific and overall ROB ratings is shown in the left panel of Figure 4. In addition, the distribution of ΔAC_1 estimates of pairwise agreement is presented in the right panel of the same figure. The detailed values of all estimates of pairwise agreements can be found in Tables S2–S6 in the supplement. The highest level of agreement, both before and after training, can be found in the domain "outcome". The median of the differences was greater than 0 for all domains, indicating a positive effect of the training. For the overall ROB rating, the median of ΔAC_1 was highest (0.427).

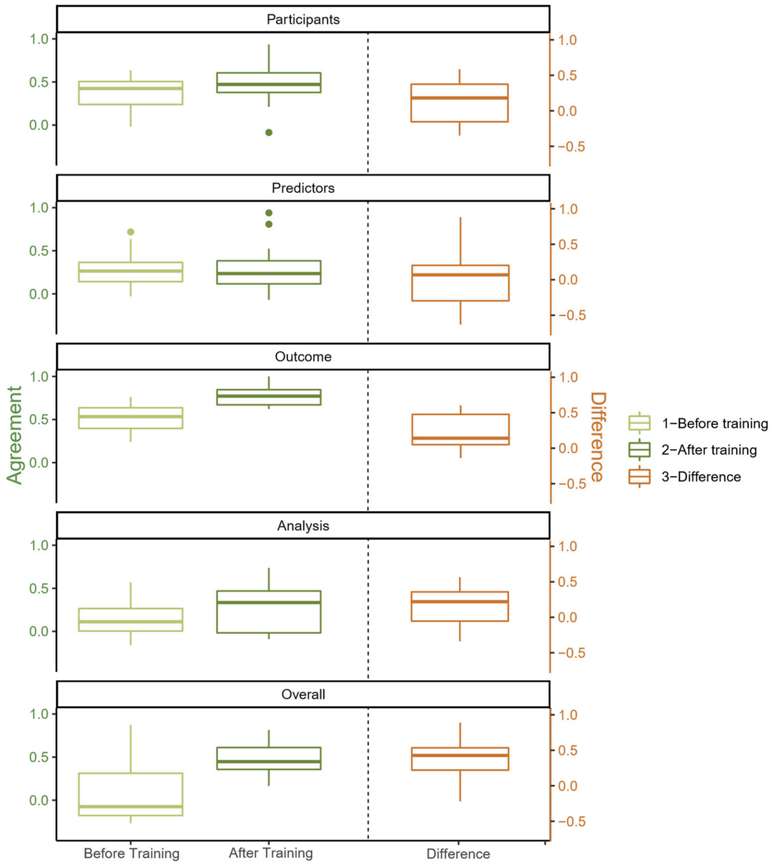


Figure 4. Distribution of pairwise inter-rater agreement in terms of box plots of AC_1 estimates before and after training for the domain-specific and overall ROB ratings (left part), as well as ΔAC_1 estimates (right part).

The agreement between individual raters and the consensual rating decision before and after training is shown in Table 1. With a few exceptions (n = 5, 17%), agreement with the consensus decision improved across all domains after training for all raters. The amount of improvement varied depending on rater and domain. The highest agreement between raters and consensus decision after training was found in the domain "outcome" (AC_1 0.683–0.947).

Table 1. Agreement in terms of AC_1 estimates between individual raters and consensus decision before and after training for the domain-specific and overall ROB rating.

	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6
Domain 1: Participants						
Before training	0.730	0.148	0.181	0.173	0.260	0.652
After training	0.675	0.805	0.549	0.546	0.074	0.679
Domain 2: Predictors						
Before training	0.428	0.125	0.394	0.214	0.202	0.643
After training	0.636	0.698	0.243	0.308	0.314	0.726
Domain 3: Outcome						
Before training	0.572	0.588	0.278	0.510	0.774	0.647
After training	0.851	0.776	0.899	0.899	0.683	0.947
Domain 4: Analysis						
Before training	0.493	0.635	0.085	0.108	0.145	0.629
After training	0.606	0.740	0.222	0.413	−0.022	0.802
Overall						
Before training	0.562	0.479	0.216	−0.313	−0.256	0.694
After training	0.711	0.713	0.423	0.537	0.392	0.893

3.4. Comparison of Raw Agreement, Gwet's AC_1 and Cohen's κ for Mean Pairwise Agreement

Table 2 compares the mean pairwise raw agreement, mean AC_1 and mean Cohen's κ for all PROBAST domains and the overall ROB rating to ensure the comparability of our results with other studies that did not use the AC_1 as measure for the IRR. Mean values of the pairwise raw agreement proportion ranged from 0.377 to 0.630 before training and from 0.494 to 0.809 after training with highest values in the domain "outcome". Due to the adjustment for random agreement between raters, the mean values of the AC_1 and Cohen's κ for pairwise agreement were lower than the mean raw agreement, with κ values usually being considerably lower than AC_1 estimates due to imbalances of marginal distribution of rating results. For the domain "outcome", where the imbalance of the marginal distribution of rating results was strongest, we observed the highest difference between AC_1 and Cohen's κ.

Table 2. Mean pairwise raw agreement, mean pairwise AC_1 and mean Cohen's κ before and after training for the domain-specific and overall ROB rating.

	Mean Raw Agreement		Mean Pairwise AC_1		Mean Cohen's κ	
	Before Training	After Training	Before Training	After Training	Before Training	After Training
Domain 1: Participants	0.530	0.615	0.357	0.464	0.167	0.396
Domain 2: Predictors	0.465	0.494	0.284	0.297	0.019	0.171
Domain 3: Outcome	0.637	0.809	0.534	0.776	0.183	0.310
Domain 4: Analysis	0.397	0.524	0.142	0.298	0.134	0.287
Overall	0.377	0.612	0.098	0.474	0.132	0.261

4. Discussion

Our results show that without guidance or specific training, the IRR of the PROBAST instrument was low, meaning that the ROB assessment of melanoma prediction studies was not reliable. Training sessions and customized guidance focusing on the implementation of the PROBAST instrument in our particular field of application, namely melanoma prediction studies, significantly improved the agreement for the overall and two domain-specific ROB ratings, which substantiates the need for intensive as well as disease- and study-type-specific training before using the tool.

Slight to moderate agreement was found before training both at two-rater (mean pairwise AC_1: 0.098–0.534) and at multi-rater level (AC_1: 0.071–0.535). However, there were substantial differences depending on the domain. In domains requiring high levels of subjective judgment and methodological expertise, such as in the domain “analysis” (mean pairwise AC_1: 0.142; multi-rater AC_1: 0.100), agreement was lowest. There was also poor agreement on the overall ROB rating (mean pairwise AC_1: 0.098, multi-rater AC_1: 0.071). We observed the highest level of agreement for the domain “outcome” (mean pairwise AC_1: 0.534, multi-rater AC_1: 0.535), which is the domain requiring less complex and subjective evaluations than other domains as there is an established definition of the outcome, here cutaneous melanoma, with standard diagnostic procedures that have been used in most studies. Furthermore, our study found that IRR varied widely depending on the pair of raters. The degree of variability was again dependent on the PROBAST domain. Especially for the overall rating, the IRR before training varied strongly across the fifteen different rater pairs (pairwise AC_1: −0.265–0.873). This clearly demonstrates the subjectivity of the PROBAST instrument and its rater dependency.

To the best of our knowledge, this is the first study assessing the reliability of the PROBAST instrument for prediction studies on a specific outcome. A previous study by Venema et al. [73], which focused on comparing a short form of PROBAST with the full-length PROBAST instrument in their capabilities to identify prediction models for cardiovascular diseases that perform poorly at external validation, also examined the IRR between two reviewers for the ROB assessment on these clinical prediction models. They reported a Cohen’s κ of 0.33, which is in line with our results and allows for the conclusion that the low IRR of the PROBAST tool is not a melanoma-specific problem. Several studies assessed the reliability of other ROB instruments, such as the Cochrane ROB tool and ROBIS. Some of them reported IRRs that were of a similarly low level as in our study [15,17,74,75]. Gates et al. [15] evaluated the IRR of the AMSTAR (A MeaSurement Tool to Assess systematic Reviews), AMSTAR 2, and ROBIS tools. While the IRR for AMSTAR/AMSTAR 2 was in a moderate to good range (AC_1: 0.5–0.8), the IRR of the ROBIS tool was similar to our results for PROBAST (AC_1: −0.2–0.6). Könsgen et al. [74] evaluated the IRR of the Cochrane ROB tool using Conger’s κ. Their results for the IRR (0.2–0.5) are slightly higher than our values for PROBAST (Conger’s κ: 0.0262–0.181), but are still in a fair range of agreement. Other studies, including Momen et al. [76] who studied the ROB-SPEO (Studies estimating Prevalence of Exposure to Occupational risk factors) tool, and Hoy et al. [77] who analyzed the IRR of the Hoy tool, report higher IRR estimates (Cohen’s κ: 0.5–0.8 and 0.5–0.9, respectively). However, in both cases the raters were familiar with the use of the tool. They had either been involved in the development of the instrument or received customized guidance before its use, so these IRR values are not comparable to our results before training.

Two possible explanations for the disagreement between raters are conceivable [75]: (i) a relevant piece of information is missed by one or more than one of the raters, (ii) interpretation of the same information is different owing to a subjective component. Training sessions and the development of a targeted and structured guidance manual address the problem of different interpretations of ROB items. Our results after training demonstrated that the IRR of the PROBAST instrument significantly improved in the second part of ROB assessments. At the start of the study, all raters were entirely inexperienced in using PROBAST, so there was a consistent baseline for quantifying the training effect. The largest net gain was achieved in the overall rating (ΔAC_1: 0.405) and the domain “outcome” (ΔAC_1: 0.245). When looking at the agreement of rater pairs, it became evident that for the vast majority of the rater pairs, the training improved the IRR. Other researchers have also shown that standardized training leads to a significant improvement in IRR for other ROB instruments [14].

However, high reliability does not imply correctness or validity of the tool. Focusing only on IRR would be insufficient, as high IRR does not necessarily imply that the ratings are correct [78]. Due to the absence of an external gold standard to validate our ROB assessments, we had no choice but to build on our consensus ratings, assuming these to be "correct". On account of a valid consensus process, where all raters jointly made final decisions, and involvement of two independent referees when no consensus could be reached through discussion, these ratings should be free of individual rater errors and bias. Our results show that, with a few exceptions, training improved agreement with the consensus decision in all domains and for all raters, making us confident that the consensus decisions were correct.

In practical applications comprising ROB assessments, it is not sufficient to simply use the checklist of a published ROB instrument. Specific guidance on how to implement a given instrument to a specific disease condition or study type is essential. Explanations, such as those available for PROBAST [2,13], can help to interpret the items correctly. However, explicit criteria for unclear and high ratings are rarely included, as they depend on the specific application. Therefore, before using the tool, it is important that users conduct training and/or create guidance manuals to address the main methodological problems common in their specific area of research. Valid decision rules for ROB ratings in a given research field require experienced epidemiologists specialized in the area of research that is involved.

Beyond defining decision rules, the rater group will achieve calibration through discussion and develop a common sense of when to apply a low or high ROB rating to a study. Beyond verifiable facts, each rater group develops its own evaluation standard for the ROB classification of studies by means of consensus discussions. Thus, a high IRR is always an indicator of a good calibration within the group. However, a high IRR for a ROB instrument in one rater group does not mean that other rater groups would arrive at the same ROB ratings with the same instrument for the same studies, as it may be that the other raters are "calibrated" differently.

Authors of systematic reviews and meta-analyses are strongly encouraged by guidelines such as PRISMA to incorporate ROB considerations into their process of research synthesis for quality improvement, namely reduction in bias in overall results [6]. However, ROB assessment and interpretation with regard to the strength of evidence assessment will be misleading if based on sub-optimal use of ROB instruments. Our results highlight that raters need to be aware of the limitations of ROB instruments. Detailed guidelines, decision rules, and transparency of the rating process are needed so that readers of systematic reviews can see how the tools were applied and are able to evaluate the results, that is, both the ROB tool and any specific thematic guidance used should ideally be published along with a systematic review.

Due to unbalanced marginal distributions in our ROB ratings, the use of any κ statistic would have potentially underestimated the IRR due to the well-known κ paradox [79,80]. In fact, individual rating categories were often disproportionately represented in some domains of our PROBAST rating. The domain "outcome" was rated as low in 82 out of 120 ratings (68%) before training and in 112 out of 132 ratings (85%) after training. The domains "participants" and "predictors" were rated as low in 63% of the ratings (75/120 and 76/120, respectively) before training. While Gwet's AC_1 offered in our case the advantage of addressing the problem of unbalanced marginal rating distributions, it also limits our comparability with other studies as this measure is used less frequently than the more widely used κ statistic. We reduced this limitation by additionally reporting Cohen's κ for pairwise agreement and Conger's multi-rater κ for our main results. Additionally, even if Gwet's AC_1 is still rather unknown, it has already been used by other researchers for the evaluation of inter-rater agreement [4,15,81].

A limitation to the generalizability of our findings regarding the training effect is that the magnitude of the effect is probably related to how detailed the decision rules were defined. These were developed based on the consensus over the first 20 studies, which evidently could not cover all possible reasons for unclear and high ratings for all domains faced in the remaining 22 studies. This translates to the notion that residual uncertainties will always be an issue arising with future primary literature, necessitating continual update of such customized guidance to original ROB tools. Furthermore, agreement may be higher among raters with a comparably high experience in research methods and epidemiology. The composition of our group was mixed in terms of the field of expertise and experience with systematic reviews and ROB assessments, which may have had some negative impact on IRR results. However, our mixed group of raters likely represents the range of raters that would typically be involved in such activities and thus our results provide a realistic impression of what can be expected from the PROBAST instrument in practice.

5. Conclusions

Without targeted guidance, the inter-rater agreement of the PROBAST instrument is low, questioning its use as an appropriate ROB instrument for prediction studies. Therefore, intensive training and guidance manuals with context-specific decision rules for high and unclear ratings are needed to correctly apply and interpret this ROB instrument and to ensure consistency of the ratings.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/jcm12051976/s1, Table S1: Number of low, high, and unclear ROB ratings, as well as consensus decisions per PROBAST domain and overall for all included studies; Figure S1: Multi-rater agreement in terms of Conger's κ before and after training for the domain-specific and overall ROB rating; Table S2: Pairwise inter-rater agreements in terms of AC_1 estimates for the ROB rating in the domain "participants"; Table S3: Pairwise inter-rater agreements in terms of AC_1 estimates for the ROB rating in the domain "predictors"; Table S4: Pairwise inter-rater agreements in terms of AC_1 estimates for the ROB rating in the domain "outcome"; Table S5: Pairwise inter-rater agreements in terms of AC_1 estimates for the ROB rating in the domain "analysis"; Table S6: Pairwise inter-rater agreement in terms of AC_1 estimates for all pairs of raters for the overall ROB rating.

Author Contributions: Conceptualization, O.G.; methodology, I.K. and O.G.; software, I.K.; validation, A.B.P., W.U. and O.G.; formal analysis, I.K.; investigation, I.K., S.M., K.D., T.S. and M.V.H.; resources, O.G.; data curation, I.K. and A.B.P.; writing—original draft preparation, I.K.; writing—review and editing, A.B.P., S.M., W.U., K.D., T.S., M.V.H. and O.G.; visualization, I.K.; supervision, O.G.; project administration, I.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. Regarding article processing charges we acknowledge financial support by Deutsche Forschungsgemeinschaft and Friedrich-Alexander-Universität Erlangen-Nürnberg within the funding programme "Open Access Publication Funding".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Most of the data is contained within the article and supplementary material. Data for the pairwise agreement are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Use of Gwet's AC_1 as Measure for Inter-Rater Reliability

A traditional, widely used measure to assess pairwise inter-rater reliability (IRR) is the kappa (κ) statistic by Cohen [26]. It corrects for "chance agreement" between the two raters by subtracting the amount of agreement resulting from a statistically independent rating of the two raters from the observed raw agreement and relating this difference to the maximally achievable one. For the case of more than two raters, several generalizations of the kappa measures exist that use different approaches to define chance agreement in the multi-rater situation [82]. An extension of Cohen's kappa which is based on a pairwise definition of chance agreement was proposed by Conger [27]. However, all kappa statistics have limitations and generate confusion in special situations. When there is strong imbalance in marginal distributions of the contingency tables describing the joint distribution of rating results kappa values tend to be low, although the absolute percentage of agreement between raters is high [83]. This phenomenon has been described in the literature as the kappa paradox [79,80]. The AC_1 statistic developed by Gwet [25] uses a different approach to capture chance agreement that is better suited for situations characterized by strong imbalance in marginal distributions. AC_1 estimates the true overall chance agreement in the presence of high agreement between reviewers, thus yielding IRR values better matching the impression of agreement observed in the contingency tables. As we faced strongly unbalanced marginal distributions in our rater data, we used Gwet's AC_1 in our analysis.

References

1. Sackett, D.L. Bias in analytic research. *J. Chronic Dis.* **1979**, *32*, 51–63. [CrossRef] [PubMed]
2. Moons, K.G.M.; Wolff, R.F.; Riley, R.D.; Whiting, P.F.; Westwood, M.; Collins, G.S.; Reitsma, J.B.; Kleijnen, J.; Mallett, S. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. *Ann. Intern. Med.* **2019**, *170*, W1–W33. [CrossRef] [PubMed]
3. The Cochrane Collaboration. *Cochrane Handbook for Systematic Reviews of Interventions*; Version 6.2; Cochrane: London, UK, 2021.
4. Jeyaraman, M.M.; Al-Yousif, N.; Robson, R.C.; Copstein, L.; Balijepalli, C.; Hofer, K.; Fazeli, M.S.; Ansari, M.T.; Tricco, A.C.; Rabbani, R.; et al. Inter-rater reliability and validity of risk of bias instrument for non-randomized studies of exposures: A study protocol. *Syst. Rev.* **2020**, *9*, 32. [CrossRef] [PubMed]
5. Bohlin, I. Formalizing Syntheses of Medical Knowledge: The Rise of Meta-Analysis and Systematic Reviews. *Perspect. Sci.* **2012**, *20*, 273–309. [CrossRef]
6. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *J. Clin. Epidemiol.* **2021**, *134*, 178–189. [CrossRef]
7. Ma, L.L.; Wang, Y.Y.; Yang, Z.H.; Huang, D.; Weng, H.; Zeng, X.T. Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: What are they and which is better? *Mil. Med. Res.* **2020**, *7*, 7. [CrossRef]
8. Wang, Z.T.K.; Allman-Farinelli, M.; Armstrong, B.; Askie, L.; Ghersi, D.; McKenzie, J.; Norris, S.; Page, M.; Rooney, A.; Woodruff, T.; et al. *A Systematic Review: Tools for Assessing Methodological Quality of Human Observational Studies*; National Health and Medical Research Council: Canberra, Australia, 2019.
9. Sterne, J.A.C.; Savovic, J.; Page, M.J.; Elbers, R.G.; Blencowe, N.S.; Boutron, I.; Cates, C.J.; Cheng, H.Y.; Corbett, M.S.; Eldridge, S.M.; et al. RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ* **2019**, *366*, l4898. [CrossRef]
10. Whiting, P.; Savovic, J.; Higgins, J.P.; Caldwell, D.M.; Reeves, B.C.; Shea, B.; Davies, P.; Kleijnen, J.; Churchill, R.; The Robis Group. ROBIS: A new tool to assess risk of bias in systematic reviews was developed. *J. Clin. Epidemiol.* **2016**, *69*, 225–234. [CrossRef]
11. Wolff, R.F.; Moons, K.G.M.; Riley, R.D.; Whiting, P.F.; Westwood, M.; Collins, G.S.; Reitsma, J.B.; Kleijnen, J.; Mallett, S.; Groupdagger, P. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann. Intern. Med.* **2019**, *170*, 51–58. [CrossRef]
12. PROBAST. Available online: https://www.probast.org/ (accessed on 22 October 2022).
13. de Jong, Y.; Ramspek, C.L.; Zoccali, C.; Jager, K.J.; Dekker, F.W.; van Diepen, M. Appraising prediction research: A guide and meta-review on bias and applicability assessment using the Prediction model Risk Of Bias ASsessment Tool (PROBAST). *Nephrology* **2021**, *26*, 939–947. [CrossRef]
14. da Costa, B.R.; Beckett, B.; Diaz, A.; Resta, N.M.; Johnston, B.C.; Egger, M.; Juni, P.; Armijo-Olivo, S. Effect of standardized training on the reliability of the Cochrane risk of bias assessment tool: A prospective study. *Syst. Rev.* **2017**, *6*, 44. [CrossRef] [PubMed]
15. Gates, M.; Gates, A.; Duarte, G.; Cary, M.; Becker, M.; Prediger, B.; Vandermeer, B.; Fernandes, R.M.; Pieper, D.; Hartling, L. Quality and risk of bias appraisals of systematic reviews are inconsistent across reviewers and centers. *J. Clin. Epidemiol.* **2020**, *125*, 9–15. [CrossRef] [PubMed]
16. Minozzi, S.; Cinquini, M.; Gianola, S.; Castellini, G.; Gerardi, C.; Banzi, R. Risk of bias in nonrandomized studies of interventions showed low inter-rater reliability and challenges in its application. *J. Clin. Epidemiol.* **2019**, *112*, 28–35. [CrossRef] [PubMed]
17. Minozzi, S.; Cinquini, M.; Gianola, S.; Gonzalez-Lorenzo, M.; Banzi, R. The revised Cochrane risk of bias tool for randomized trials (RoB 2) showed low interrater reliability and challenges in its application. *J. Clin. Epidemiol.* **2020**, *126*, 37–44. [CrossRef]
18. Kim, S.Y.; Park, J.E.; Lee, Y.J.; Seo, H.J.; Sheen, S.S.; Hahn, S.; Jang, B.H.; Son, H.J. Testing a tool for assessing the risk of bias for nonrandomized studies showed moderate reliability and promising validity. *J. Clin. Epidemiol.* **2013**, *66*, 408–414. [CrossRef]
19. Hartling, L.; Hamm, M.; Milne, A.; Vandermeer, B.; Santaguida, P.L.; Ansari, M.; Tsertsvadze, A.; Hempel, S.; Shekelle, P.; Dryden, D.M. *Validity and Inter-Rater Reliability Testing of Quality Assessment Instruments*; Agency for Healthcare Research and Quality: Rockville, MD, USA, 2012.

20. McHugh, M.L. Interrater reliability: The kappa statistic. *Biochem. Med.* **2012**, *22*, 276–282. [CrossRef]
21. Kaiser, I.; Pfahlberg, A.B.; Uter, W.; Heppt, M.V.; Veierod, M.B.; Gefeller, O. Risk Prediction Models for Melanoma: A Systematic Review on the Heterogeneity in Model Development and Validation. *Int. J. Environ. Res. Public Health* **2020**, *17*, 7919. [CrossRef]
22. Usher-Smith, J.A.; Emery, J.; Kassianos, A.P.; Walter, F.M. Risk prediction models for melanoma: A systematic review. *Cancer Epidemiol. Biomark. Prev.* **2014**, *23*, 1450–1463. [CrossRef]
23. Vuong, K.; McGeechan, K.; Armstrong, B.K.; Cust, A.E. Risk prediction models for incident primary cutaneous melanoma: A systematic review. *JAMA Derm.* **2014**, *150*, 434–444. [CrossRef]
24. Kaiser, I.; Mathes, S.; Pfahlberg, A.B.; Uter, W.; Berking, C.; Heppt, M.V.; Steeb, T.; Diehl, K.; Gefeller, O. Using the Prediction Model Risk of Bias Assessment Tool (PROBAST) to Evaluate Melanoma Prediction Studies. *Cancers* **2022**, *14*, 33. [CrossRef]
25. Gwet, K.L. Computing inter-rater reliability and its variance in the presence of high agreement. *Br. J. Math. Stat. Psychol.* **2008**, *61*, 29–48. [CrossRef] [PubMed]
26. Cohen, J.F. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
27. Conger, A.J. Integration and Generalization of Kappas for Multiple Raters. *Psychol. Bull.* **1980**, *88*, 322–328. [CrossRef]
28. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. [CrossRef]
29. Efron, B. Better Bootstrap Confidence-Intervals. *J. Am. Stat. Assoc.* **1987**, *82*, 171–185. [CrossRef]
30. *R Development Core Team R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2022.
31. Augustsson, A. Melanocytic naevi, melanoma and sun exposure. *Acta Derm. Venereol. Suppl.* **1991**, *166*, 1–34.
32. Bakos, L.; Mastroeni, S.; Bonamigo, R.R.; Melchi, F.; Pasquini, P.; Fortes, C. A melanoma risk score in a Brazilian population. *Bras. Derm.* **2013**, *88*, 226–232. [CrossRef]
33. Bakshi, A.; Yan, M.; Riaz, M.; Polekhina, G.; Orchard, S.G.; Tiller, J.; Wolfe, R.; Joshi, A.; Cao, Y.; McInerney-Leo, A.M.; et al. Genomic Risk Score for Melanoma in a Prospective Study of Older Individuals. *J. Natl. Cancer Inst.* **2021**, *113*, 1379–1385. [CrossRef]
34. Barbini, P.; Cevenini, G.; Rubegni, P.; Massai, M.R.; Flori, M.L.; Carli, P.; Andreassi, L. Instrumental measurement of skin colour and skin type as risk factors for melanoma: A statistical classification procedure. *Melanoma Res.* **1998**, *8*, 439–447. [CrossRef]
35. Cho, E.; Rosner, B.A.; Feskanich, D.; Colditz, G.A. Risk factors and individual probabilities of melanoma for whites. *J. Clin. Oncol.* **2005**, *23*, 2669–2675. [CrossRef]
36. Cho, H.G.; Ransohoff, K.J.; Yang, L.; Hedlin, H.; Assimes, T.; Han, J.; Stefanick, M.; Tang, J.Y.; Sarin, K.Y. Melanoma risk prediction using a multilocus genetic risk score in the Women's Health Initiative cohort. *J. Am. Acad. Derm.* **2018**, *79*, 36–41.e10. [CrossRef] [PubMed]
37. Cust, A.E.; Drummond, M.; Kanetsky, P.A.; Australian Melanoma Family Study, I.; Leeds Case-Control Study, I.; Goldstein, A.M.; Barrett, J.H.; MacGregor, S.; Law, M.H.; Iles, M.M.; et al. Assessing the Incremental Contribution of Common Genomic Variants to Melanoma Risk Prediction in Two Population-Based Studies. *J. Investig. Derm.* **2018**, *138*, 2617–2624. [CrossRef] [PubMed]
38. Cust, A.E.; Goumas, C.; Vuong, K.; Davies, J.R.; Barrett, J.H.; Holland, E.A.; Schmid, H.; Agha-Hamilton, C.; Armstrong, B.K.; Kefford, R.F.; et al. MC1R genotype as a predictor of early-onset melanoma, compared with self-reported and physician-measured traditional risk factors: An Australian case-control-family study. *BMC Cancer* **2013**, *13*, 406. [CrossRef] [PubMed]
39. Davies, J.R.; Chang, Y.M.; Bishop, D.T.; Armstrong, B.K.; Bataille, V.; Bergman, W.; Berwick, M.; Bracci, P.M.; Elwood, J.M.; Ernstoff, M.S.; et al. Development and validation of a melanoma risk score based on pooled data from 16 case-control studies. *Cancer Epidemiol. Biomark. Prev.* **2015**, *24*, 817–824. [CrossRef] [PubMed]
40. Dwyer, T.; Stankovich, J.M.; Blizzard, L.; FitzGerald, L.M.; Dickinson, J.L.; Reilly, A.; Williamson, J.; Ashbolt, R.; Berwick, M.; Sale, M.M. Does the addition of information on genotype improve prediction of the risk of melanoma and nonmelanoma skin cancer beyond that obtained from skin phenotype? *Am. J. Epidemiol.* **2004**, *159*, 826–833. [CrossRef] [PubMed]
41. English, D.R.; Armstrong, B.K. Identifying people at high risk of cutaneous malignant melanoma: Results from a case-control study in Western Australia. *Br. Med. J. Clin. Res. Ed.* **1988**, *296*, 1285–1288. [CrossRef]
42. Fang, S.; Han, J.; Zhang, M.; Wang, L.E.; Wei, Q.; Amos, C.I.; Lee, J.E. Joint effect of multiple common SNPs predicts melanoma susceptibility. *PLoS ONE* **2013**, *8*, e85642. [CrossRef]
43. Fargnoli, M.C.; Piccolo, D.; Altobelli, E.; Formicone, F.; Chimenti, S.; Peris, K. Constitutional and environmental risk factors for cutaneous melanoma in an Italian population. A case-control study. *Melanoma Res.* **2004**, *14*, 151–157. [CrossRef]
44. Fears, T.R.; Guerry, D.T.; Pfeiffer, R.M.; Sagebiel, R.W.; Elder, D.E.; Halpern, A.; Holly, E.A.; Hartge, P.; Tucker, M.A. Identifying individuals at high risk of melanoma: A practical predictor of absolute risk. *J. Clin. Oncol.* **2006**, *24*, 3590–3596. [CrossRef]
45. Fontanillas, P.; Alipanahi, B.; Furlotte, N.A.; Johnson, M.; Wilson, C.H.; andMe Research, T.; Pitts, S.J.; Gentleman, R.; Auton, A. Disease risk scores for skin cancers. *Nat. Commun.* **2021**, *12*, 160. [CrossRef]
46. Fortes, C.; Mastroeni, S.; Bakos, L.; Antonelli, G.; Alessandroni, L.; Pilla, M.A.; Alotto, M.; Zappala, A.; Manoorannparampill, T.; Bonamigo, R.; et al. Identifying individuals at high risk of melanoma: A simple tool. *Eur. J. Cancer. Prev.* **2010**, *19*, 393–400. [CrossRef] [PubMed]
47. Garbe, C.; Buttner, P.; Weiss, J.; Soyer, H.P.; Stocker, U.; Kruger, S.; Roser, M.; Weckbecker, J.; Panizzon, R.; Bahmer, F.; et al. Risk factors for developing cutaneous melanoma and criteria for identifying persons at risk: Multicenter case-control study of the Central Malignant Melanoma Registry of the German Dermatological Society. *J. Investig. Derm.* **1994**, *102*, 695–699. [CrossRef] [PubMed]

48. Garbe, C.; Kruger, S.; Stadler, R.; Guggenmoos-Holzmann, I.; Orfanos, C.E. Markers and relative risk in a German population for developing malignant melanoma. *Int. J. Derm.* **1989**, *28*, 517–523. [CrossRef] [PubMed]
49. Goldberg, M.S.; Doucette, J.T.; Lim, H.W.; Spencer, J.; Carucci, J.A.; Rigel, D.S. Risk factors for presumptive melanoma in skin cancer screening: American Academy of Dermatology National Melanoma/Skin Cancer Screening Program experience 2001–2005. *J. Am. Acad. Derm.* **2007**, *57*, 60–66. [CrossRef] [PubMed]
50. Gu, F.; Chen, T.H.; Pfeiffer, R.M.; Fargnoli, M.C.; Calista, D.; Ghiorzo, P.; Peris, K.; Puig, S.; Menin, C.; De Nicolo, A.; et al. Combining common genetic variants and non-genetic risk factors to predict risk of cutaneous melanoma. *Hum. Mol. Genet.* **2018**, *27*, 4145–4156. [CrossRef] [PubMed]
51. Guther, S.; Ramrath, K.; Dyall-Smith, D.; Landthaler, M.; Stolz, W. Development of a targeted risk-group model for skin cancer screening based on more than 100,000 total skin examinations. *J. Eur. Acad. Derm. Venereol.* **2012**, *26*, 86–94. [CrossRef]
52. Harbauer, A.; Binder, M.; Pehamberger, H.; Wolff, K.; Kittler, H. Validity of an unsupervised self-administered questionnaire for self-assessment of melanoma risk. *Melanoma Res.* **2003**, *13*, 537–542. [CrossRef]
53. Hubner, J.; Waldmann, A.; Eisemann, N.; Noftz, M.; Geller, A.C.; Weinstock, M.A.; Volkmer, B.; Greinert, R.; Breitbart, E.W.; Katalinic, A. Association between risk factors and detection of cutaneous melanoma in the setting of a population-based skin cancer screening. *Eur. J. Cancer Prev.* **2018**, *27*, 563–569. [CrossRef]
54. Kypreou, K.P.; Stefanaki, I.; Antonopoulou, K.; Karagianni, F.; Ntritsos, G.; Zaras, A.; Nikolaou, V.; Kalfa, I.; Chasapi, V.; Polydorou, D.; et al. Prediction of Melanoma Risk in a Southern European Population Based on a Weighted Genetic Risk Score. *J. Invest. Derm.* **2016**, *136*, 690–695. [CrossRef]
55. Landi, M.T.; Baccarelli, A.; Calista, D.; Pesatori, A.; Fears, T.; Tucker, M.A.; Landi, G. Combined risk factors for melanoma in a Mediterranean population. *Br. J. Cancer* **2001**, *85*, 1304–1310. [CrossRef]
56. MacKie, R.M.; Freudenberger, T.; Aitchison, T.C. Personal risk-factor chart for cutaneous melanoma. *Lancet* **1989**, *2*, 487–490. [CrossRef] [PubMed]
57. Mar, V.; Wolfe, R.; Kelly, J.W. Predicting melanoma risk for the Australian population. *Australas J. Derm.* **2011**, *52*, 109–116. [CrossRef] [PubMed]
58. Marrett, L.D.; King, W.D.; Walter, S.D.; From, L. Use of Host Factors to Identify People at High-Risk for Cutaneous Malignant-Melanoma. *Can. Med. Assoc. J.* **1992**, *147*, 445–452.
59. Nielsen, K.; Masback, A.; Olsson, H.; Ingvar, C. A prospective, population-based study of 40,000 women regarding host factors, UV exposure and sunbed use in relation to risk and anatomic site of cutaneous melanoma. *Int. J. Cancer* **2012**, *131*, 706–715. [CrossRef] [PubMed]
60. Nikolic, J.; Loncar-Turukalo, T.; Sladojevic, S.; Marinkovic, M.; Janjic, Z. Melanoma risk prediction models. *Vojn. Pregl.* **2014**, *71*, 757–766. [CrossRef] [PubMed]
61. Olsen, C.M.; Pandeya, N.; Thompson, B.S.; Dusingize, J.C.; Webb, P.M.; Green, A.C.; Neale, R.E.; Whiteman, D.C.; Study, Q.S. Risk Stratification for Melanoma: Models Derived and Validated in a Purpose-Designed Prospective Cohort. *J. Natl. Cancer Inst.* **2018**, *110*, 1075–1083. [CrossRef] [PubMed]
62. Penn, L.A.; Qian, M.; Zhang, E.; Ng, E.; Shao, Y.; Berwick, M.; Lazovich, D.; Polsky, D. Development of a melanoma risk prediction model incorporating MC1R genotype and indoor tanning exposure: Impact of mole phenotype on model performance. *PLoS ONE* **2014**, *9*, e101507. [CrossRef]
63. Quereux, G.; Moyse, D.; Lequeux, Y.; Jumbou, O.; Brocard, A.; Antonioli, D.; Dreno, B.; Nguyen, J.M. Development of an individual score for melanoma risk. *Eur. J. Cancer. Prev.* **2011**, *20*, 217–224. [CrossRef]
64. Richter, A.; Khoshgoftaar, T. Melanoma Risk Prediction with Structured Electronic Health Records. In Proceedings of the ACM-BCB'18: 9th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, Washington, DC, USA, 29 August–1 September 2018.
65. Smith, L.A.; Qian, M.; Ng, E.; Shao, Y.Z.; Berwick, M.; Lazovich, D.; Polsky, D. Development of a melanoma risk prediction model incorporating MC1R genotype and indoor tanning exposure. *J. Clin. Oncol.* **2012**, *30*, 8574. [CrossRef]
66. Sneyd, M.J.; Cameron, C.; Cox, B. Individual risk of cutaneous melanoma in New Zealand: Developing a clinical prediction aid. *BMC Cancer* **2014**, *14*, 359. [CrossRef]
67. Stefanaki, I.; Panagiotou, O.A.; Kodela, E.; Gogas, H.; Kypreou, K.P.; Chatzinasiou, F.; Nikolaou, V.; Plaka, M.; Kalfa, I.; Antoniou, C.; et al. Replication and predictive value of SNPs associated with melanoma and pigmentation traits in a Southern European case-control study. *PLoS ONE* **2013**, *8*, e55712. [CrossRef] [PubMed]
68. Tagliabue, E.; Gandini, S.; Bellocco, R.; Maisonneuve, P.; Newton-Bishop, J.; Polsky, D.; Lazovich, D.; Kanetsky, P.A.; Ghiorzo, P.; Gruis, N.A.; et al. MC1R variants as melanoma risk factors independent of at-risk phenotypic characteristics: A pooled analysis from the M-SKIP project. *Cancer Manag. Res.* **2018**, *10*, 1143–1154. [CrossRef]
69. Vuong, K.; Armstrong, B.K.; Drummond, M.; Hopper, J.L.; Barrett, J.H.; Davies, J.R.; Bishop, D.T.; Newton-Bishop, J.; Aitken, J.F.; Giles, G.G.; et al. Development and external validation study of a melanoma risk prediction model incorporating clinically assessed naevi and solar lentigines. *Br. J. Derm.* **2020**, *182*, 1262–1268. [CrossRef] [PubMed]
70. Vuong, K.; Armstrong, B.K.; Weiderpass, E.; Lund, E.; Adami, H.O.; Veierod, M.B.; Barrett, J.H.; Davies, J.R.; Bishop, D.T.; Whiteman, D.C.; et al. Development and External Validation of a Melanoma Risk Prediction Model Based on Self-assessed Risk Factors. *JAMA Derm.* **2016**, *152*, 889–896. [CrossRef]
71. Whiteman, D.C.; Green, A.C. A risk prediction tool for melanoma? *Cancer Epidemiol. Biomark. Prev.* **2005**, *14*, 761–763. [CrossRef]

72. Williams, L.H.; Shors, A.R.; Barlow, W.E.; Solomon, C.; White, E. Identifying Persons at Highest Risk of Melanoma Using Self-Assessed Risk Factors. *J. Clin. Exp. Derm. Res.* **2011**, *2*, 1000129. [CrossRef]
73. Venema, E.; Wessler, B.S.; Paulus, J.K.; Salah, R.; Raman, G.; Leung, L.Y.; Koethe, B.C.; Nelson, J.; Park, J.G.; van Klaveren, D.; et al. Large-scale validation of the prediction model risk of bias assessment Tool (PROBAST) using a short form: High risk of bias models show poorer discrimination. *J. Clin. Epidemiol.* **2021**, *138*, 32–39. [CrossRef]
74. Konsgen, N.; Barcot, O.; Hess, S.; Puljak, L.; Goossen, K.; Rombey, T.; Pieper, D. Inter-review agreement of risk-of-bias judgments varied in Cochrane reviews. *J. Clin. Epidemiol.* **2020**, *120*, 25–32. [CrossRef]
75. Hartling, L.; Hamm, M.P.; Milne, A.; Vandermeer, B.; Santaguida, P.L.; Ansari, M.; Tsertsvadze, A.; Hempel, S.; Shekelle, P.; Dryden, D.M. Testing the risk of bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. *J. Clin. Epidemiol.* **2013**, *66*, 973–981. [CrossRef]
76. Momen, N.C.; Streicher, K.N.; da Silva, D.T.C.; Descatha, A.; Frings-Dresen, M.H.W.; Gagliardi, D.; Godderis, L.; Loney, T.; Mandrioli, D.; Modenese, A.; et al. Assessor burden, inter-rater agreement and user experience of the RoB-SPEO tool for assessing risk of bias in studies estimating prevalence of exposure to occupational risk factors: An analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury. *Environ. Int.* **2022**, *158*, 107005. [CrossRef]
77. Hoy, D.; Brooks, P.; Woolf, A.; Blyth, F.; March, L.; Bain, C.; Baker, P.; Smith, E.; Buchbinder, R. Assessing risk of bias in prevalence studies: Modification of an existing tool and evidence of interrater agreement. *J. Clin. Epidemiol.* **2012**, *65*, 934–939. [CrossRef] [PubMed]
78. Pieper, D.; Jacobs, A.; Weikert, B.; Fishta, A.; Wegewitz, U. Inter-rater reliability of AMSTAR is dependent on the pair of reviewers. *Bmc Med. Res. Methodol.* **2017**, *17*, 98. [CrossRef]
79. Byrt, T.; Bishop, J.; Carlin, J.B. Bias, Prevalence and Kappa. *J. Clin. Epidemiol.* **1993**, *46*, 423–429. [CrossRef]
80. Feinstein, A.R.; Cicchetti, D.V. High Agreement but Low Kappa. 1. The Problems of 2 Paradoxes. *J. Clin. Epidemiol.* **1990**, *43*, 543–549. [CrossRef]
81. Jeyaraman, M.M.; Rabbani, R.; Al-Yousif, N.; Robson, R.C.; Copstein, L.; Xia, J.; Pollock, M.; Mansour, S.; Ansari, M.T.; Tricco, A.C.; et al. Inter-rater reliability and concurrent validity of ROBINS-I: Protocol for a cross-sectional study. *Syst. Rev.* **2020**, *9*, 12. [CrossRef]
82. Martin Andres, A.; Alvarez Hernandez, M. Hubert's multi-rater kappa revisited. *Br. J. Math. Stat. Psychol.* **2020**, *73*, 1–22. [CrossRef]
83. Konstantinidis, M.; Le, L.W.; Gao, X. An Empirical Comparative Assessment of Inter-Rater Agreement of Binary Outcomes and Multiple Raters. *Symmetry* **2022**, *14*, 262. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.