

## Extent, impact, and mitigation of batch effects in tumor biomarker studies using tissue microarrays

Konrad H. Stopsack,<sup>1,2</sup> Svitlana Tyekucheva,<sup>3,4</sup> Molin Wang,<sup>1,3,5</sup> Travis A. Gerke,<sup>6</sup> J. Bailey Vaselkiv,<sup>1</sup> Kathryn L. Penney,<sup>1,5</sup> Philip W. Kantoff,<sup>2</sup> Stephen P. Finn,<sup>7,8</sup> Michelangelo Fiorentino,<sup>1,9</sup> Massimo Loda,<sup>10</sup> Tamara L. Lotan,<sup>11</sup> Giovanni Parmigiani,<sup>3,4#</sup> Lorelei A. Mucci<sup>1#</sup>

# joint senior authors

<sup>1</sup> Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA

<sup>2</sup> Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY

<sup>3</sup> Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA,

<sup>4</sup> Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA

<sup>5</sup> Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA

<sup>6</sup> Department of Cancer Epidemiology, Moffitt Cancer Center, Tampa, FL

<sup>7</sup> Department of Pathology, St. James's Hospital, Dublin, Ireland

<sup>8</sup> Trinity College, Dublin, Ireland

<sup>9</sup> Pathology Unit, Addaari Institute, S. Orsola-Malpighi Hospital, Bologna, Italy

<sup>10</sup> Department of Pathology, Weill Cornell Medical College, New York, NY

<sup>11</sup> Department of Pathology, Johns Hopkins Medical Institutions, Baltimore, MD

*Correspondence:* Konrad H. Stopsack, Department of Epidemiology, Harvard T.H. Chan School of Public Health, 677 Huntington Ave, Boston, MA 02115, [stopsack@post.harvard.edu](mailto:stopsack@post.harvard.edu)

*Keywords:* tissue microarray; batch effects; measurement error; batchtma R package

*Running head:* Batch effects between tissue microarrays

1 **Abstract**

2 Tissue microarrays (TMAs) have been used in thousands of cancer biomarker studies. To what extent batch effects,  
3 measurement error in biomarker levels between slides, affects TMA-based studies has not been assessed  
4 systematically. We evaluated 20 protein biomarkers on 14 TMAs with prospectively collected tumor tissue from  
5 1,448 primary prostate cancers. In half of the biomarkers, more than 10% of biomarker variance was attributable to  
6 between-TMA differences (range, 1–48%). We implemented different methods to mitigate batch effects (R package  
7 *batchtma*), tested in plasmode simulation. Biomarker levels were more similar between mitigation approaches  
8 compared to uncorrected values. For some biomarkers, associations with clinical features changed substantially after  
9 addressing batch effects. Batch effects and resulting bias are not an error of an individual study but an inherent feature  
10 of TMA-based protein biomarker studies. They always need to be considered during study design and addressed  
11 analytically in studies using more than one TMA.

## 12 Introduction

13 Tissue microarrays (TMAs) were first developed in the 1990s as an efficient way to examine tissue-based  
14 biomarkers (1). Since then, TMAs have been used in thousands of studies to evaluate histologic and molecular  
15 biomarkers, mostly in cancer tissue. Individual TMAs consist of cylindrical cores from hundreds of tissue samples  
16 embedded in one recipient block (1, 2). Studies often include more than one TMA. Even when biomarker assays are  
17 well standardized and run conditions are diligently kept fixed, some TMA slides (batches) may have measurements  
18 systematically too low or too high, and some batches may have wider spread around the true values of the biomarker  
19 than others. In general, such batch effects can have a profound impact on the validity of biomarker studies, such those  
20 using RNA microarrays (3, 4). Contrary to popular belief, whether such measurement error induces upward or  
21 downward bias in results is not guaranteed to follow simple heuristics (5).

22 Whether and to what extent TMAs are affected by batch effects has not been empirically assessed. TMAs  
23 pose unique challenges. For example, when tumor tissue is collected prospectively for inclusion on TMAs, tumor  
24 characteristics may differ between batches due to nonrandom assignment of cases, as well as temporal trends in tumor  
25 risk factors, screening, and diagnosis. Differences in tissue processing or storage across tissue specimens may have  
26 differential impact on biomarkers. Including calibration samples for quality control is also more challenging for  
27 TMAs than, for example, assaying of blood samples, because repeat sections from a tumor may differ due to  
28 intratumoral heterogeneity rather than only batch effects.

29 In this study, we assess batch effects in a large set of centrally constructed TMAs from prostate cancer tissue  
30 from 1,448 men in two nationwide cohort studies. We quantify the extent to which protein biomarker variation could  
31 be explained by batch effects. We probe different methods for mitigating batch effects while maintaining true,  
32 “biological,” between-TMA variation, including in a plasmode simulation. Finally, we demonstrate the impact of  
33 handling batch effects on commonly performed biomarker analyses.

## 34 Results

35 **Extent and type of batch effects.** To evaluate the presence of batch effects in studies using TMAs, we studied tumor  
36 tissue from 1,448 men with primary prostate cancer on 14 TMAs (labeled “A” through “N”), each including multiple  
37 tumor cores from 47 to 158 patients per TMA (Figure 1). Multiple cores from the same tumor (usually 3) were always  
38 located on the same TMA.

39 TMAs were used to quantify 20 protein biomarkers (Figure 2). Biomarker values showed noticeable between-  
40 TMA variation, despite immunohistochemical staining having been conducted at the same time for all 14 TMAs. We  
41 estimated that across the 20 biomarkers, between-TMA variation explained between 1% and 48% of overall variation  
42 in biomarker levels (intraclass correlation coefficient, ICC), with half of the biomarkers having ICCs greater than 10%  
43 (Figure 2).

44 In an example biomarker, estrogen receptor alpha in nuclei of stromal cells (Figure 3), the means of the most  
45 extreme TMAs differed by 2.2 standard deviations in intensity of expression and variances differed up to 9.3-fold.  
46 Other biomarkers showed similar between-TMA variation by magnitude and by which TMAs had the most extreme  
47 values (Figure 4A). Likewise, we observed that not only means, but also variances of biomarker levels differed  
48 between TMAs, although patterns of heteroskedasticity appeared weaker than for means (Suppl. Figure 1). In contrast,  
49 we found little evidence for more complex patterns of batch effects, such that tumors with specific grade, stage, or  
50 year of diagnosis would have been particularly affected by between-TMA differences (Suppl. Table 1). Nevertheless,  
51 observations from the same TMAs tended to be clustered together when projected onto the first two principal  
52 components, capturing 27% of variance in all biomarkers (Figure 4B).

53 Some biomarkers were stained using automated staining systems, other stains were done manually (Figure 2).  
54 Moreover, the method of scoring, including human (eye) scoring and computer-assisted quantification, differed  
55 between biomarkers, as did the main quantitative score, typically a measure of staining intensity, a proportion of cells  
56 above an intensity threshold, or a combination of both (Figure 2). Notably, between-TMA differences were present  
57 with any of these approaches. For example, batch effects were not only present when considering intensities of  
58 biomarker staining, as for the estrogen receptor alpha and beta example. Even when setting cut-offs for staining

60 visible by eye and quantifying the number of stain-positive cells, 8% (95% CI, 2 to 15) of variance in estrogen  
61 receptor alpha positivity and 27% (95% CI, 11 to 42) of estrogen receptor beta positivity were attributable to between-  
62 TMA variation (Suppl. Figure 2). Our data do not allow distinguishing which of these approaches, if any, were less  
63 prone to batch effects.

64 In summary, we observed a large and concerning degree of between-TMA variation for several biomarkers  
65 that were quantified using different approaches, suggesting that addressing batch effects could significantly impact  
66 scientific inference.

67  
68 **Source of batch effects.** The noticeable proportion of variance attributable to TMAs could have two possibly co-  
69 existing explanations. First, that between-TMA differences in biomarkers reflect different patient and tumor  
70 characteristics that need to be retained. Second, that between-TMA differences are artifacts due to systematic  
71 measurement error that need to be removed (batch effects).

72 In support of the first hypothesis, there were noticeable differences in patient and tumor characteristics  
73 between TMAs that are likely associated with biomarker levels (Figure 1). Along with a 14-year range between the  
74 per-TMA medians of cancer diagnosis year, there were differences in the proportion of tumors with a Gleason score  
75 of 8 or higher (between 11% and 33%) and rates of lethal disease (between 2 and 16 events per 1000 person-years of  
76 follow-up).

77 In support of the second hypothesis, we observed that certain TMAs had consistently higher or lower  
78 biomarker values for the majority of tested biomarkers (Figure 4A). For example, the same batches that showed  
79 higher-than-average biomarker values for stathmin also had higher-than-average values for PTEN. This example is  
80 noteworthy because both markers were assayed together on the same section of each TMA using multiplex  
81 immunofluorescence, and stathmin would be expected to be expressed in more aggressive tumors with activation of  
82 the PI3K signaling pathway while PTEN expression would be expected to be low in the same tumors (6).

83 Further supporting the second hypothesis, we did not observe any meaningful reduction in ICCs when we  
84 considered tumors that had the same clinical features in terms of Gleason score and stage (Suppl. Figure 3). Moreover,  
85 the association between Gleason score and biomarker levels (Figure 2D) was considerably lower than between TMAs  
86 and biomarker levels, as underscored by less pronounced visual separation of principal components by Gleason score  
87 (Figure 4C) than by TMA (Figure 4B). Gleason score differences explained no more than 13% of variance in  
88 biomarker levels (for prostate-specific membrane antigen, PSMA; 95% CI for ICC, 0.02 to 0.29), and 13 of the 20  
89 biomarkers had ICCs by Gleason score of 1% or less (Suppl. Figure 4).

90 To directly disentangle both hypotheses, we further examined data on 10 tumors with a total of 53 tumor  
91 cores for which some cores were included on different TMAs (Figure 4D). These were not included in the previous  
92 analyses and had estrogen receptor scoring data. This design allowed us to estimate biomarker differences directly  
93 attributable to between-TMA variability within the same tumors while controlling for the between-core variability  
94 expected due to intratumoral heterogeneity. Of the total variance in estrogen receptor alpha levels, 30% (95% CI, 0 to  
95 67) was explained by between-TMA variation; for estrogen receptor beta, 24% (95% CI, 0 to 60) was explained by  
96 between-TMA variation. For comparison, between-tumor variation explained 37% (95% CI, 4 to 68) of the variance  
97 of estrogen receptor alpha levels and 26% (95% CI, 0 to 57) of the variance of estrogen receptor beta levels.

98 Collectively, while these observations highlighted moderate differences in clinical and pathological  
99 characteristics between TMAs, they suggested that between-TMA differences were largely due to batch effects.

100  
101 **Mitigation of batch effects.** We implemented six different approaches for batch effects mitigation and compared  
102 these to the uncorrected biomarker levels (Figure 3, Suppl. Figure 5). Two mitigation approaches, batch means  
103 (approach 2) and quantile normalization (approach 6), assumed no true difference between TMAs is arising from  
104 patient and tumor characteristics, while all other approaches attempted to retain such differences between TMAs. It is  
105 possible that the choice of mitigation approaches may be optimized using knowledge of the source of the batch effect.  
106 This would be the case if each method “specialized” in mitigating effect from specific sources. We have not  
107 investigated this possibility here. Overall, correlations between values adjusted by different approaches were higher

108 (mean Pearson  $r$ , 0.97 to 1.00) than between uncorrected values and corrected values ( $r$ , 0.90 to 0.95), regardless of  
109 mitigation approach (Figure 4E).

110 Approaches 2–7 reduced visible separation by batch on plots of the first two principal components (Suppl.  
111 Figure 6). Variance attributable to between-TMA differences decreased to ICCs of <1% for all markers (Suppl.  
112 Table 2). An exception was the quantile regression-based approach 5; the ICCs after this approach remained up to  
113 10%. This method does not explicitly address differences in means between batches but allows associations between  
114 clinical and pathological factors and biomarker levels to differ at high and low quantiles (Suppl. Figure 7).

115 The differences between uncorrected values and batch effect-corrected values were remarkably similar  
116 between the mean-based approaches using approaches 2 (simple means), 3 (standardized batch means), and 4 (inverse  
117 probability-weighted batch means; Suppl. Figure 8). Consequently, batch effect-corrected values by approaches 2–4  
118 were highly correlated (Figure 4E). All mean-only batch effect mitigations also gave the same results when fitting  
119 outcome models stratified by batch (Suppl. Figure 9). However, batch-specific results differed for approaches that  
120 targeted between-batch differences in the variance of biomarkers.

121  
122 **Validating batch effect mitigation in plasmode simulation.** To compare the performance of the different batch  
123 mitigation approaches in a time-to-event analysis, we applied plasmode simulation (7) to fix the expected strength of  
124 the biomarker exposure–outcome relationship *a priori* before artificially introducing batch effects. The correlation  
125 structure between biomarker and confounders and between confounders and batches from the actual data (Suppl.  
126 Figure 10A, C) was preserved in the plasmode-simulated data. Likewise, across a range of hazard ratios for the  
127 biomarker–outcome association, confounder–outcome associations remained unchanged (Suppl. Figure 10B, D).

128 We first evaluated a setting in which we did not introduce batch effects (Figure 5A). Here, the observed  
129 hazard ratios without batch effect mitigation equaled the expected. When performing (unnecessary) batch effect  
130 mitigation, observed hazard ratios were still comparable with the expected hazard ratios (Figure 5D; see Suppl.  
131 Table 3 for confidence intervals).

132 We then introduced batch effects by adding batch-specific mean differences to the observed biomarker levels,  
133 yet without introducing differences in variance by batch (Figure 5B). Without batch effect mitigation, for a true  
134 hazard ratio of 3.0, the observed hazard ratio, averaged over simulations, was 2.17 (95% CI, 1.86 to 2.53), an  
135 underestimate by 28% (Figure 5E; Suppl. Table 3). In contrast, all mitigation approaches produced CIs that covered  
136 the expected hazard ratio (*e.g.*, approach 6 quantile normalization: hazard ratio, 3.03; 95% CI, 2.48 to 3.69).

137 When we introduced batch-specific differences in both means and in variances (Figure 5C), the observed  
138 hazard ratio without batch effect mitigation decreased to 1.90 (95% CI, 1.66 to 2.16) compared to the expected hazard  
139 ratio of 3.0 (Figure 5F; Suppl. Table 3). Batch effect mitigation methods that only focus on means (approaches 2–4)  
140 reduced but did not fully eliminate bias, with hazard ratios ranging between 2.67 and 2.70. Methods that address  
141 differences in both mean and variance resulted in less bias, with an observed hazard ratio of 3.11 (95% CI, 2.54 to  
142 3.81) for approach 6 (quantile normalization).

143 We also included two stratification-based approaches. Fitting survival models separately by batch, followed  
144 by inverse-variance pooling (approach 8) resulted in approximately unbiased estimates but was less efficient than  
145 other approaches, comes with a risk of sparse-data bias, and resulted in considerably wider confidence intervals in our  
146 simulation. Including batch as a stratification variable in a single Cox model (approach 9) was unbiased and efficient.  
147 A drawback of both stratification-based approaches is that they do not explicitly estimate batch effect-adjusted  
148 biomarker values that could be visualized directly.

149 Scenarios evaluated thus far were based on the actual, modest imbalance of confounders between batches and  
150 at most weak associations between the biomarker and confounders, resulting in weak confounding overall. We  
151 additionally introduced both modest and strong associations between biomarker and confounders and created more  
152 severe imbalance between batches (Suppl. Figure 11). In all scenarios, the ranking of mitigation methods was  
153 preserved (Suppl. Figure 12, Suppl. Tables 3 and 4), with the least bias obtained through quantile normalization  
154 (approach 6). Bias occurred when using uncorrected biomarker levels in the presence of any batch effects, except if  
155 there was no association between biomarker and outcome (*i.e.*, a hazard ratio of 1), and with mean-only approaches 2–  
156 4 if variance was also affected by batch effects. In no situation, except possibly with the quantile regression-based

157 approach 5, were estimates after batch effect mitigation farther from the expected values than results based on  
158 uncorrected biomarker levels.

159  
160 **Impact of batch effects.** To illustrate how batch effect mitigations alter the results of commonly conducted tumor  
161 biomarker analyses, we estimated how uncorrected and corrected biomarker levels were associated with Gleason  
162 score and with rates of lethal disease. For markers with little between-TMA variability (low ICCs) such as beta-  
163 catenin, there were no noticeable differences in associations between using unadjusted and adjusted biomarker levels  
164 irrespective of adjustment model, as expected from plasmode simulation. However, for markers with higher between-  
165 TMA variability (higher ICC) and stronger associations with the outcome, adjustment approaches led to noticeable  
166 differences (Figure 6). For example, uncorrected stathmin expression levels were not associated Gleason score  
167 (difference, 0.00 standard deviations per 1 grade-group increase; 95% CI, -0.05 to 0.05), while the difference in levels  
168 corrected according to approach 6 was 0.04 (95% CI, 0.00 to 0.07), suggesting a potentially qualitatively different  
169 interpretation (Figure 6A; Suppl. Table 5). In models for lethal disease (Figure 6B), the otherwise unadjusted hazard  
170 ratio for the highest quartile of the vitamin D receptor, compared to the lowest quartile, was 0.44 (95% CI, 0.23 to  
171 0.86); after mitigation using approach 6, the hazard ratio was 0.19 (95% CI, 0.09 to 0.40), suggesting that unadjusted  
172 biomarker levels could underestimate the prognostic association by 2.3-fold (Suppl. Table 6 and 7).

173  
174 **Discussion**

175 The key strength of using TMAs is their utility in parallelizing the assessment of biomarkers on a large number of  
176 tissue specimens (1). Similar to other high-throughput platforms, batch effects have to be considered in every TMA  
177 biomarker study. As we demonstrated, for some of the biomarkers, batch effects can be of substantial magnitude. We  
178 show that batch effect mitigation is possible and can enhance study findings.

179 In our study of prostate tumor specimens, between-TMA differences explained 10% or more of the variance  
180 in biomarker levels for half of the included biomarkers, considerably more than one of the strongest pathological  
181 features in prostate cancer, Gleason grade. All analytical mitigation approaches to reduce batch effects, whether they  
182 attempted to retain real differences between tumors from different TMAs or not, led to corrected biomarker levels that  
183 were more similar to each other than they were, in general, to the uncorrected biomarker levels. In drawing from a  
184 large set of protein tumors biomarkers in prostate cancer, we show how appropriately mitigating batch effects  
185 strengthens results and their validity for biomarkers affected by batch effects.

186 Ideally, batch effects between TMAs are minimized when designing a study. Standardizing how tumor  
187 samples are obtained, stored, processed, and assayed is critical, as are stratified or random allocation of tumor samples  
188 to different TMAs (3) when possible. However, the batch effects that we observed occurred despite all feasible  
189 standardization efforts. Moreover, samples will be collected sequentially, and TMAs may be constructed sequentially  
190 in large-scale prospective studies over time. There were modest differences in the clinical and pathological  
191 characteristics between our TMAs, an issue that may be inevitable in larger-scale biobank studies. Allocation schemes  
192 of tumors to TMAs that appear ideal retrospectively, for example by matching “cases” of lethal tumors with  
193 “controls” of non-lethal tumors, may not be feasible prospectively. Likewise, in few of the thousands of studies using  
194 TMAs will it be possible to reallocate tumors to different TMAs and repeat all pathology work merely to reduce  
195 implications of batch effects.

196 An additional challenge in the design phase is that tissue samples are inherently heterogeneous and cannot  
197 simply be diluted, like blood samples. “Quality control” tumor samples that could serve as a quantitative calibration  
198 series suitable for all future biomarkers do not exist. One potential strategy is to include cell lines that have been  
199 formalin-fixed and paraffin-embedded on each TMA. While cell lines address issues of heterogeneity, the cell lines  
200 are often genetically unique and as such may not be relevant for all biomarkers. Another potential approach is to  
201 include samples from the same tumor case across TMAs, which would allow for direct estimation of batch effects. For  
202 these reasons, a principled approach that anticipates batch effects and addresses them analytically is critical.

203 Beyond efforts to prevent batch effects during the study design phase, we suggest the following best practices  
204 when undertaking TMA-based tissue biomarker studies (Figure 7). First, the extent of potential batch effects should

205 be explored and reported in any study of cancer tissue using TMAs. Inspecting TMA slides and plots (Figure 3) (8) is  
206 important. Between-TMA variation should be quantified, for example by calculating ICCs, *i.e.*, to contrast variation of  
207 biomarker levels between TMAs compared to that between or within tumors (9). In our study, for half of the  
208 biomarkers, ICCs for between-TMA variation were low, at less than 10%, although the proportion of tolerable batch  
209 variation should be chosen based on the context. Whether TMAs differ in terms of average biomarker levels, low  
210 levels (possibly reflective of background), or variability between tumors will also inform what impact of between-  
211 TMA differences to expect.

212 Second, the source of between-TMA differences should be elucidated. Ideally, including multiple cores from  
213 the same tumors in more than one TMA will help estimating, again using ICCs, how biomarker levels vary between  
214 TMAs, between tumors, and within tumors. Alternatively, ICCs between TMAs can be estimated by restricting to or  
215 adjusting for tumor features associated with differences in the biomarker, if known. In our study, both approaches  
216 indicated that the largest share of between-TMA differences was likely due to batch effects rather than due to true  
217 differences between tumors on different TMAs. However, one should not simply assume this to be the case in other  
218 settings, and also explore between-tumor differences as one source of between-TMA differences.

219 In multidisciplinary team discussions (10), it may be possible to directly pinpoint the source of batch effects  
220 and eliminate its cause. All study steps, including the pre-analytic, analytic, and post-analytic phases, should be  
221 considered. If sources of batch effects can be identified, it is preferable that they be addressed directly during the pre-  
222 analytical or analytical phase, rather than applying the post-analytical methods that we have described here and that  
223 may not adequately incorporate knowledge on the source of batch effects. For example, if immunohistochemical  
224 staining was performed separately for each TMA, then immunohistochemistry and quantification should be repeated  
225 using new sections from all TMAs at once. Imaging of pathology slides can also be a source of batch effects (11), as  
226 could be image analysis. In other cases, particularly if such obvious reasons for batch effects were avoided through  
227 standardized processing, as in our examples, it may remain elusive whether batch effects were induced through subtle  
228 differences in how tumors were cored and embedded during TMA construction, how long they had been stored, how  
229 they were sectioned, how well the staining process was standardized, or how successfully background signal was  
230 eliminated during software-based quantification. Yet even biomarkers scored by manual quantification were not free  
231 from batch effects.

232 Third, if a biomarker is affected by batch effects and no “physical” remediation is possible, then analytical  
233 approaches should be used to reduce bias in results (3, 4). We demonstrate that in all plausible or exaggerated real-  
234 world scenarios, estimates after applying batch effect mitigations were consistently closer to the true underlying  
235 values than they were without. If batches do not only differ in terms of mean values, but also in terms of their  
236 variances, then methods that focus solely on means may be insufficient. A simple quantile-normalization-based  
237 approach was successful in reducing bias in real-world scenarios and could be preferred for its simplicity. It is  
238 important to note that any method tested in this study is preferable over not addressing batch effects, and thus the  
239 choice between methods should be secondary to the choice to address batch effects altogether. Only results for  
240 biomarkers that are affected by batch effects and that are associated with the outcome of interest will show large  
241 changes in estimates, as the vitamin D receptor in our example. In contrast, for the majority of our example  
242 biomarkers, results did not change appreciably because batch effects were low, associations with the outcome were  
243 close to null, or both (Figure 6).

244 We recommend that researchers openly address batch effects in their TMA-based studies: they are not an  
245 error of an individual study, but an inherent feature of TMA-based studies. Batch effects have long been recognized in  
246 studies of the transcriptome using microarrays and next-generation sequencing, where batch effect mitigations are a  
247 component of standard workflows (4, 12). Our data strongly suggest that protein biomarker studies using multiple  
248 TMAs are at risk of batch effects just like any other biomarker study. The extent of batch effects is difficult to predict,  
249 and empirical evaluation is necessary each time. Future studies should quantify between-TMA differences and, if they  
250 deem batch effect mitigations to be unnecessary, provide evidence for absence of batch effects, rather than merely  
251 assuming their absence. The methods that we provide facilitate appropriate migration of batch effects between TMAs  
252 and help strengthen scientific inference. It may be prudent to extend this approach to *in-situ* tissue biomarkers other  
253 than proteins, such as RNA *in-situ* hybridization, even if our study only demonstrated batch effects for proteins.

254 Having mitigated batch effects will allow researchers to focus on increasing study validity by addressing other sources  
255 of measurement error (5), selection bias (for example, from tumor biospecimen availability) (13), and confounding.

256  
257 **Methods**

258 **TMA and biomarkers.** Tumor tissue in this study was from men who were diagnosed with primary prostate cancer  
259 during prospective follow-up of two nationwide cohort studies. The Health Professionals Follow-up Study is an  
260 ongoing cohort study that enrolled 51,529 male health professionals across the United States in 1986. The Physicians'  
261 Health Study 1 and 2 were randomized-controlled trials of aspirin and dietary supplements, starting in 1982 with  
262 22,071 male physicians. Participants were diagnosed with and treated for prostate cancer at local health care providers  
263 across the United States. The study team collected formalin-fixed paraffin-embedded tissue specimens from radical  
264 prostatectomy and transurethral resection of the prostate (TURP), and study genitourinary pathologists performed  
265 central re-review, including standardized Gleason grading of full hematoxylin–eosin-stained slides from the tumor  
266 blocks (14). The study protocol was approved by the institutional review boards of the Brigham and Women's  
267 Hospital and Harvard T.H. Chan School of Public Health, and those of participating registries as required.

268 TMA were constructed using 0.6-mm tissue cores of the primary nodule or the nodule with the highest  
269 Gleason score (15), including three or more cores of tumor tissue per participant (tumor). For a subset of tumors,  
270 additional cores of tumor-adjacent, histologically normal-appearing prostate tissue were included. TMA were  
271 constructed at the same laboratory across a 10-year period, as tissue from cohort participants became available,  
272 without matching on patient or tumor characteristics and without randomization. Cores from the same participant  
273 were generally included on the same TMA, with exceptions noted below, and summarized as the mean. We include  
274 information from 14 prostate tumor tissue microarrays.

275 Immunostaining was generally performed separately for individual biomarkers yet always for all TMA at the  
276 same time. Detailed immunohistochemistry staining and quantification procedures for each marker have been  
277 published (6, 16–26) or are in preparation for estrogen receptor alpha (antibody SP1; Thermo Scientific, Waltham,  
278 MA) and an antibody (PPG5/10; Bio-Rad Laboratories, Hercules, CA) widely used to measure estrogen receptor beta.  
279 If batch effect mitigation approaches had been applied in the original studies, the uncorrected levels were retrieved.  
280 Right-skewed biomarker scores (Ki-67, pS6, TUNEL) were  $\log_e$  transformed. All biomarkers were scaled to mean 0  
281 and standard deviation 1 solely to facilitate comparisons of batch effects across markers; batch effect mitigation does  
282 not necessitate scaling and preserves absolute biomarker values.

283 **Extent and type of batch effects.** To visualize the extent of biomarker variation between TMA, we plotted  
284 uncorrected biomarker values by tumor, biomarker, and TMA. We summarized biomarker variation using the first two  
285 principal components (27). We calculated between-TMA mean differences and ratios of variances versus the first  
286 TMA. We tested if tumors with different clinical/pathological characteristics had higher biomarker levels in TMA  
287 with higher means (*i.e.*, multiplicative effect modification). For each biomarker and each clinical/pathological feature  
288 (ordinal Gleason score, ordinal stage, or calendar year of diagnosis), let  $Z_{ij}$  be the within-TMA  $z$ -score (mean 0,  
289 standard deviation 1) for tumor  $i$  from TMA  $j$ ;  $A_i$ , the clinical/pathological feature of tumor  $i$ ;  $B_j$ , the TMA-specific  
290 biomarker mean,  $r_j$ , the TMA-specific random effect, and  $e_{ij}$ , residual error. In the regression model  
291  $Z_{ij} = \beta_0 + \beta_1 A_i + \beta_2 B_j + \beta_3 A_i B_j + r_j + e_{ij}$ , we evaluated the  $\beta_3$  term to assess for multiplicative effect measure  
292 modification.

293 We calculated the proportion of variation in biomarker levels attributable to TMA using intra-class  
294 correlations (ICCs, also “repeatability” (9)) based on one-way random effects linear mixed models with an  
295 independent variance–covariance structure (9, 28) for  $Y_{ij}$ , the biomarker level per tumor  $i$  and TMA  $j$ ; where  $\beta_0$  is the  
296 biomarker mean;  $r_j$ , the random effect for TMA  $j$ ; and  $e_{ij}$ , the residual error:  $Y_{ij} = \beta_0 + r_j + e_{ij}$ . The ICC was defined  
297 as the proportion of between-TMA variance in the total variance:  $ICC = \frac{\text{var}(r)}{\text{var}(r) + \text{var}(e)}$ . 95% CIs for ICCs were  
298 obtained using parametric bootstrapping using 500 repeats (29).

299  
300

301 **Source of batch effects.** To directly distinguish between-TMA variation caused by batch effects from variation  
302 caused by differences in patient and tumor characteristics, we compared ICCs per biomarker overall to ICCs per  
303 biomarker when restricting analyses to a subset of tumors with the same clinical features. We also leveraged a small  
304 subset of tumors that had cores included on more than one TMA. Here, we used two-way random effects linear mixed  
305 models with independent variance-covariance structure (30) to separate between-TMA variation from between-core  
306 variation (*i.e.*, intratumoral heterogeneity) and residual modeling error:  $Y_{ijk} = \beta_0 + r_j + s_i + e_{ijk}$ . Compared to the  
307 model described earlier, this model additionally includes tumor-specific random effects  $s_i$ , and thus

308 
$$ICC = \frac{\text{var}(r)}{\text{var}(r) + \text{var}(s) + \text{var}(e)}$$

309  
310 **Mitigation of batch effects.** In addition to (1) *using uncorrected values*, we implemented eight different approaches  
311 to handle between-TMA batch effects:

312 (2) *Simple means.* This approach assumes that all TMAs, if not affected by batch effects, would have the  
313 same mean biomarker value. Differences in mean biomarker values per batch are corrected by estimating batch-  
314 specific mean effects (differences from the overall mean level) using a linear regression model with uncorrected  
315 biomarker values as the outcome and batch indicators as predictors. Corrected biomarker values are then obtained by  
316 subtracting batch-specific effects from the uncorrected biomarker values. Mean differences can either indicate the  
317 difference of each batch mean to the overall mean, as implemented here, or be defined by comparison to a reference  
318 batch.

319 (3) *Standardized means.* This approach estimates marginal means per batch using model-based  
320 standardization (in the epidemiologic sense). It assumes that batches with similar characteristics have the same means  
321 if not affected by batch effects. A linear regression model for a specific biomarker is fit, adjusting for tumor variables  
322 that differ in distribution between TMAs, similar to an approach described in the epidemiology literature by  
323 Rosner (31). Let  $Y_{ij}$  indicate the biomarker value for tumor  $i$  on TMA  $j$ ;  $B_j$ , TMA  $j$ ;  $C_1$  to  $C_m$ , the  $m$  covariates to be  
324 retained; and  $e_{ij}$ , the residuals. Then  $Y_{ij} = \beta_0 + \beta_j B_i + \gamma_1 C_1 + \dots + \gamma_n C_n + e_{ij}$ . Batch effect-corrected biomarker  
325 values can be obtained by subtracting batch-specific effects  $\beta_j$  predicted from the model above from uncorrected  
326 biomarker values.

327 We included the following clinical and pathologic variables as plausible sources of real between-TMA  
328 differences that should be retained in this approach, as well approaches 4–7: calendar year of diagnosis (linear),  
329 Gleason score (categorical: 5–6; 3+4; 4+3; 8; 9–10), and pathologic tumor stage (categorical: pT1/T2, pT3/T3a,  
330 pT3b/T4/N1, missing/tissue from transurethral resection of the prostate).

331 (4) *Inverse-probability weighted batch means.* Like the preceding approach, this approach assumes that  
332 batches with similar characteristics have the same means if not affected by batch effects. While the preceding  
333 approach assumes a constant association between covariates and biomarker levels across batches, this approach allows  
334 for associations to differ between batches. We first used inverse probability weighting for marginal standardization of  
335 the distribution of clinical and pathological features per batch to the distribution in the entire study population.  
336 Stabilized weights (32), truncated at the 2.5<sup>th</sup> and 97.5<sup>th</sup> percentile, were obtained through multinomial regression  
337 models, modeling the probability of assignment to a specific batch based on same clinical and pathological variables  
338 as in (3). In the weighted pseudopopulation, we then used a marginal linear model to estimate batch-specific mean  
339 differences, which were further used as in approaches 2 and 3.

340 (5) *Quantile regression.* This approach assumes that batches with similar characteristics have the same values  
341 for a selected set of batch-specific quantiles, in this application the upper and lower quartile. The lower quartile may  
342 be particularly affected by background noise, while the upper quartile may more likely reflect differences in batches  
343 due to covariates. A corollary of separately modeling the two differently is that clinical and pathological variables are  
344 allowed to have different effects on these quartiles (33). These assumptions contrast with approaches 2–4 that focus  
345 on mean levels only. We used quantile regression with the Frisch-Newton approach (34) separately for the first and  
346 third quartile of biomarker values with batch indicators to predict adjusted batch-specific quantile values with the  
347 same confounders as above. We then used the batch-specific 25<sup>th</sup> percentiles ( $y^{\tau=0.25}$ ) as the offset and the

348 interquartile range between the 25<sup>th</sup> and 75<sup>th</sup> percentiles ( $y^{\tau=0.75}$ ) as the scaling factor when batch-correcting  
349 biomarker levels. Let  $y_{ij}^*$  indicate the batch effect-corrected biomarker level for tumor  $i$  on TMA  $j$ ;  $y_{ij}$ , the  
350 uncorrected biomarker level for tumor  $i$  on TMA  $j$ ;  $\hat{y}_i^{\tau=x}$ ,  $x^{\text{th}}$  quantile of  $y$  for batch  $j$  (predicted value for  $y_j$  from  
351 unadjusted quantile regression);  $\hat{y}_j^{\tau=x,*}$  is  $\hat{y}_j^{\tau=x}$  with adjustment for confounders (predicted value for  $y_j$  from adjusted  
352 quantile regression); and  $\bar{y}^{\tau=x}$ , the  $x^{\text{th}}$  quantile of  $y$  overall. Then the corrected biomarker level is

$$y_{ij}^* = \frac{(y_{ij} - \hat{y}_j^{\tau=0.25}) (\bar{y}^{\tau=0.75} - \bar{y}^{\tau=0.25})}{(\hat{y}_j^{\tau=0.75,*} - \hat{y}_j^{\tau=0.25,*})} + \bar{y}^{\tau=0.25} - \hat{y}_j^{\tau=0.25,*} + \hat{y}_j^{\tau=0.25}$$

353 (6) *Quantile normalization*. This approach assumes that samples on all batches, if not affected by batch  
354 effects, would not only have the same mean and variance but also the same distribution of individual biomarker  
355 values. Uncorrected biomarker values are ranked within each batch and then ranks are replaced by the mean of values  
356 with the same rank across batches. We implemented quantile normalization using *limma* (35, 36).

357 A conceptually related approach, for example employed in molecular epidemiology (3, 10), would be to use  
358 within-batch ranks as the batch-corrected biomarker, often grouped into data-driven categories such as batch-specific  
359 quartiles. We did not further consider these derivatives because they do not retain absolute biomarker levels and can  
360 distort rank distances.

361 (7) *ComBat*. For comparison, we additionally included the ComBat algorithm, which like approach 4  
362 attempts to retain differences in batch means due to covariate differences; it is frequently applied together with  
363 approach 6. ComBat and its derivatives (12, 37, 38) were initially designed for microarray studies of gene expression,  
364 which include considerably more than one biomarker per sample. This property would typically limit their use for a  
365 protein biomarker quantified on a TMA unless a large number of biomarkers is available, as in our study. Mitigation  
366 depends on values of other biomarkers on the same batches. Even if multiple protein biomarkers were available, the  
367 non-randomly selected set of concomitantly available biomarkers may influence how batch effects are corrected.  
368 ComBat scales means and (optionally) variances while (optionally) retaining adjustment variables. ComBat is  
369 implemented using an empirical Bayes approach to achieve more favorable properties for small batches. The  
370 underlying model is similar to the regression above and has been emulated by a two-way analysis of variance (39). In  
371 using ComBat, we scaled both means and variances, adjusting for the same clinical and pathological variables as  
372 before. Because ComBat cannot handle biomarkers if they are missing on entire batches, we ran ComBat separately  
373 for groups of biomarkers measured on 8, 9, 10, or 14 TMAs.

374 (8) *Stratification with inverse-variance pooling*. An alternative approach to treating batch effects is to  
375 estimate outcome regression models separately by batch. This approach can be applied for a variety of regression  
376 models but does not result in corrected values. We pooled estimates with inverse variance-weighting to obtain  
377 summary estimates.

378 (9) *Stratification in Cox proportional hazards regression*. In a special case of stratification for time-to-event  
379 outcomes, Cox proportional hazards models allow for nonparametric batch effect mitigation by including batch as a  
380 stratification factor in the model specification. Comparisons are performed within batches. Unlike approach 8, only  
381 batch-specific baseline hazard functions but no batch-specific effects are estimated.

382 For approaches 1–7, we calculated Pearson correlation coefficients between uncorrected and corrected  
383 biomarker levels. Additionally, we repeated ICC and principal components analyses with corrected levels, and we  
384 estimated associations between Gleason score and biomarker levels after batch effect mitigation, stratifying by batch  
385 using approach 8.

386 Approaches 2–6, which result in batch effect-adjusted biomarker levels, are implemented in the R package  
387 *batchtma*, available at <https://stopsack.github.io/batchtma>.

388 **Plasmode simulation.** We evaluated the impact of batch effect mitigation approaches on known, investigator-  
389 determined biomarker–outcome associations using plasmode simulation, an approach used, for example, for  
390 evaluating confounding control for binary exposures in pharmacoepidemiology (7). We used observed data from all  
391 tumors included on the 14 TMAs to determine covariates (Gleason grade, pathological stage) and outcome (lethal  
392 disease), preserving the observed correlation structure (e.g., joint distribution of clinical characteristics across TMAs).

394 The only simulated elements were the biomarker levels and the strengths of biomarker–outcome associations (hazard  
395 ratios ranging from  $1/3$  to 3) that we fixed by simulating event times with flexible parametric survival models (40).  
396 Models used a baseline hazard function consisting of cubic splines with three knots (41). Group differences were  
397 based on proportional hazards for the observed confounder–outcome coefficients in the real data and the fixed  
398 biomarker (exposure)–outcome hazard ratios.

399 First, we used plasmode simulation to generate the fixed associations of the biomarker levels with the  
400 outcome, which are unknowable outside simulation studies, generating 200 plasmode datasets for each association.  
401 Second, we introduced batch effects. Batch effects were either only for the mean or for both mean and variance, using  
402 the actual standardized between-TMA differences in means and variances for the estrogen receptor-alpha protein, a  
403 biomarker with high ICCs. We also added batch effects for mean and variance with effect modification, making mean  
404 and variance changes due to batch effects strongly correlated with Gleason scores. Third, we calculated batch effect-  
405 adjusted biomarker levels using approaches 2–6. Lastly, we compared the expected hazard ratios for the biomarker–  
406 outcome association, fixed during simulations, with the estimated hazard ratios from Cox regression (with normality-  
407 based 95% CIs) before and after batch effect mitigation approaches 2–6 and using the two stratification-based  
408 approaches 8 and 9.

409 In sensitivity analyses, we simulated “moderate” associations between the biomarker and confounders  
410 (0.2 standard deviations difference in biomarker levels per Gleason grade group, 0.1 per stage category), “strong”  
411 associations (differences of 0.4 and 0.2 standard deviations, respectively; stronger than observed for any biomarker in  
412 our study), as well as “strong” associations and additional imbalance in Gleason grade and stage between TMAs (by  
413 excluding tumors with low grades from TMAs with higher-than-average means and excluding tumors with high stage  
414 from TMAs with low-than-average means), all before the four steps described above.

415  
416 **Impact of batch effects.** To quantify the impact of different approaches to batch-effect handling on scientific  
417 inference, we focused on two commonly employed types of analyses in biomarker research in prostate cancer: first, a  
418 cross-sectional analysis of Gleason score and biomarker levels, using linear regression models; second, a time-to-  
419 event analysis of biomarker levels and rates of lethal disease, using Cox proportional hazards regression. For  
420 graphing, exposures were modeled in five categories (Gleason scores) or using restricted cubic splines with three  
421 knots (all biomarkers in models for lethal disease). For numeric comparisons, Gleason scores were modeled as ordinal  
422 variables and biomarkers as linear variables to obtain one single estimate per model. We also categorized biomarkers  
423 into four quartiles and compared hazard ratios for lethal disease of the extreme quartiles. Models were designed only  
424 for investigating issues of batch effects and not for subject matter inference on specific biomarkers.

425  
426 **Data availability.** The batchtma R package is available at <https://stopsack.github.io/batchtma>. Code used to produce  
427 results this manuscript is at [https://github.com/stopsack/batchtma\\_manuscript](https://github.com/stopsack/batchtma_manuscript). Data are available for analysis on the  
428 Harvard FAS computing cluster through a project proposal for the Health Professionals Follow-up Study  
429 (<https://sites.sph.harvard.edu/hpfs/for-collaborators>).

### 430 431 Acknowledgments

432 We thank the participants and staff of the HPFS and the PHS for their valuable contributions. In particular, we would  
433 like to recognize the contributions of Liza Gazeeva, Siobhan Saint-Surin, Robert Sheahan, Betsy Frost-Hawes, and  
434 Eleni Konstantis. We would like to thank the following state cancer registries for their help: AL, AZ, AR, CA, CO,  
435 CT, DE, FL, GA, ID, IL, IN, IA, KY, LA, ME, MD, MA, MI, NE, NH, NJ, NY, NC, ND, OH, OK, OR, PA, RI, SC,  
436 TN, TX, VA, WA, and WY. The authors assume full responsibility for analyses and interpretation of these data. The  
437 HPFS is supported by the NIH (U01 CA167552). This research was funded in part by the Specialized Programs of  
438 Research Excellence program in Prostate Cancer P50 CA090381 and P50 CA211024, the NIH/NCI Cancer Center  
439 Support Grants P30 CA008748 and P30 CA006516, NIH/NCI grants 5R37 CA227190-02 (S. Tyekucheva,  
440 K.L. Penney, G. Parmigiani, and L.A. Mucci), R03 CA212799 (M.W.), R35 CA212799 (M.W.), and R01 CA131945

441 (M. Loda). The Department of Defense supported K.H. Stopsack (W81XWH-18-1-0330). K.H. Stopsack,  
442 K.L. Penney, S.P. Finn, T.L. Lotan, and L.A. Mucci are Prostate Cancer Foundation Young Investigators.  
443

444 **Competing Interests**

445 P.W. Kantoff reports the following disclosures for the last 24-month period: he has investment interest in Context  
446 Therapeutics LLC, DRGT, Placon, and Seer Biosciences; he is a company board member for Context Therapeutics  
447 LLC; he is a consultant/scientific advisory board member for Bavarian Nordic Immunotherapeutics, DRGT, GE  
448 Healthcare, Janssen, OncoCellMDX, Progenity, Seer Biosciences, and Tarveda Therapeutics; and he serves on data  
449 safety monitoring boards for Genentech/Roche and Merck.

450  
451 G. Parmigiani reports the following disclosures for the last 24-month period: he had investment interest in CRA  
452 Health; he is a co-founder and company board member of Phaeno Biotechnology; he is a consultant / scientific  
453 advisory board member for Konica-Minolta, Delfi Diagnostics and Foundation Medicine; he serves on a data safety  
454 monitoring board for Geisinger. None of these activities are related to the content of this article.

455 **References**

- 456 1. J. Kononen *et al.*, Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat. Med.* **4**, 844-847  
457 (1998).
- 458 2. O. P. Kallioniemi, U. Wagner, J. Kononen, G. Sauter, Tissue microarray technology for high-throughput molecular profiling of  
459 cancer. *Hum. Mol. Genet.* **10**, 657-662 (2001).
- 460 3. S. S. Tworoger, S. E. Hankinson, Use of biomarkers in epidemiologic studies: minimizing the influence of measurement error  
461 in the study design and analysis. *Cancer Causes Control* **17**, 889-899 (2006).
- 462 4. J. T. Leek *et al.*, Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* **11**, 733-  
463 739 (2010).
- 464 5. M. van Smeden, T. L. Lash, R. H. H. Groenwold, Reflection on modern methods: five myths about measurement error in  
465 epidemiological research. *Int. J. Epidemiol.* **49**, 338-347 (2020).
- 466 6. K. H. Stopsack *et al.*, Multiplex Immunofluorescence in Formalin-Fixed Paraffin-Embedded Tumor Tissue to Identify Single-  
467 Cell-Level PI3K Pathway Activation. *Clin. Cancer Res.* **26**, 5903-5913 (2020).
- 468 7. J. M. Franklin, S. Schneeweiss, J. M. Polinski, J. A. Rassen, Plasmode simulation for the evaluation of pharmacoepidemiologic  
469 methods in complex healthcare databases. *Comput Stat Data Anal* **72**, 219-226 (2014).
- 470 8. S. Manimaran *et al.*, BatchQC: interactive software for evaluating sample and batch effects in genomic data. *Bioinformatics* **32**,  
471 3836-3838 (2016).
- 472 9. S. Nakagawa, H. Schielzeth, Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. *Biol. Rev. Camb. Philos. Soc.* **85**, 935-956 (2010).
- 473 10. M. T. Marrone *et al.*, Adding the Team into T1 Translational Research: A Case Study of Multidisciplinary Team Science in the  
474 Evaluation of Biomarkers of Prostate Cancer Risk and Prognosis. *Clin. Chem.* **65**, 189-198 (2019).
- 475 11. S. Kothari *et al.*, Removing batch effects from histopathological images for enhanced cancer diagnosis. *IEEE J Biomed Health  
476 Inform* **18**, 765-772 (2014).
- 477 12. J. Leek, W. E. Johnson, A. Jaffe, H. Parker, J. Storey (2011) The SVA package for removing batch effects and other unwanted  
478 variation in high-throughput experiments.
- 479 13. L. Liu *et al.*, Utility of inverse probability weighting in molecular pathological epidemiology. *Eur. J. Epidemiol.* **33**, 381-392  
480 (2018).
- 481 14. J. R. Stark *et al.*, Gleason score and lethal prostate cancer: does  $3 + 4 = 4 + 3$ ? *J. Clin. Oncol.* **27**, 3459-3464 (2009).
- 482 15. A. Pettersson *et al.*, The TMPRSS2:ERG rearrangement, ERG expression, and prostate cancer outcomes: a cohort study and  
483 meta-analysis. *Cancer Epidemiol. Biomarkers Prev.* **21**, 1497-1509 (2012).
- 484 16. J. R. Rider *et al.*, Tumor expression of adiponectin receptor 2 and lethal prostate cancer. *Carcinogenesis* **36**, 639-647 (2015).
- 485 17. R. Flavin *et al.*, SPINK1 protein expression and prostate cancer progression. *Clin. Cancer Res.* **20**, 4904-4911 (2014).
- 486 18. M. Fiorentino *et al.*, Overexpression of fatty acid synthase is associated with palmitoylation of Wnt1 and cytoplasmic  
487 stabilization of beta-catenin in prostate cancer. *Lab. Invest.* **88**, 1340-1348 (2008).
- 488 19. T. U. Ahearn *et al.*, Calcium sensing receptor tumor expression and lethal prostate cancer progression. *J. Clin. Endocrinol.  
489 Metab.* 10.1210/jc.2016-1082, jc20161082 (2016).
- 490 20. Z. Ding *et al.*, SMAD4-dependent barrier constrains prostate cancer growth and metastatic progression. *Nature* **470**, 269-273  
491 (2011).
- 492 21. P. L. Nguyen *et al.*, Fatty acid synthase polymorphisms, tumor expression, body mass index, prostate cancer risk, and survival.  
493 *J. Clin. Oncol.* **28**, 3958-3964 (2010).
- 494 22. A. Pettersson *et al.*, MYC Overexpression at the Protein and mRNA Level and Cancer Outcomes among Men Treated with  
495 Radical Prostatectomy for Prostate Cancer. *Cancer Epidemiol. Biomarkers Prev.* **27**, 201-207 (2018).
- 496 23. J. L. Kasperzyk *et al.*, Prostate-specific membrane antigen protein expression in tumor tissue and risk of lethal prostate cancer.  
497 *Cancer Epidemiol. Biomarkers Prev.* **22**, 2354-2363 (2013).
- 498 24. P. K. Dhillon *et al.*, Aberrant Cytoplasmic Expression of p63 and Prostate Cancer Mortality. *Cancer Epidemiology Biomarkers  
499 & Prevention* **18**, 595-600 (2009).
- 500 25. W. K. Hendrickson *et al.*, Vitamin D receptor protein expression in tumor tissue and prostate cancer progression. *J. Clin. Oncol.*  
501 **29**, 2378-2385 (2011).
- 502 26. K. Zu *et al.*, Protein expression of PTEN, insulin-like growth factor I receptor (IGF-IR), and lethal prostate cancer: a  
503 prospective study. *Cancer Epidemiol. Biomarkers Prev.* **22**, 1984-1993 (2013).
- 504 27. S. Lê, F. Josse, F. Husson, FactoMineR: A Package for Multivariate Analysis. *Journal of Statistical Software* **25**, 1-18 (2008).
- 505 28. S. E. Hankinson *et al.*, Reproducibility of plasma hormone levels in postmenopausal women over a 2-3-year period. *Cancer  
506 Epidemiol. Biomarkers Prev.* **4**, 649-654 (1995).
- 507 29. M. A. Stoffel, S. Nakagawa, H. Schielzeth, S. Goslee, rptR: repeatability estimation and variance decomposition by generalized  
508 linear mixed-effects models. *Methods Ecol. Evol.* **8**, 1639-1644 (2017).
- 509 30. D. Bates, M. Machler, B. M. Bolker, S. C. Walker, Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical  
510 Software* **67**, 1-48 (2015).
- 511 31. B. Rosner, N. Cook, R. Portman, S. Daniels, B. Falkner, Determination of blood pressure percentiles in normal-weight children:  
512 some methodological issues. *Am. J. Epidemiol.* **167**, 653-666 (2008).
- 513 32. S. R. Cole, M. A. Hernan, Constructing inverse probability weights for marginal structural models. *Am. J. Epidemiol.* **168**, 656-  
514 664 (2008).
- 515 33. D. Bann, E. Fitzsimons, W. Johnson, Determinants of the population health distribution: an illustration examining body mass  
516 index. *Int. J. Epidemiol.* **49**, 731-737 (2020).

518 34. S. Portnoy, R. Koenker, The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error  
519 estimators. *Statistical Science* **12**, 279-300 (1997).

520 35. M. E. Ritchie *et al.*, limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids*  
521 *Res.* **43**, e47 (2015).

522 36. B. M. Bolstad, R. A. Irizarry, M. Astrand, T. P. Speed, A comparison of normalization methods for high density oligonucleotide  
523 array data based on bias and variance. *Bioinformatics* **19**, 185-193 (2003).

524 37. W. E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods.  
525 *Biostatistics* **8**, 118-127 (2007).

526 38. Y. Zhang, D. F. Jenkins, S. Manimaran, W. E. Johnson, Alternative empirical Bayes models for adjusting for batch effects in  
527 genomic studies. *BMC Bioinformatics* **19**, 262 (2018).

528 39. V. Nygaard, E. A. Rodland, E. Hovig, Methods that remove batch effects while retaining group differences may lead to  
529 exaggerated confidence in downstream analyses. *Biostatistics* **17**, 29-39 (2016).

530 40. M. J. Crowther, P. C. Lambert, Simulating biologically plausible complex survival data. *Stat. Med.* **32**, 4118-4134 (2013).

531 41. C. Jackson, flexsurv: A Platform for Parametric Survival Modeling in R. *Journal of Statistical Software* **70**, 1-33 (2016).

532

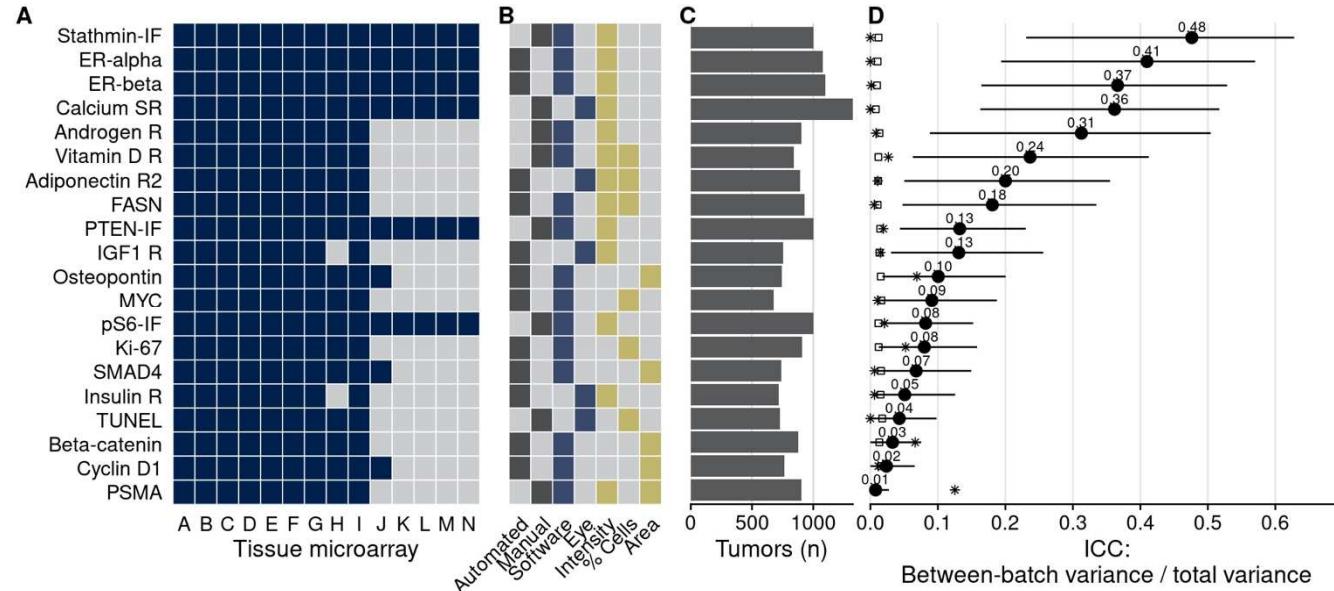
533 **Figures**

534

535 **Figure 1. Characteristics of men with prostate cancer with tissue included on the 14 tumor tissue microarrays.**

536

537 **A**, Calendar years of cancer diagnosis, with thick lines indicating median, boxes interquartile ranges, and whiskers 1.5  
538 times the interquartile range. **B**, Counts of tumors by Gleason score. **C**, Counts of tumors by pathological TNM stage  
539 (RP: radical prostatectomy). **D**, Rates of lethal disease (metastases or prostate cancer-specific death over long-term  
follow-up), with bars indicating 95% confidence intervals. As throughout, multiple cores are summarized per tumor.

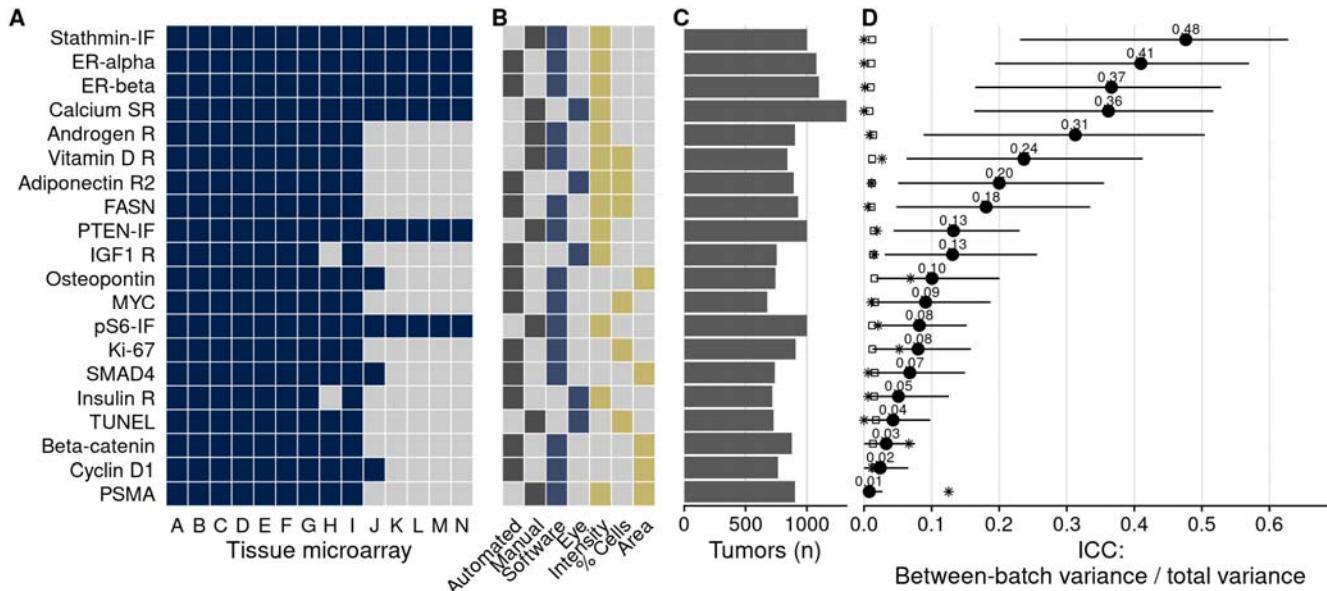


540

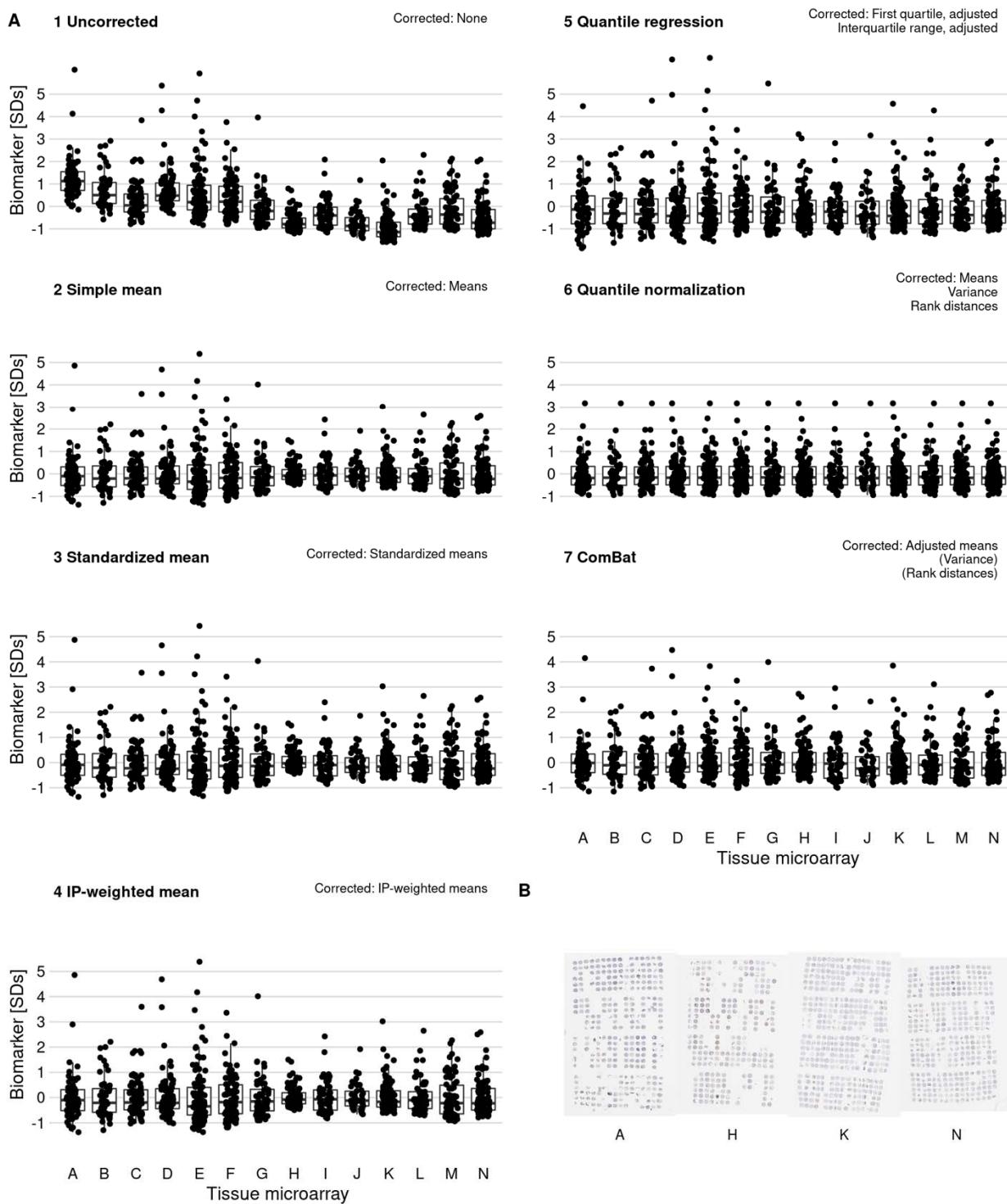
541

542 **Figure 2. Biomarkers stained, staining and scoring methods, and intraclass correlation coefficients (ICCs).** **A,**  
543 Tissue microarrays assessed for each marker (dark blue, yes). **B, Approach to staining biomarkers:** automated staining  
544 system vs. manual staining (gray, yes); quantification of biomarkers: software-based scoring vs. eye scoring (blue,  
545 yes); biomarker quality assessed: staining intensity, proportion of cells positive for the biomarker, area of tissue  
546 positive for the biomarker (yellow, yes). **C, Counts of tumors assessed for each biomarker.** **D, Between-tissue**

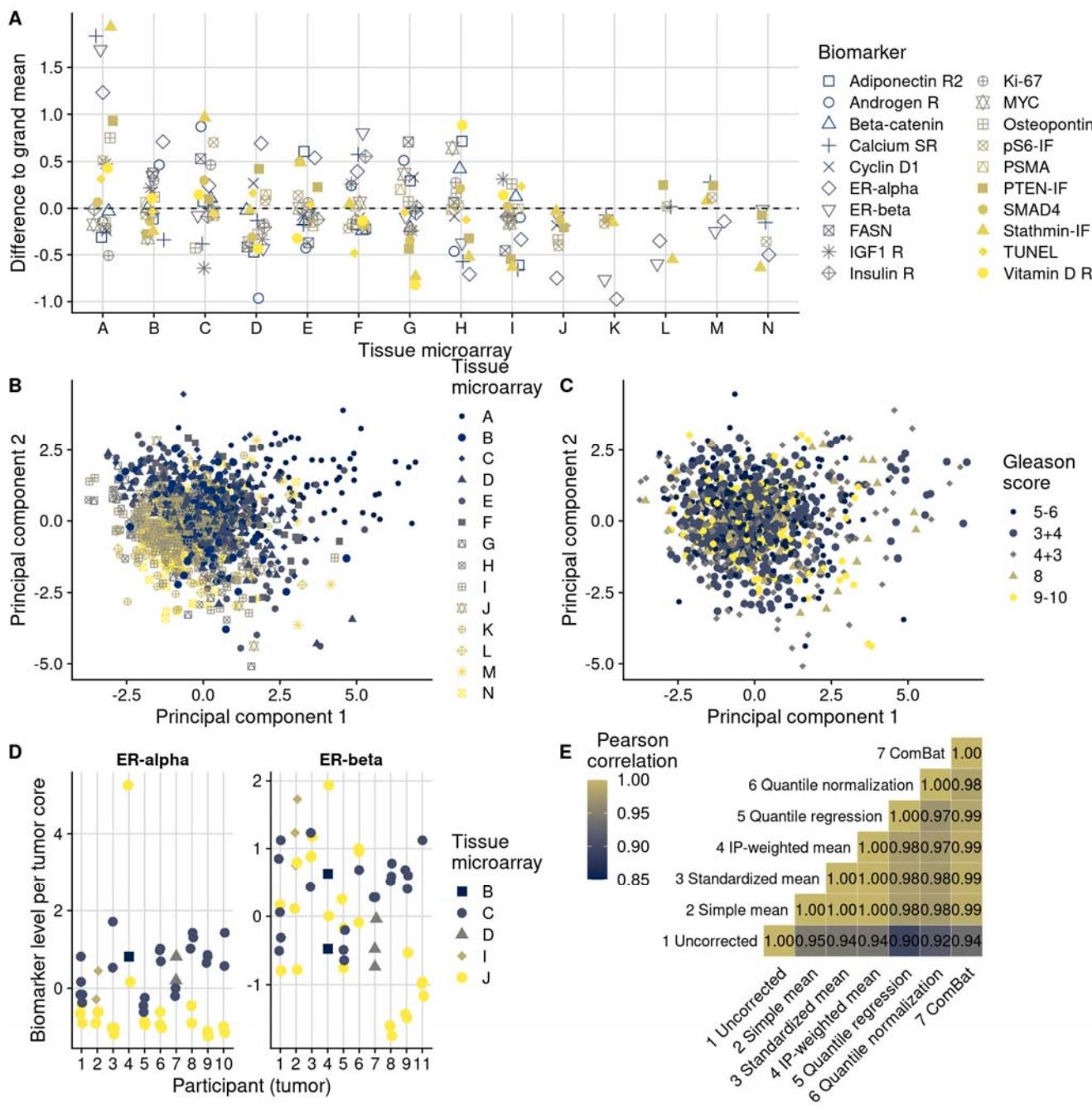
547 microarray ICCs (*i.e.*, proportion of variance explained by between-tissue microarray differences) for each biomarker,  
548 with 95% confidence intervals. Empty symbols indicate the 97.5<sup>th</sup> percentile of the null distribution of the ICC  
549 obtained by permuting tumor assignments to TMAs; asterisks indicate between-Gleason grade group ICCs.  
550 Biomarkers are arranged by descending between-tissue microarray ICC.



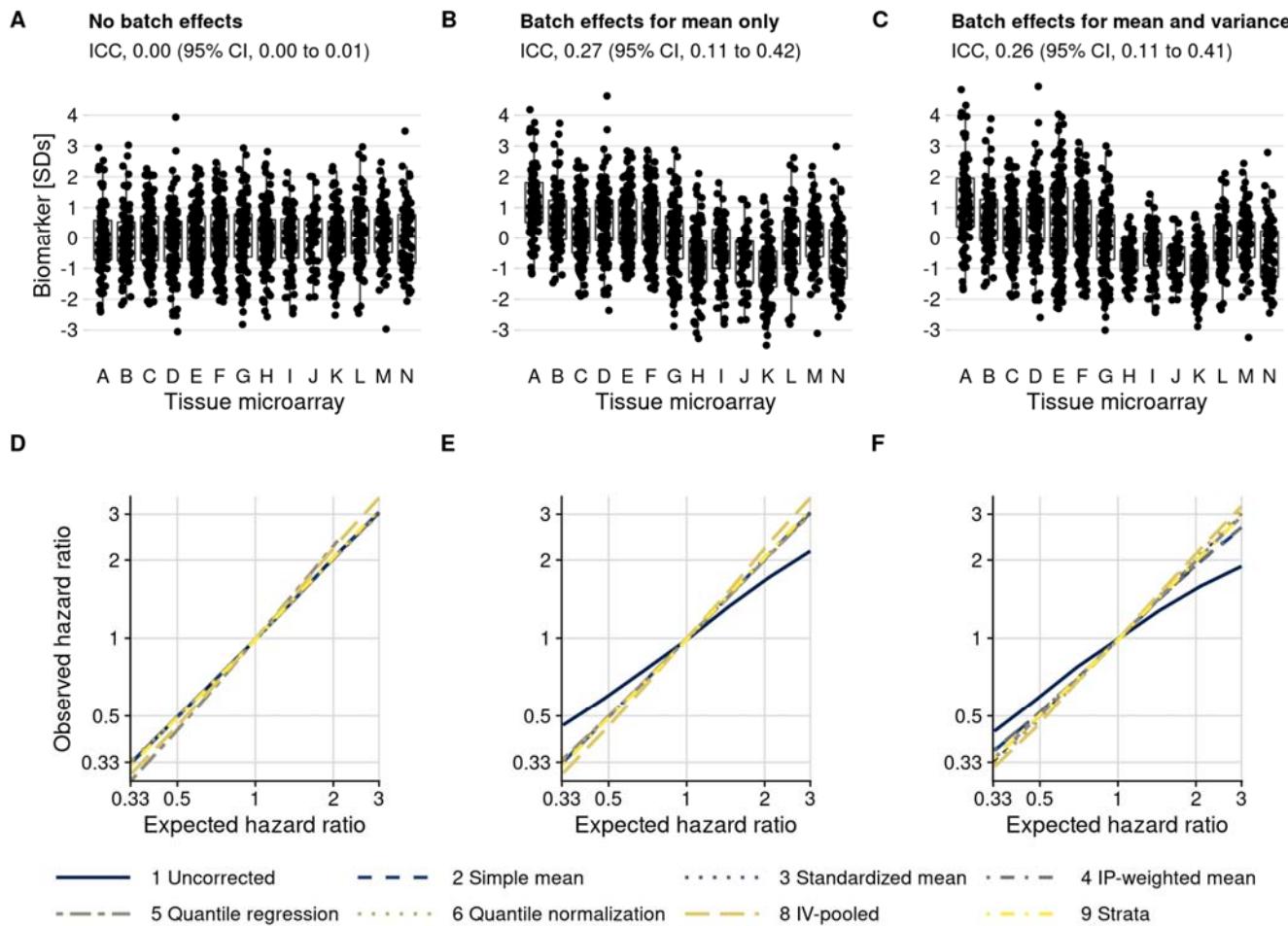
553 **Figure 3. Effect of batch effect mitigation on a biomarker with strong between-tissue microarray variation. A,**  
554 The protein biomarker estrogen receptor-alpha was quantified as staining intensity in nuclei of epithelial cells,  
555 averaged over all cores of each tumor. Each panel shows processed data for a specific approach to correcting batch  
556 effects. Notes in the upper right corner indicate which properties of batch effects were potentially addressed. Each  
557 data point represents one tumor. y-axes are standard deviations of the combined data for the specific method. Thick  
558 lines indicate medians, boxes interquartile ranges, and whisker length is 1.5 times the interquartile range. **B**, Example  
559 photographs of tissue microarrays; brown color indicates positive staining.



561 **Figure 4. Patterns, source, and remediation of batch effects.** **A**, Biomarker mean levels by tissue microarray, in  
 562 biomarker-specific standard deviations (y-axis). Each point is one tissue microarray. **B**, First two principal  
 563 components of biomarkers levels on all 14 tissue microarrays, with color/shape denoting tissue microarray. Each point  
 564 is one tumor. **C**, The same first two principal components, with color/shape denoting Gleason score. **D**, Per-core  
 565 biomarker levels for tumors with multiple cores included on two separate tissue microarrays, for estrogen receptor  
 566 (ER) alpha and beta, both in standard deviations. Each point is one tumor core. **E**, Pearson correlation coefficients  $r$   
 567 between uncorrected and corrected biomarker levels. Entries are averages across all markers.

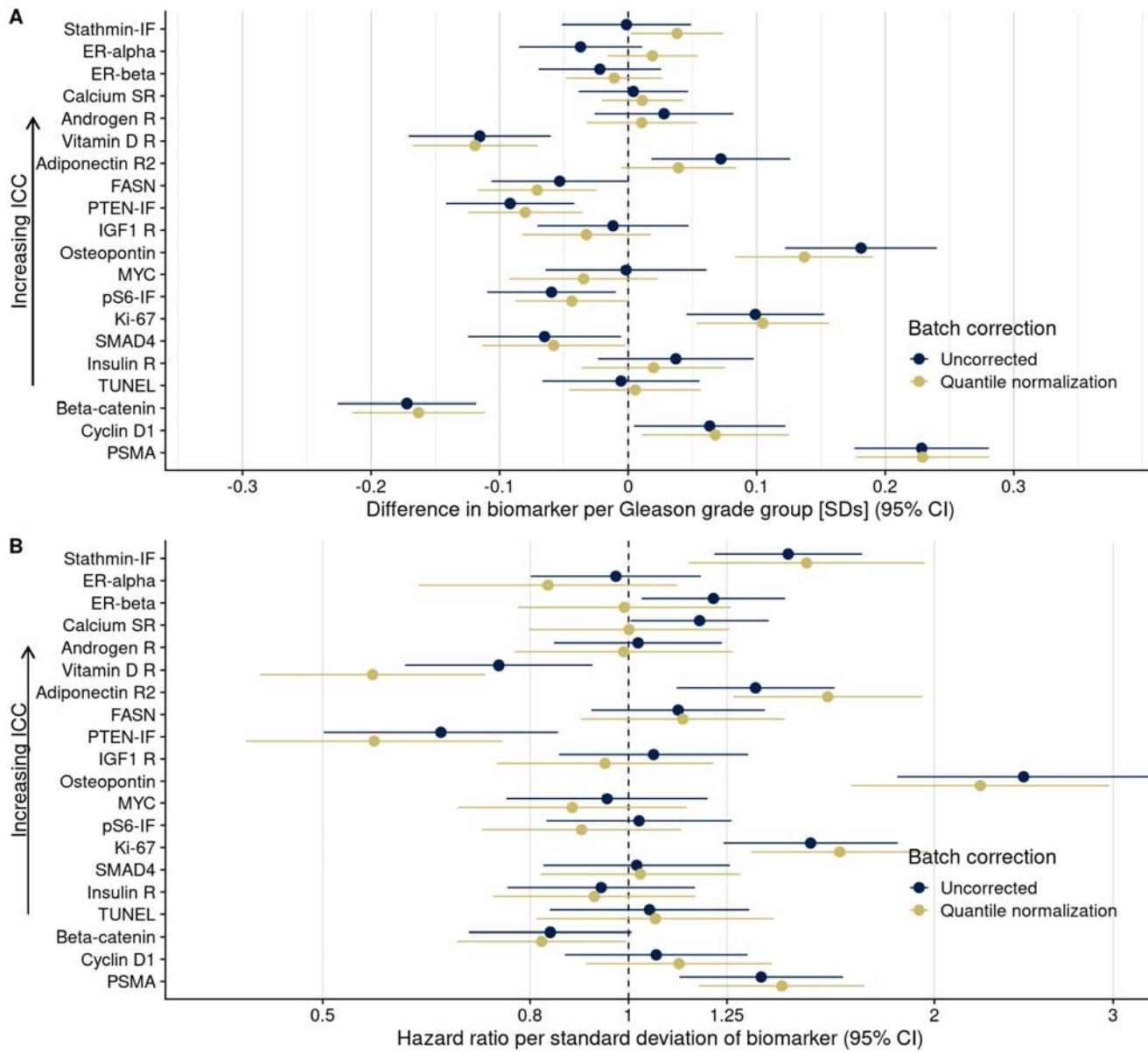


570 **Figure 5. Plasmode simulation results. A–C.** Biomarker levels by tissue microarray in three simulation scenarios;  
571 **D–F,** true versus observed hazard ratios for the biomarker–outcome association after alternative approaches to batch  
572 effect correction, with correction methods being numbered as in the Methods section. The solid blue line indicates no  
573 correction for batch effects. **A** and **D**, no batch effects; **B** and **E**, means-only batch effects; **C** and **F**, means and  
574 variance batch effects.



575  
576

577 **Figure 6. Consequences of batch effect mitigation on scientific inference. A**, Gleason score and biomarker levels  
578 (in standard deviations, per Gleason grade group). **B**, Biomarker levels and progression to lethal disease, with hazard  
579 ratios per one standard deviation increase in biomarker levels from univariable Cox regression models. In both panels,  
580 blue dots indicate estimates using uncorrected biomarker levels, yellow dots indicate batch effect-corrected levels,  
581 applying approach (5), quantile regression. Lines are 95% confidence intervals. Biomarkers are ordered by decreasing  
582 between-tissue microarray intraclass correlation coefficient (ICC).



583  
584

585 **Figure 7. Recommended workflow for anticipating and handling batch effects between tissue microarrays.**

586 Following prevention approaches at the design phase of a project, all tissue microarray-based studies should explore  
587 the potential for batch effects once a biomarker has been measured. Addressing batch effects should only be skipped  
588 there is no indication for their presence. Batch effect-corrected biomarker levels can easily be generated by the  
589 *batchtma* R package.

### Prevent batch effects

*Design Phase* 

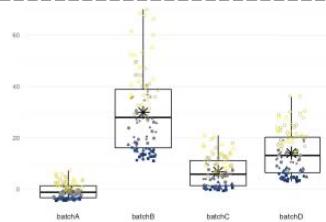
- If possible, consider during tissue allocation to TMAs:
  - Randomization *and/or* matching
  - Replicate cores from the same tumor on different TMAs

589 *Analysis Phase*

### Explore batch effects

`batchtma::: plot_batch()` 

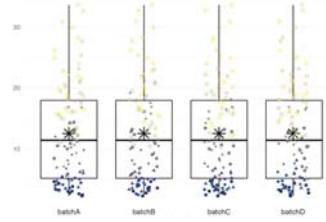
- Plot biomarker values by batch
- Quantify potential batch effects
- Elucidate potential sources



### Address batch effects

`batchtma::: adjust_batch()` 

- Fit outcome regression models separately by batch *or*
- Generate batch effect-corrected biomarker values



### Main analysis

590

591 **Legends for Supplementary Tables and Figures**

592 See separate markdown document, also available at

593 [https://stopsack.github.io/batchtma\\_manuscript/batchtma\\_manuscript\\_211203.html](https://stopsack.github.io/batchtma_manuscript/batchtma_manuscript_211203.html)

594

595 **Supplementary Table 1.** Interaction terms to test for multiplicative effect modification, *i.e.* whether batch effects  
596 more strongly affect tumors with more extreme clinical/pathological characteristics. The table shows point estimates  
597 (differences in biomarker levels), 95% confidence interval bounds, p-values, and false-discovery rates (FDR, in  
598 ascending order) for interaction terms between the within-batch normalized biomarker level and the potential effect  
599 modifier in linear models that have absolute biomarker levels in standard deviation units per biomarker as the  
600 outcome and also include main effects for the biomarker and the effect modifier (terms not shown).

601 **Supplementary Table 2.** Intraclass correlation coefficient (ICC) for between-batch variance for uncorrected  
602 biomarker levels (“1 Raw”) and biomarker levels after applying different correction methods.

603 **Supplementary Table 3.** Results from plasmode simulation according to type of induced batch effect, using the data  
604 correlation structure “moderate confounding.” For three fixed (“true”) hazard ratios for the biomarker–outcome  
605 association ( $1/3$ , 1, and 3), the observed hazard ratios after batch correction (with 95% confidence intervals) are shown.

606 **Supplementary Table 4.** Results from plasmode simulation according to data correlation structure, using the batch  
607 effect “mean and variance.” For three fixed (“true”) hazard ratios for the biomarker–outcome association ( $1/3$ , 1, and  
608 3), the observed hazard ratios after batch correction (with 95% confidence intervals) are shown.

609 **Supplementary Table 5.** Gleason grade—biomarker associations according to batch effect correction method. Point  
610 estimates from unadjusted linear regression models for biomarker values with Gleason score categories per 1 “grade  
611 group” increase as the predictor are shown (with 95% confidence intervals). For  $\log_e$ -transformed markers like Ki-67,  
612 estimates are differences on the  $\log_e$  scale.

613 **Supplementary Table 6.** Biomarker levels and lethal disease according to batch effect correction method. Hazard  
614 ratios (with 95% confidence intervals) per 1 standard deviation increase in the biomarker (linear) from unadjusted Cox  
615 regression models are shown.

616 **Supplementary Table 7.** Biomarker levels and lethal disease according to batch effect correction method. Unlike in  
617 the preceding table, the hazard ratios (with 95% confidence intervals) are contrasts comparing extreme quartiles  
618 (fourth compared to first quartile) from unadjusted Cox regression models.

619

620 **Supplementary Figure 1.** Ratios of variance per tissue microarray to the mean variance for each marker.

621 **Supplementary Figure 2.** Tissue microarrays and differences in % positivity, at the example of estrogen receptor  
622 alpha and beta, and variance in biomarker levels explained by between-tissue microarray differences (ICC).

623 **Supplementary Figure 3.** Intraclass correlation coefficients (ICCs), quantifying the proportion of variance in  
624 biomarker levels attributable to between-tissue microarray differences, across all tumors and after restriction to those  
625 378 tumors across tissue microarrays that have the same clinical/pathological characteristics in terms of Gleason score  
626 3+4 and prostatectomy stage pT1/T2.

627 **Supplementary Figure 4.** Intraclass correlation coefficients (ICCs), quantifying the proportion of variance in  
628 biomarker levels attributable to between-Gleason grade differences, by increasing ICC.

629 **Supplementary Figure 5.** Uncorrected compared with batch effect-corrected biomarker levels, for estrogen receptor  
630 alpha. Symbols and color indicate the tissue microarray.

631 **Supplementary Figure 6.** Principal components 1 and 2 after batch effect correction using (5) quantile regression for  
632 biomarkers available on all tissue microarrays. Symbol color and shape indicate the tissue microarray.

633 **Supplementary Figure 7.** Quantile-specific associations of confounders (clinical/pathological differences) with  
634 (uncorrected) biomarker levels of estrogen receptor alpha. Shown are regression coefficients for the 10<sup>th</sup>, 50<sup>th</sup>, and 90<sup>th</sup>  
635 percentiles as the outcomes of quantile regression models.

636 **Supplementary Figure 8.** Batch corrections per tissue microarray and method. The plot shows the difference  
637 between uncorrected and corrected values per batch, averaged across all biomarkers. IGF1-R was excluded because of  
638 missing values for some correction approaches. For batch correction approaches that only address the mean (*i.e.*, that  
639 subtract the same correction value from all biomarker values within each batch), only that difference is visible; for  
640 correction methods that address individual values within batches differently, batch-specific medians and interquartile  
641 ranges of differences between uncorrected and corrected values are visible.

642 **Supplementary Figure 9.** Biomarker differences, after batch effect correction methods, for a one-unit increment in  
643 Gleason score, stratified by tissue microarray. “Pooled” indicates estimates pooled over batches (TMAs) using  
644 inverse-variance weighting. “No stratification” indicates estimates without stratification. Note that for batch effect  
645 correction approaches that only address between-batch differences in means (approaches 2–4), estimates stratified by  
646 batch (and pooled estimates thereof) are the same.

647 **Supplementary Figure 10.** Data structures in the actual data and in 200 plasmode simulation datasets. **A**, Gleason  
648 scores and lethal prostate cancer (metastasis-free survival) in the actual data. **B**, Gleason scores and lethal prostate  
649 cancer in an example simulated dataset. Shaded areas indicate 95% confidence intervals. **C**, Pearson correlation  
650 coefficients between biomarker levels and confounders, and between confounders, across all simulated datasets.  
651 Correlation coefficients observed in the actual data are noted in the legend. **D**, Hazard ratios for the biomarker and the  
652 confounders in relation to lethal prostate cancer, pooling all simulated data sets. Confounder–outcome associations  
653 were simulated to correspond to their observed values in the actual data; exposure–outcome associations were  
654 simulated to a range of hazard ratios (x axis). Lines indicate medians across simulations with the same exposure–  
655 outcome hazard ratio, shaded areas range from the 2.5<sup>th</sup> to 97.5<sup>th</sup> percentile.

656 **Supplementary Figure 11.** The data correlation structure “confounding and imbalance.” Tumors with more extreme  
657 Gleason scores were set to be more extremely influenced by batch effects in terms of mean and variances.

658 **Supplementary Figure 12.** Plasmode simulation results for all scenarios. Observed hazard ratios after different  
659 approaches to batch effect correction (y axis) are compared to true (fixed) hazard ratios for the biomarker–outcome  
660 association (x axis; solid blue line: no correction for batch effects). Columns are different batch effects that were  
661 added; rows are different data correlation structures.