Batch Effect Correction in a Functional Colorectal Cancer Organoid Clinical Correlation Study

Gavin R. Oliver *, António Miguel de Jesus Domingues, Carlton C. Barnett

Posted Date: 13 February 2026

doi: 10.20944/preprints202602.1067.v1

Keywords: organoids; confounding; technical artifacts; batch-effects; clinical correlation; colorectal cancer; batch correction; experimental design; new approach methodologies

 doi:10.20944/preprints202602.1067.v1

Article

Batch Effect Correction in a Functional Colorectal Cancer Organoid Clinical Correlation Study

Gavin R. Oliver [1,*,+], **António Miguel de Jesus Domingues** [2,+] **and Carlton C. Barnett** [1,3]

[1] Xilis, Inc. 31 Alexandria Way, Durham, NC27712, USA

[2] Xilis, BV, Universiteitsweg 101, 3584 CG, Utrecht, Netherlands

[3] University of Colorado School of Medicine, Aurora CO 80045, USA

* Correspondence: gavin.oliver@xilis.net

[+] Authors contributed equally

Abstract

Batch effects are recognized as major sources of technical confounding in high-throughput assays. However, their impact on organoid studies receives little attention in the literature. As organoids gain prominence as a class of emerging new approach methodologies (NAMs), consideration of batch variation will become increasingly important to ensure data reproducibility and accurate interpretation in pre-clinical and clinical studies. In this manuscript, we provide a practical description of our work in detecting, characterizing, and correcting batch effects in a prior published retrospective clinical colorectal cancer organoid drug-response study. We outline the workflow employed, including exploratory diagnostics, experimental drift detection, and statistical adjustment. We detail methods employed to evaluate batch effects, monitor longitudinal drift, and select approaches to remove technical artifacts, preserve biological signal and test for robustness. Our experience demonstrates that in even modestly sized studies, results can be adversely affected by insufficient consideration and attempts at ameliorating batch effects. By documenting the challenges encountered and the solutions implemented within our study, we hope that we can provide a seminal practical reference for organoid researchers and enable increased discussion and adoption of robust batch-compensation practices in the organoid field, ensuring that the topic is more routinely addressed, improved, and eventually standardized.

Keywords: organoids; confounding; technical artifacts; batch-effects; clinical correlation; colorectal cancer; batch correction; experimental design; new approach methodologies

1. Introduction

Organoids are self-organizing 3D multicellular culture systems arising from stem cells [1], or in the case of patient derived tumor organoids (PDTOs), from dissociated tumor cells [2]. Organoids possess the ability to mimic the genetics, molecular characteristics and physiological behaviors of their tissue of origin and thus promise biomimicry exceeding traditional cell-line or animal models [2]. Organoids and their various applications have seen a surge of interest in recent years. Since Sato *et al.* successfully developed crypt-villus organoids from Lgr5+ intestinal stem cells in 2009 [3], the modern organoid field has undergone explosive growth. Following the crypt-villus organoids, many other tissue types have been successfully developed including pancreas, lung, liver, skin, heart, stomach and colon [4]. Not only have the available tissues expanded, but organoids have experienced ever-broadening applicability. Organoids are routinely used in the fields of regenerative medicine, reproductive biology, infectious, neurodegenerative and genetic disease research, and oncology [5]. Within the field of oncology, their clinical relevance is becoming increasingly prominent [6]. Studies have demonstrated the potential of organoids to significantly correlate with or predict patient treatment response in malignancies including colorectal cancer, gastroesophageal cancer and ovarian cancer, and have reported encouraging correlations in a wide selection of other tumor types [6].

 doi:10.20944/preprints202602.1067.v1

Furthermore, since 2014, there have been a growing number of clinical trials in oncology, utilizing organoids for purposes including the assessment of genetic makeup and drug sensitivity compared to the original tissue, testing of new treatments and establishment of biobanks [2]. Existing large scale organoid biobanks including the Hubrecht Institute and Hub, promise to make organoids more broadly accessible and standardized for commercial and academic initiatives [7]. In the wake of the FDA Modernization Act 2.0 in 2022 [8] and the the movement away from animal-only studies by the NIH [9], as well as the foundation of the NIH Standardized Organoid Modeling (SOM) Center [9] in 2025, the organoid field continues to see widening interest and the investigation and utilization of organoids and related technologies are certain to continue to grow in prominence. While the existing impact and untapped potential of organoid technology are clear, challenges unquestionably remain before that their utility can be fully realized [1]. Tackling the problems of standardization, reproducibility, and reliability will all be core to the successful widespread adoption and realized benefit of organoid technology [10]. Differences in cell procurement, sample storage, media composition and downstream experimental protocols are only a few examples of considerations that have the potential to introduce variability sufficient to render ostensibly congruous experiments incomparable [7]. The introduction of standards, shared protocols and industrial developments including automation and miniaturization will be at the core of addressing such challenges [2,4,6,11,12]. Mature analysis protocols will also be needed to ensure robust data analysis that circumvents analytical pitfalls and produces accurate, biologically meaningful and interpretable outputs from organoid-based experiments and single or multi-site studies.

The phenomenon of batch effects is well known and described in the scientific literature [13]. These effects are broadly defined as variability introduced to scientific measurements by technical factors that are irrelevant to the hypothesized driver of the experimentally assessed signal. This systematic variability can vary in cause and prominence and depending on its underlying cause it may serve to either attenuate the actual signal being measured or alternatively create the illusion of signal correlated with experimental variables of interest where none actually exists [14]. When this occurs, it is said that the signal of interest has been confounded by the unwanted technical factor, or confounder [15]. When the biological variable of interest becomes impossible to differentiate from technical factors due to inadequate experimental design causing them to become inseparably correlated, the experiment can be said to be affected by perfect confounding [16]. While consideration of batch effects in the literature appears heavily skewed toward high-dimensionality data, awareness of batch effects is necessary in any area of science when samples of any nature are being gathered, and measurements of any class are being collected and compared. An historical example that illustrates this point was provided by WJ Youden in 1972 and describes how estimates of Astronomical Unit collected by physicists at different times, in different laboratories, varied both within and across labs and differed from the accepted value [17].

While there has been documented awareness of nuisance factors or batch effects dating from the early 20th century [18], the advent of microarrays and broad-scale multicenter studies in the late 1990s and early 2000s brought about a renaissance in their consideration [19]. During this period multiple high profile scientific studies saw their conclusions thrown into disarray by the batch-aware third-party analysis [20]. This led to retraction or damaged credibility for studies reporting genetic differences in ethnic groupings [21,22] and genetic variants associated with exceptional longevity in humans [23,24] among other examples. The highlighting of these issues on a broad scientific stage had the benefit not only of preventing erroneous scientific conclusions being cemented in common belief, but also of spurring the creation of methodologies, initiatives and standards for their consideration and control [25–27].

In an era of multiomics and single cell-level analyses, the awareness of batch effects in the scientific literature appears higher than ever before. Since 2020 there have been over 150 published manuscripts incorporating the term batch effects in their title, which is more than in the 30 years preceding. Despite this previously unseen prominence of literature considering the dangers of batch effects in modern science, and the parallel increase in organoid literature, the two topics rarely

 doi:10.20944/preprints202602.1067.v1

intersect. To our knowledge there is no published literature that considers batch effects in organoid studies either critically or in depth. There is undoubtedly awareness of the variability inherent to organoid studies. Published works discuss batch-to-batch variability with specific mention of challenges including extracellular culture matrices [4,10,28–31], manual handling [11], media variability [4,31], clonal differentiation [31,32], organoid morphology and gene expression [12], and stability of drug sensitivity [33,34]. While some advice is given toward attempting avoidance of variability upfront, published works discussing analytical methods of control or correction are seemingly absent from modern scientific literature. When analytical methods are discussed they tend to focus upon on-plate microenvironmental conditions most associated with high-throughput screens, while inter-plate variability tends to mainly consider vehicle and kill controls for dose response curve normalization [35], rather than experimental confounding. Manuscripts detailing PDTO-based clinical correlation studies appear to make little or no mention of batch effects or their control. An example of the paucity of attention given to batch-effects is a recent systematic review of clinical correlation in colorectal cancer organoid studies [36]. The review itself made no mention of batch effects, nor did the twenty individual studies described within it mention batch effects as considerations in their experimental design or analysis. The lack of described handling of the issue is perhaps partially due to the fact that the majority of recent published literature on the topic of batch effect handling and correction has been heavily biased toward high-throughput data modalities [13,14,37,38]. However, preparedness for batch effects in organoid studies is vitally important, particularly as the organoids undergo such explosive growth in use and edge ever closer to status as a clinical technology.

We recently described a collaborative study with the German Cancer Research Center (DKFZ) [39], utilizing our micro-fluidics based MOSGen technology [40,41] as the foundation for a retrospective clinical correlation study in colorectal cancer patients treated with neoadjuvant standard-of-care agents. Patient-derived tumor tissue was distributed within 3D hydrogel spheres termed MicroOrganospheres (MOS) with size, shape and consistency under a high degree of control from mature, automated equipment and protocols. The study outcomes demonstrated clinical correlation between patient clinical outcomes and drug-treated MOS models. Despite benefiting from mature automation and state of the art micro-fluidics technology, our study was nonetheless designed from the outset to monitor for batch effects and was able to detect and correct for effects that would otherwise have caused negative impacts on clinical correlation analysis. In this manuscript we detail the analytical steps employed for detection, the nature of the effects observed, and the approaches taken to correct them and ensure robustness of the post-correction results. The relatively small scale of an organoid dataset compared to high-dimensionality omics and multiomics data, and the disparate structure of the data itself meant that deviations from methodologies considered standard in other fields were required. It is our belief that effects similar to or more severe than those we observed are of an even higher likelihood in studies that lack the standardized automated components that we have developed, and we hope that the work we describe will inform others in the field to maximize the integrity of results from their future organoid studies.

2. Materials and Methods

Study Design

The original study is described fully in our recently published manuscript [39] but in summary: Primary or metastatic lesions were collected from colorectal cancer patients that had been treated clinically with neoadjuvant standard-of-care therapies. Organoids were established from patient tumor samples and deposited in MOS droplets following sample dissociation. MOS droplets were deposited at an average density of 40 per well in 384 well plates, drug-treated with a nine-point dose gradient and dose response (negative logarithm of the half maximal inhibitory concentration i.e. pIC50) calculated using fluorescence intensity from brightfield microscopy with EpCAM staining, following 7 days in-assay. EpCAM signal intensity at assay day 0 was used to normalize EpCAM endpoint signals to account for minor differences in starting biomass. Percentiles of response to each

 doi:10.20944/preprints202602.1067.v1

standard-of-care therapy were calculated and utilized in subsequent clinical correlation analyses. Clinical correlation was based on percentile of in-assay MOS model dose response to the patient's standard of care treatment, based on both binary clinical lesion-level clinical response (radiological RECIST-like scoring for metastatic tumors and pathological Dworak score for primary lesions), and disease-free survival (DFS). Endpoint CellTiter-Glo assays were run in parallel to EpCAM imaging as an orthogonal measurement of MOS model viability.

On-Plate Controls

Platemaps were used to standardize plate layout and drug dosing. Edge wells were used as buffer rows and platemaps were designed in order to avoid spatial biases. All doses and conditions were run in quadruplicate. In all cases, dose response curves were normalized to multiple control wells (zero killing) and staurosporine kill wells (full killing).

Dose Response Analysis

Sample quality control and dose response analysis were conducted as described previously [39]**.** In brief, MOS droplets were individually segmented in brightfield microscopy images and EpCAM fluorescence signal quantified per-MOS droplet using computational analysis. Endpoint signal was normalized by the initial timepoint signal to normalize for variations in biomass. Whole-well signal was calculated as the sum of individual MOS droplet signals per-well. Dose response curves were constructed using a 4-parameter log logistic fit with the DRDA R package [66]. The curve was normalized to the median vehicle and kill control well signals. Curves were manually inspected for quality control. Notably assays were treated with multiple drugs in order to provide increased data points for assay monitoring and batch-compensation. Samples treated clinically with FOLFOX also had FOLIRI or FOLFOXIRI included in their assay (50/50 split approximately). Samples treated clinically with FOLFIRI received both FOLFIRI and FOLFOX in their assay. Finally, samples treated clinically with FOLFOXIRI received FOLFOXIRI and FOLFOX in their assay, with one case also receiving FOLFIRI.

Study Monitoring

After laboratory processing of each sample batch, dose response metrics were calculated and plotted in the absence of clinical response information to enable monitoring of undesired trends or drift. The study was performed in 3 temporal phases (referred to as Phase 1-3). Exploratory analysis incorporating clinical response data was conducted at the conclusion of each phase to assist in detection of potential batch effects.

Rolling Mean Analysis

Rolling-mean analysis was performed using the rollapply function from the R zoo package [67]. A window size of 3 was selected to compute right-aligned rolling means, enabling the inclusion of single observations at the beginning of each data series. Single observation inclusion was achieved using the partial = TRUE argument, which enables windows with fewer than three observations to be evaluated.

Linear Regression Analysis

Sequential univariate linear regression models were generated, stratified by drug and phase. Separate models were fit for each candidate confounder. Models were fit using the lm function in the base R stats package. Each model regressed pIC50 on a single independent variable. Coefficient estimate, confidence interval, and p-value were extracted and ranked by p-value to identify the confounders associated most strongly with pIC50.

 doi:10.20944/preprints202602.1067.v1

Tumor-Level Clinical Correlation Analysis

Clinical correlation was performed as described previously [39] but in brief, a percentile-based MOS drug response score was calculated for each treatment category using pIC50 values (pre and post batch compensation). Scores were combined across treatments for inter-group comparison. Binary tumor-level clinical response was determined by combining radiological response for metastatic tumors (≥20% growth = Progression, consistent with RECIST thresholds) with pathological Dworak score (TRG0 = Progression; TRG1–4 = Response). Wilcoxon rank-sum tests were used to compare response groups, and p-values were adjusted for multiple comparisons using the Benjamini–Hochberg method. To account for non-independence of multiple lesions contributed by the same patient, we two-sided p-values from 2,000 patient-level permutations in which responder status labels were randomly reassigned to patients (and their lesions) under the null-hypothesis of no association. The final p-value corresponded to the proportion of permuted statistics as or more extreme than the observed value.

Receiver Operating Curve Analysis

Receiver operating characteristic (ROC) analysis was performed using the pROC package [68] in R to evaluate the ability of pIC50 values to discriminate clinical response. Area under the curve (AUC) was calculated with 95% confidence intervals estimated by bootstrap resampling under a binormal model. Uncertainty around sensitivity and specificity were calculated using a nonparametric, patient-level cluster bootstrap. Sensitivity, specificity, and accuracy were recalculated at a fixed classification threshold for each bootstrap dataset and 95% confidence intervals calculated for 2000 replicates. The classification threshold was derived from ROC analysis where candidate thresholds were ranked by balanced accuracy (average of sensitivity and specificity). In cases where the threshold with the highest accuracy corresponded to extreme imbalance (e.g. near-perfect sensitivity but poor specificity), the next highest accuracy threshold that provided a more balanced trade-off between sensitivity and specificity was selected.

Batch Correction

Batch compensation for pIC50 values was performed for each drug individually using the ComBat algorithm from the sva package in R [48]. The processing batch was provided as the batch label (batch = processing_batch). ComBat was run in mean-only mode (mean.only=TRUE) and responder status was supplied as a biological covariate (mod = model.matrix(~responder_status, data = pIC50)).

e.g.

```
ComBat(dat = data_matrix, batch = batch_factor, mod =
model.matrix(~responder_status, data = pIC50), mean.only = TRUE)
```

Permutation Analysis

Batch labels were randomly shuffled among samples before the full batch correction process was rerun using the newly labeled samples and the Wilcoxon rank sum test was applied using the wilcox.test function in base R, and effect size calculated. This analysis was repeated for 1000 iterations, generating empirical null distributions, and the location of the original batch correction p-value and effect sizes in the null distribution were calculated.

Time-to-Event Analysis

Individual MOS drug response scores were correlated with patient-level DFS using Kaplan–Meier visualizations to explore trends in time to progression. A 50th percentile cutoff of the MOS Response Score was used to split the cohort. The log rank test was run to determine significance of

 doi:10.20944/preprints202602.1067.v1

group differences. Survival curves and log-rank p-values were generated using the survminer::ggsurvplot function [69] in R.

3. Results

3.1. Important Elements of Experimental Design

Prior to describing the purely analytical results of our study, it is necessary to consider the elements of experimental design that are required to minimize experimental confounding, or maximize the chances of compensating for problems if they occur. Besides analytics to monitor or post-correct batch effects, core good practices are vital in experimental design. Post-hoc correction of batch effects is not always possible, nor is it necessarily straightforward. There is no substitute for careful planning and good experimental design. We will not describe all elements of good design in depth since they are covered comprehensively elsewhere [15,19], but we will give brief consideration to a selection of some vital design considerations and how they influenced the analysis and results of our study.

3.1.1. Randomization

Ensuring that samples representing the experimental variables of interest are well randomized across time, instruments, etc. is important to avoid high levels of confounding. For example, clustering most responder samples in one month and most non-responders in the following month introduces the potential of observing differences in experimental measurements that can be mistaken as being due to response status, while they are in fact due to temporal differences that may be linked to various underlying causative phenomena e.g. a change in machine calibration, or media batch. Clinical studies like our own can be inherently problematic, since sample receipt and processing is frequently dictated by a patient's disease progression and clinical scheduling. Despite being unable to formally randomize conditions, we were able to ensure a relatively randomized order of sample processing as shown in Table 1.

Table 1. Clinical and experimental metadata for all samples (n=37) included in clinical correlation analysis.

Sample	Study Phase	Processing Batch	Tumor-level clinical response	Clinical treatment	Media Batch	Localization	Disease free survival (years)
Sample 1	Phase 1	1.1	Progression	FOLFOX	NA	Lymph node	0.76
Sample 2	Phase 1	1.1	Response	FOLFIRI	NA	Liver metastasis	0.58
Sample 3	Phase 1	1.1	Response	FOLFOX	NA	Primary tumor	1.95
Sample 4	Phase 1	1.2	Response	FOLFOX	NA	Liver metastasis	0.2
Sample 5	Phase 1	1.2	Response	FOLFOX	NA	Primary tumor	0.54
Sample 6	Phase 1	1.2	Progression	FOLFOX	NA	Primary tumor	3.41
Sample 7	Phase 1	1.3	Response	FOLFOX	NA	Liver metastasis	0.3
Sample 8	Phase 1	1.3	Response	FOLFOX	NA	Liver metastasis	1.06
Sample 9	Phase 1	1.4	Response	FOLFOX	NA	Liver metastasis	0.3
Sample 10	Phase 1	1.4	Response	FOLFOX	NA	Liver metastasis	1.06
Sample 11	Phase 1	1.4	Response	FOLFOX	NA	Liver metastasis	1.11
Sample 12	Phase 2	2.1	Response	FOLFOX	1	Primary tumor	3.42
Sample 13	Phase 2	2.1	Response	FOLFOX	1	Primary tumor	0.33
Sample 14	Phase 2	2.1	Progression	FOLFOXIRI	1	Primary tumor	0.71
Sample 15	Phase 2	2.1	Response	FOLFOXIRI	1	Primary tumor	0.81
Sample 16	Phase 2	2.1	Response	FOLFOX	1	Liver metastasis	1.68
Sample 17	Phase 2	2.1	Response	FOLFOX	1	Primary tumor	0.25
Sample 18	Phase 2	2.1	Response	FOLFOX	1	Liver metastasis	0.54
Sample 19	Phase 2	2.2	Progression	FOLFOX	2	Liver metastasis	0.33
Sample 20	Phase 2	2.2	Response	FOLFOXIRI	2	Liver metastasis	0.71
Sample 21	Phase 2	2.2	Response	FOLFOXIRI	2	Liver metastasis	0.71

 doi:10.20944/preprints202602.1067.v1

Sample 22	Phase 2	2.2	Response	FOLFOXIRI	2	Liver metastasis	0.81
Sample 23	Phase 2	2.2	Response	FOLFOX	2	Liver metastasis	1.11
Sample 24	Phase 2	2.3	Response	FOLFIRI	3	Liver metastasis	0.04
Sample 25	Phase 2	2.3	Response	FOLFOXIRI	3	Liver metastasis	0.81
Sample 26	Phase 2	2.4	Response	FOLFOX	3	Liver metastasis	2.63
Sample 27	Phase 2	2.4	Response	FOLFOX	3	Liver metastasis	2.63
Sample 28	Phase 3	3.1	Progression	FOLFOX	4	Liver metastasis	0.33
Sample 29	Phase 3	3.1	Response	FOLFIRI	4	Liver metastasis	0.45
Sample 30	Phase 3	3.2	Progression	FOLFOX	5	Liver metastasis	0.33
Sample 31	Phase 3	3.2	Progression	FOLFOX	5	Liver metastasis	0.33
Sample 32	Phase 3	3.3	Response	FOLFOX	5	Lymph node	0.35
Sample 33	Phase 3	3.3	Progression	FOLFOX	5	Liver metastasis	0.33
Sample 34	Phase 3	3.4	Progression	FOLFIRI	6	Primary tumor	0.02
Sample 35	Phase 3	3.4	Progression	FOLFIRI	6	Liver metastasis	0.02
Sample 36	Phase 3	3.5	Response	FOLFOX	6	Liver metastasis	0.35
Sample 37	Phase 3	3.5	Progression	FOLFIRI	6	Primary tumor	0.02

3.1.2. Avoiding Perfect Confounding

This relates to the previous point but deserves specific mention - perfect confounding should be avoided at all costs. In a study like ours where drug response is the variable of interest, it is key to ensure that clinical responders and non-responders are sufficiently randomized across potential confounding conditions to ensure that should batch effects occur, the opportunity exists to attempt disambiguation of a variable of interest and a confounder. In the case of our study, while non-responders were less frequently observed than responders, they were effectively distributed across the timeline of the study to avoid over-clustering that might have been problematic. Nonetheless it was not possible to wholly avoid batches containing only one response group of interest, nor was it possible to perfectly balance the number of responders and non-responders within or between batches (see Table 1).

3.1.3. Standardized Operating Procedures

Ensuring that sample procurement, handling and processing procedures are well documented and standardized, with qualified operators trained in their use is important in reducing technical variability that can affect a study. Within both the clinical setting and in our laboratory, standard procedures were frequently in place and relevant individuals were appropriately trained to follow them. Nonetheless, the study was conducted in a non-operations environment while the laboratory was early in its lifecycle and thus some protocols were in flux, meaning potential for inconsistencies existed.

3.1.4. Automating Processes Where Possible

Automation has the potential to reduce manual error and variability throughout an experiment and can thus attenuate specific batch effects caused by inaccuracies, or inconsistencies introduced by operators or manually operated equipment. Our experimental setup was widely automated. The MOSgen apparatus automated and controlled cell deposition and MOS droplet formation. Furthermore, automated liquid handling was employed for MOS droplet deposition in experimental plates, addition of fluorophores and drug dosing.

3.1.5. Avoiding Procedural Changes During the Study

As much as possible, procedures, methodologies, reagents etc should be kept as constant as possible throughout a study. In the context of our study, formal change control procedures were followed. Material changes to experiments were made as infrequently as possible, were discussed by

 doi:10.20944/preprints202602.1067.v1

a multidisciplinary team prior to implementation, and impact assessments made before changes believed to be acceptable were implemented and documented.

3.1.6. Maintaining Records of Possible Confounders

It is not always possible to be aware of every confounder that might affect an experiment, but prior to a study taking place, potential confounders should be discussed and documented. Furthermore, a record of these should be kept for every sample processed to enable post-hoc analysis of potential confounding features, and correction if possible. In the case of our study, potential confounders including media batch and processing batch, were widely recorded and ultimately empowered downstream analysis when the emergence of batch effects was suspected (see Table 1).

3.1.7. Including Bridging Samples or Technical Replicate Controls

A recommended approach to monitoring and potentially correcting for batch effects is inclusion of a control sample which is run in every experimental batch and whose results are expected to be as close as possible to identical with every run. This can expedite the detection of batch effects since any marked deviation in its results are an obvious sign of an issue, and in some cases the sample itself can be used to calculate a corrective factor that can be used to normalize the other samples in the batch. While our experiments did utilize bridge samples, due to a combination of logistical factors, they were not always available for every batch and drug of interest. Thus, while they could provide us with some indication of batch effects, they were insufficient in isolation to correct those effects.

3.2. Cohort Details and Metadata

Full cohort details are provided in Table 1. A total of 37 neoadjuvant treated colorectal cancer samples from 21 individual patients passed QC and were included in the final clinical correlation analysis. The three study phases comprised a total of thirteen processing batches. Patients were treated clinically with FOLFOX, FOLFIRI or FOLFOXIRI. Samples included primary tumor, liver metastases or lymph node metastases. Primary location was classified as either rectal, or left or right colon. Disease free survival time ranged from 0.02 to 3.42 years. Eleven samples were clinically classified as non-responsive to treatment, while 26 were classified as responsive. Processing batch metadata was available for all samples. Media batch was recorded for study Phases 2-3 only. Drug batch was not recorded since a policy of drug discardment after three freeze-thaw cycles was adopted and this practice was believed to be sufficiently rigorous to avoid drug-related issues within the experiment.

3.3. Longitudinal Trend Analysis

Log logistic dose response curves were generated for all sample-treatment combinations and pIC50 values were calculated as a measure of overall sample sensitivity to a treatment. pIC50s for each individual sample-treatment combination were plotted temporally, post-processing of each batch, to serve as an indicator of potential trends (Figure 1A). A 3-sample rolling mean was plotted to aid with the identification of temporal trends amidst expected natural variability in pIC50 values. Rolling-mean smoothing is widely used in identifying temporal trends and detecting drift across many scientific and applied fields that include finance [42], signal processing [43], climate science [44], epidemiology [45], business forecasting [46], and standard time-series analysis [47]. Although rolling-window methods can be sensitive to irregular time spacing, in this study the elapsed time between runs had no expected biological or technical effect. Therefore, sequence order rather than absolute time was the relevant axis for drift detection.

 doi:10.20944/preprints202602.1067.v1

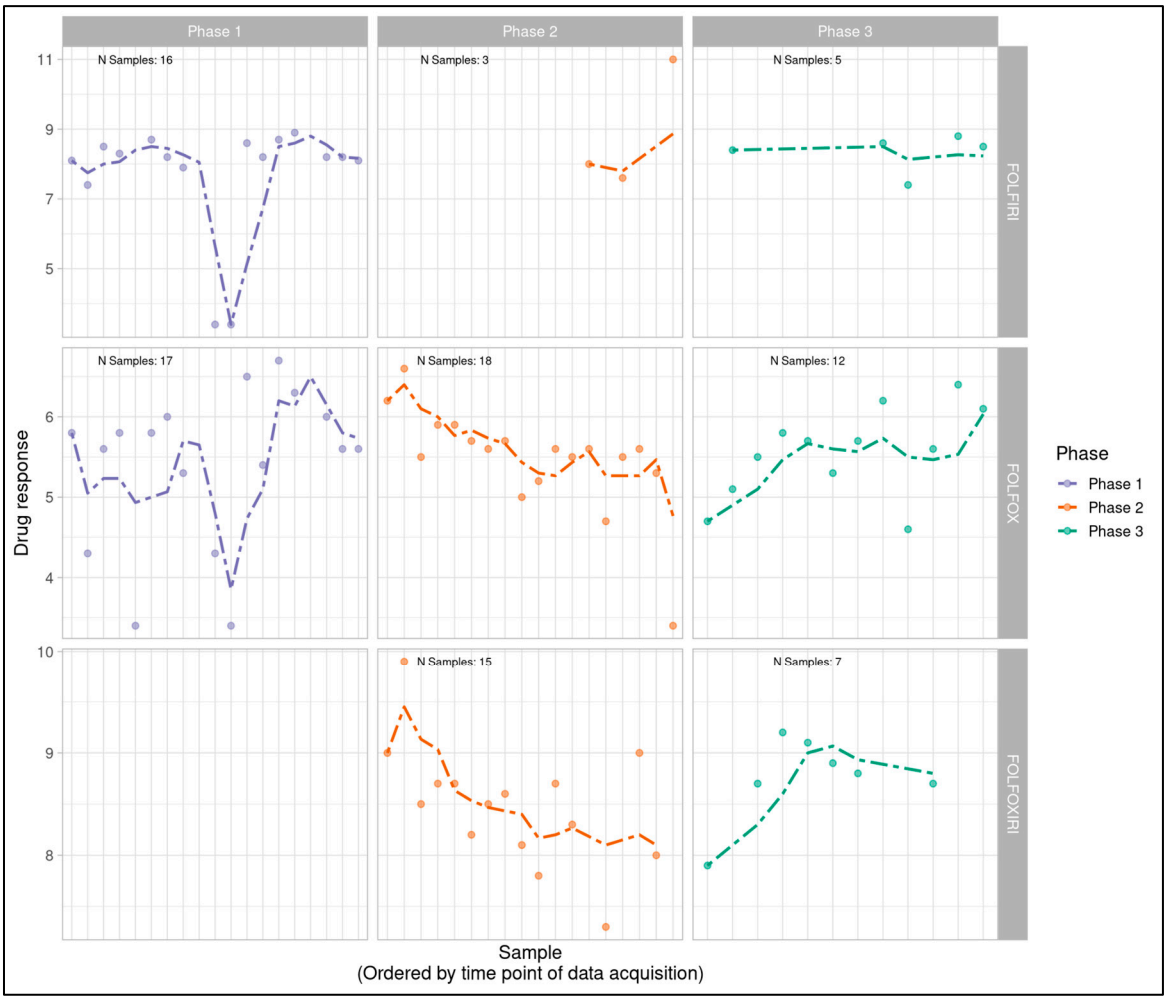


Figure 1A. Longitudinal monitoring of patient-derived MOS model drug response values (pIC50) across all phases of the project timeline with phases (Phase 1, 2 and 3) displayed from left to right and treatment type (FOLFIRI, FOLFOX, FOLFOXIRI) displayed from top to bottom. Dotted lines display a right-aligned 3-sample window rolling mean pIC50 value to account for fluctuations and aid with trend identification. Phase 1 does not appear to show a clear trend while FOLFOX and FOLFIRI show potential downward and upward trends in Phase 2 and Phase 3 respectively.

While Phase 1 data revealed minimal visual trends, Phases 2 and 3 appeared to indicate potential linear trends with FOLFOX and FOLFOXIRI data suggestive of a downward trend in Phase 2 (decreasingly responsive) and an upward trend in Phase 3 (increasingly responsive). FOLFIRI in Phase 2 suggested an upward trend but was heavily influenced by a small number of samples and a single outlier, while FOLFIRI in Phase 3 appeared flat. FOLFOXIRI was not included in Phase 1 of the study.

While this analysis in isolation was insufficient to conclude a problem, it was nonetheless a valuable representation of the project timeline that indicated potential patterns warranting further investigation. Since endpoint CellTiter-Glo assays were run in parallel to EpCAM imaging, we were able to determine that similar longitudinal patterns existed using this orthogonal measurement of viability. This enabled us to conclude that any trend being observed was independent of EpCAM batch, brightfield microscopy conditions, or related variables.

Principal components analysis was attempted due to its traditional use in batch effect exploration, but was unrevealing, likely due to the limited scale and dimensionality of the dataset, and limited batch sizes (data not shown).

 doi:10.20944/preprints202602.1067.v1

3.4. Linear Modeling

Linear modeling was performed on a per-drug, per-phase basis to follow up robustly on visual trend inspection and determine if any recorded variables showed significant correlation with the MOS model drug responses. Variables that were considered included processing batch, media batch, clinical resistance status, and DFS. Phase 1 showed statistically significant correlation with clinical tumor responses for both FOLFOX and FOLFIRI ($p=0.007$ and $p=0.0007$ respectively). Phase 2 demonstrated statistically significant or close to significant correlation with an assortment of processing batches and media batches for FOLFOX, FOLFIRI, and FOLFOXIRI (p-values ranging from 0.011 to 0.103). Phase 3 also primarily showed correlations with media and processing batches for all drugs, although not statistically significant. On the basis of this analysis we concluded that, as suggested by trend visualization, Phase 1 was likely effectively free from batch effects while Phases 2 and 3 were each impacted to varying degrees. Both media batch and processing batch appeared relevant to latter stage batch effects, and since these frequently varied alongside one another and media batch was incompletely captured, processing batch was identified as the likely optimal surrogate metadata field to inform attempted batch compensation.

3.5. Batch Effect Analysis

Batch compensation was subsequently investigated using the ComBat function from the sva package in R [48]. While ComBat's origins lie in microarray studies, it has shown versatility in broader settings and smaller datasets [49–52], and has been recommended for drug response studies [53]. Batch variables suspected to affect experimental measurements are passed to the function using the provided 'batch' argument, with processing batch used in our case. Output is a set of corrected measurements with batch effects removed, where downstream analysis techniques can subsequently be applied to the data. Removal of batch effects and use of surrogate variables has been demonstrated to improve reproducibility, stabilize error rate estimates and reduce dependence [48].

We implemented the ComBat in mean-only mode which applies mean centering across batches without variance/scale adjustment. In this mode, batch effects are modeled as additive offsets and removed through estimation of batch-specific parameters. The mean-only option was chosen because the dataset had low dimensionality, and heterogeneous variance was expected across batches due to variable and unpredictable drug responses [54]. This approach reduces systematic batch effects while preserving within-batch variance that could reflect true biological differences. Responder status was supplied to ComBat as a biological covariate to prevent overcorrection and maintain response-group-related relevant signals, as has been recommended by the original authors and elsewhere [48,55]. ComBat builds a linear model on batch and biological covariates to estimate additive batch offsets. When mean-only mode is selected, the implementation subtracts only those offsets, leaving scale and variance untouched [56].

Post-ComBat batch compensation, dose response data was replotted longitudinally and visualized in the absence of clinical correlation data. The trends suggested by the original rolling mean visualization appeared attenuated by ComBat, while Phase 1 results appeared relatively unaffected (Figure 1B).

 doi:10.20944/preprints202602.1067.v1

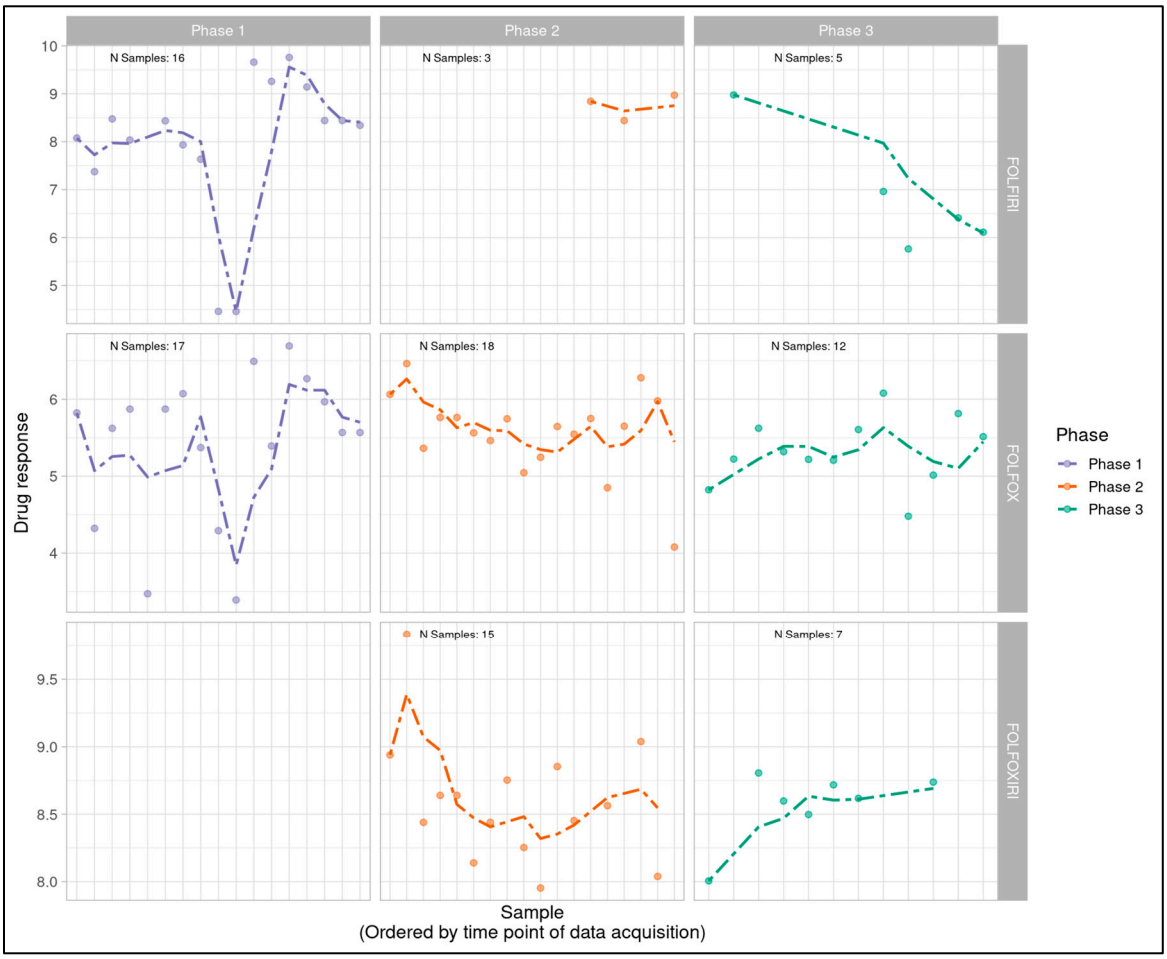


Figure 1B. Longitudinal monitoring of post-ComBat patient-derived MOS model drug response values (pIC50) across all phases of the project timeline with phases (Phase 1, 2 and 3) displayed from left to right and treatment type (FOLFIRI, FOLFOX, FOLFOXIRI) displayed from top to bottom. Dotted lines display a right-aligned 3-sample window rolling mean pIC50 value to account for fluctuations and aid with trend identification. Trends observed with a rolling average prior to ComBat appear visually attenuated after processing by ComBat.

The magnitude and direction of change in pIC50 for each sample post-ComBat was also calculated and visualized per sample (Figure 1C) and per batch (Figure 1D). These results again showed minimal effects on Phase 1 samples, and more pronounced effects in Phases 2 and 3 individually, with some batches showing more modest effects than others. The directions of the observed pIC50 changes per sample were in broad agreement with an expected reversal relative to observed trends in the longitudinal analysis.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 February 2026 doi:10.20944/preprints202602.1067.v1

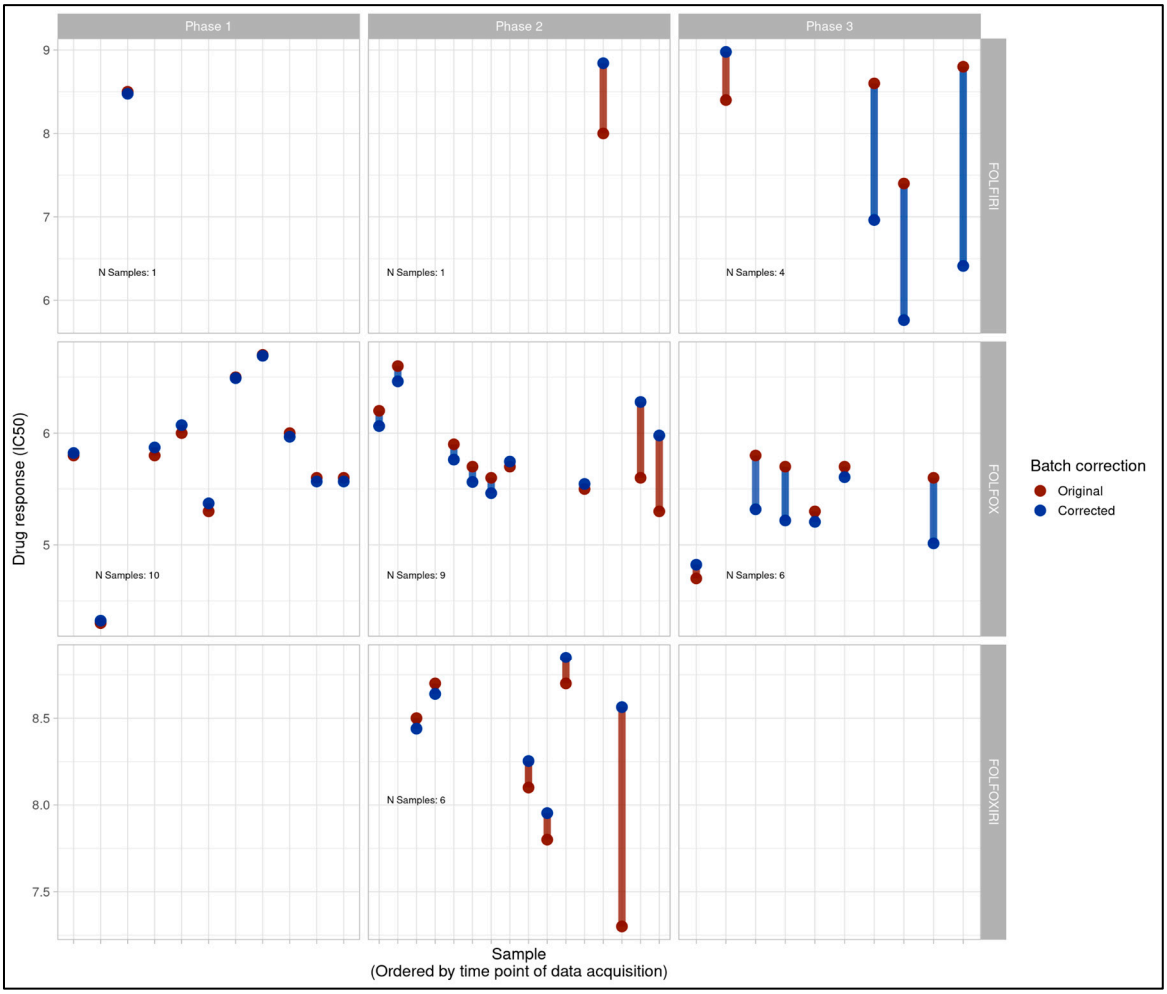


Figure 1C. Per-sample change in pIC50 for all samples entering clinical correlation analysis and their clinical treatment only (n=37) pre- and post-ComBat batch compensation across all project phases with phases (Phase 1, 2 and 3) displayed from left to right and treatment type (FOLFIRI, FOLFOX, FOLFOXIRI) displayed from top to bottom. Phase 1 shows minimal shifts while larger changes are visible in Phases 2 and 3, and directionally correspond largely to what would be expected if attenuating the trends observed in longitudinal analysis.

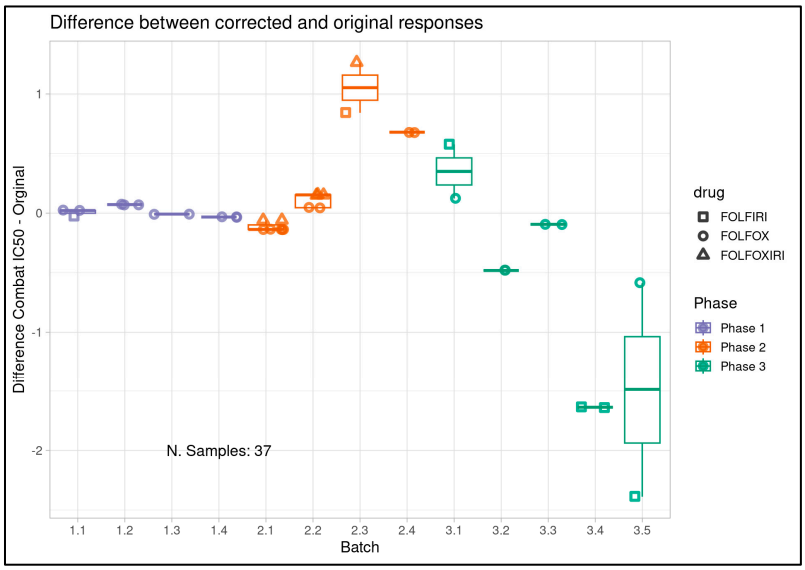


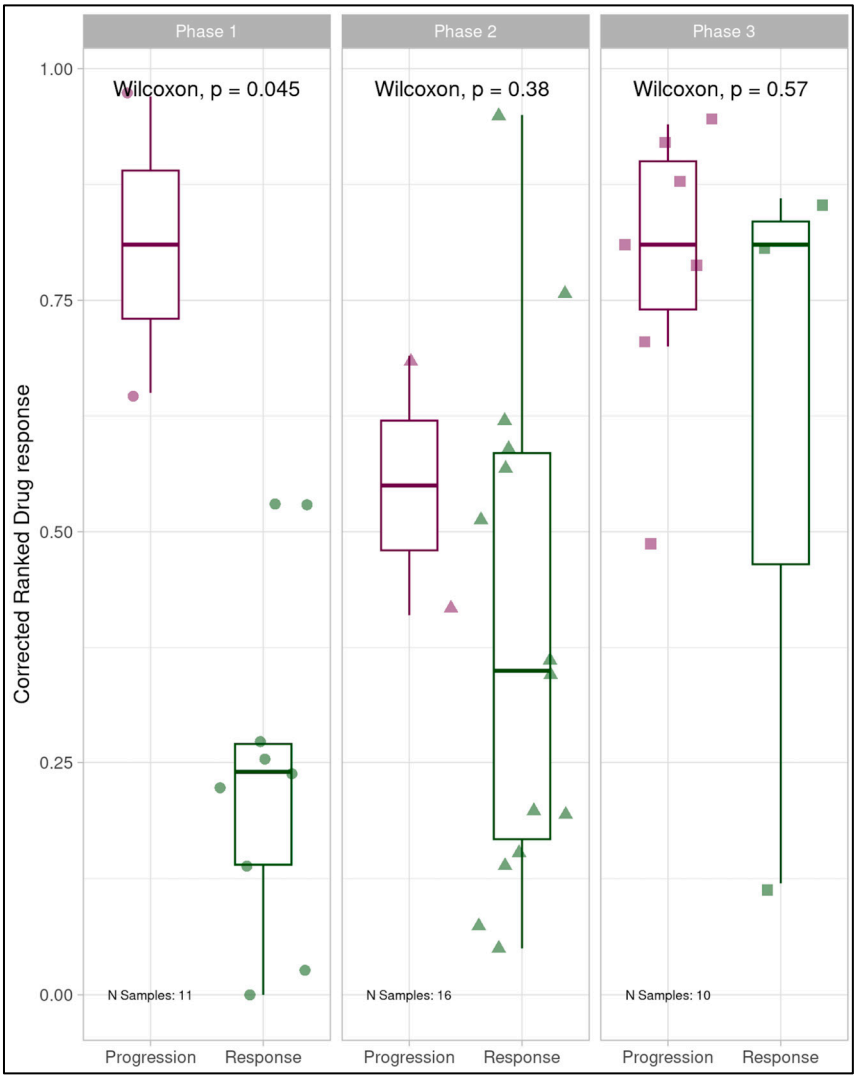
Figure 1D. Per batch delta of pre- and post-ComBat PIC50 values for all samples entering clinical correlation analysis (and their clinical treatment only (n=37). Batches are displayed on the x-axis while pre- and post-ComBat pIC50 delta is displayed on the y-axis. Drug combinations tested in-assay are represented by point shape and

 doi:10.20944/preprints202602.1067.v1

color represents study phase. Correction varies by batch, but larger deltas are observed in Phases 2 and 3 while Phase 1 shows minimal changes.

3.6. Clinical Correlation by Project Phase

Subsequent to the investigative longitudinal analyses, we plotted boxplots displaying correspondence between patient clinical response and ComBat-compensated patient tumor-derived MOS model dose response. MOS model pIC50 values were expressed as per-drug percentiles to facilitate combination of alternative drugs with disparate potencies in a single analysis. Clear discriminative ability was observed across all study phases, both individually (Figure 2) and combined as a single cohort, as demonstrated in our original study (see Figure 3B in Gobits et al.[39]). A cluster-naive Wilcoxon rank sum test was performed for all phases combined and produced a statistically significant p-value of 0.0007. The p-value remained significant following multiple-testing correction using the Benjamini-Hochberg method (p=0.0051). Since the Wilcoxon rank sum test assumes independence of samples and our study included multiple lesions per patient in some cases, we performed a patient-level permutation analysis where response labels underwent permutation for 2000 iterations at the patient level to generate the null distribution. This again yielded a significant p-value (p=0.012). ROC analysis was conducted using the post-ComBat MOS model response values (see Figure 3C in Gobits et al.[39]) and yielded an AUC of 0.86 for the full cohort (95%CI: 0.74 - 0.98). For the cutoff on the ROC curve maximizing balanced sensitivity and specificity, the assay showed 83% accuracy (95%CI: 69%-100%), 82% sensitivity (95%CI: 64%-100%), and 85% specificity (95%CI: 64%-100%).



 doi:10.20944/preprints202602.1067.v1

Figure 2. Boxplots displaying correspondence of patient-derived MOS model dose response to clinical treatment response for each project phase, post batch compensation with ComBat. All phases show separation of responders and non-responders, with patients responsive to clinical treatment having the most drug-responsive MOS models.

Clinical correlation was subsequently assessed utilizing retrospective binary tumor-level patient clinical treatment response data and corresponding ComBat-uncompensated dose response data for drug-treated patient-derived MOS models. Phase 1 demonstrated clean separation of responders and non-responders with more drug-responsive MOS models corresponding to patients' clinical responsiveness to treatment (cluster naive Wilcoxon rank sum p=0.044). This provided evidence of an initially predictive assay even in the absence of batch-compensation (Figure 3) however the apparent predictivity deteriorated in both Phase 2 (p=0.75) and Phase 3 (p=0.65) of the study with responder and non-responder groupings showing high degrees of overlap and data from all three phases combined similarly showing little distinction between responders and non-responders. Receiver operating characteristic (ROC) analysis for the full study data produced an AUC of 0.421 further highlighting the lack of predictivity (Figure 4). The coincidence of unexpected linear trends in longitudinal monitoring for Phase 2 and 3, accompanied by promising clinical discriminative ability in Phase 1 and subsequent degradation of the assay's discriminative ability during Phases 2 and 3 provided further indication of likely technical issues impacting the study, as indicated by the original longitudinal analysis.

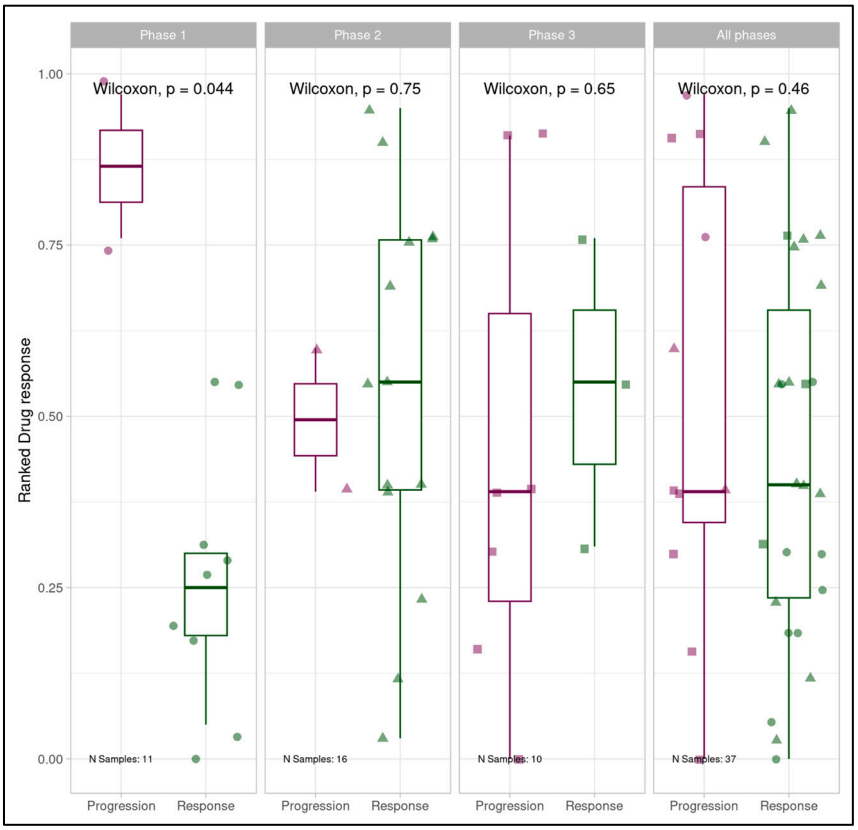


Figure 3. Boxplots displaying correspondence of patient-derived MOS model dose response to clinical treatment response for each project phase individually and in summation, prior to batch consideration. Phase 1 showed clear separation of responders and non-responders, with patients responsive to clinical treatment having the most drug-responsive MOS models. Phase 2 and 3 individually, and all phases combined showed no discriminative ability.

 doi:10.20944/preprints202602.1067.v1

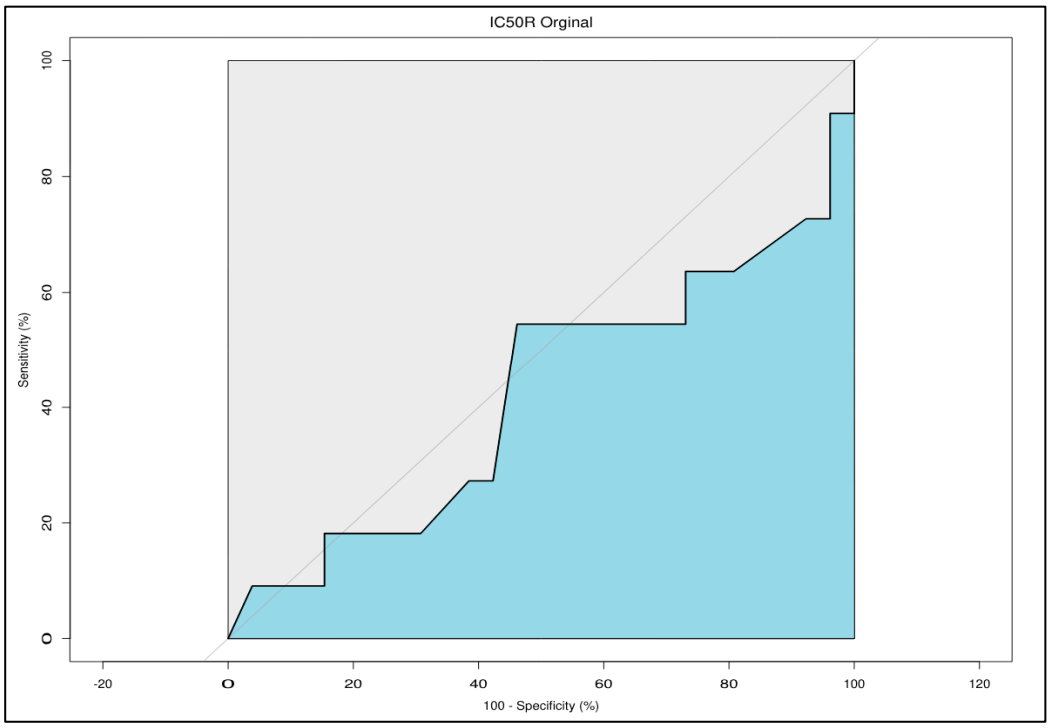


Figure 4. Receiver operating characteristic analysis for all project phases combined showing predictivity of clinical responders (specificity) and non-responders (sensitivity) by patient tumor-derived MOS assay, prior to batch consideration. AUC was 0.421 (95%CI: 0.197-0.645). An AUC of 0.421 reflects predictiveness lower than 0.5 which would be expected by random classification.

3.7. Batch Compensation Robustness Analysis

While post-ComBat outcomes in Phase 2 and 3 matched the clinical discriminative ability observed in pre-ComBat Phase 1 results, and batch compensation appeared to attenuate the longitudinal trends observed across Phases 2 and 3, it remained necessary to confirm the robustness of the post-batch compensation results. Others have questioned the potential for batch effect correction to introduce signals that favor the biological outcome of interest [57]. These concerns have primarily been identified as affecting high-dimensionality multivariate data with variance-based batch correction and unbalanced batches, and are less likely to apply to a univariate, mean-only correction scenario like ours. Nonetheless, we applied permutation analysis to ensure that genuine compensation of batch differences was driving the correction toward a discriminative outcome and that no artifact of the ComBat analysis could be producing the result. In detail, batch labels were randomly shuffled among samples before the full batch correction process was rerun using the newly labeled samples and the Wilcoxon rank sum test was applied, and effect size calculated. This analysis was repeated for 1000 iterations, generating empirical null distributions. Only one simulation attained a p-value as extreme as our ComBat-compensated analysis (Figure 5A). Furthermore, no iteration of the permutation analysis could produce an outcome with an effect size matching or exceeding that of our original ComBat-compensated analysis (Figure 5B). Collectively these findings indicate that correctly labeled batches were the central driver of the batch correction and no spurious behavior of batch compensation was responsible.

 doi:10.20944/preprints202602.1067.v1

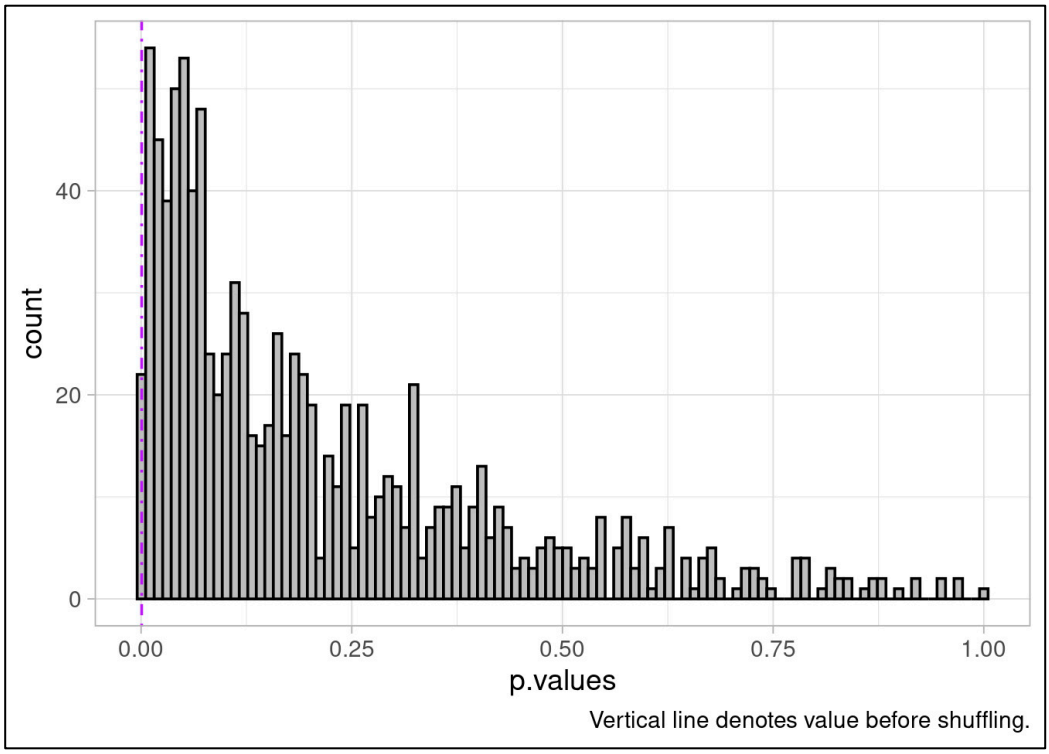


Figure 5A. Null distribution of p-values based on 1000 iterations batch-label permutation followed by ComBat batch correction and Wilcoxon rank sum test. Only one iteration produced a p-value as significant as the outcome of the batch correction applied to the study cohort indicating the reliance of batch correction on our experimental batches for correction resulting in strong discriminative ability. This outcome provides confidence that batch correction behaves as expected and does not spuriously introduce differences between experimental response classes.

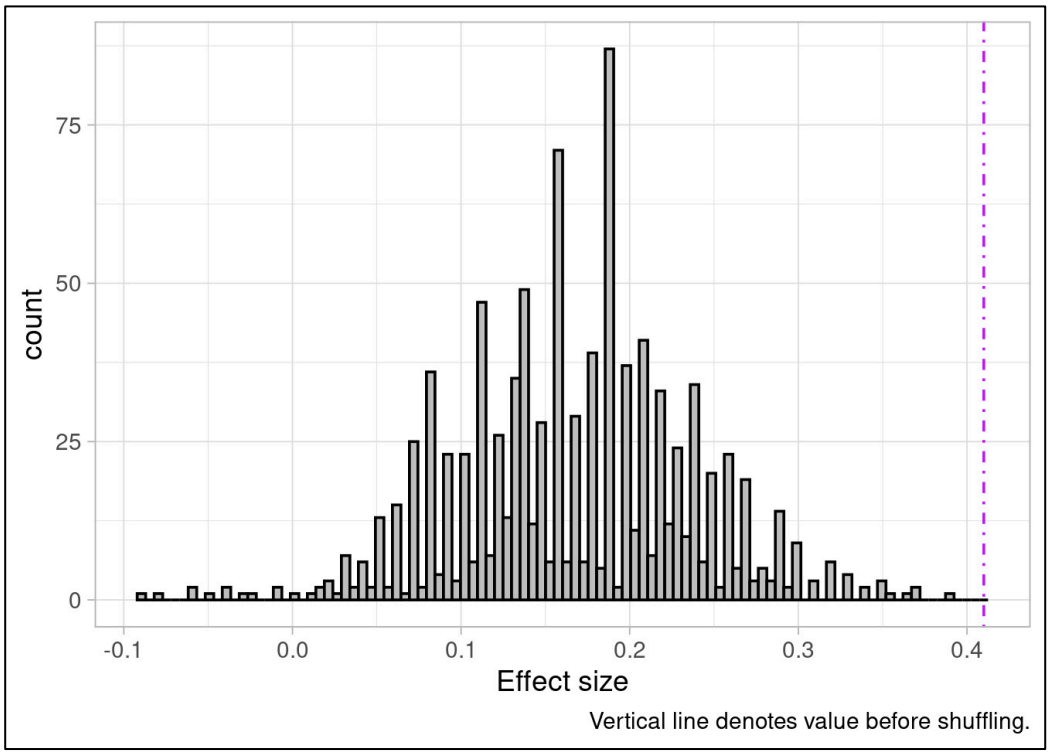


Figure 5B. Null distribution of effect sizes based on 1000 iterations batch-label permutation followed by ComBat batch correction. No iteration produced an effect size equal to the outcome of the batch correction applied to the

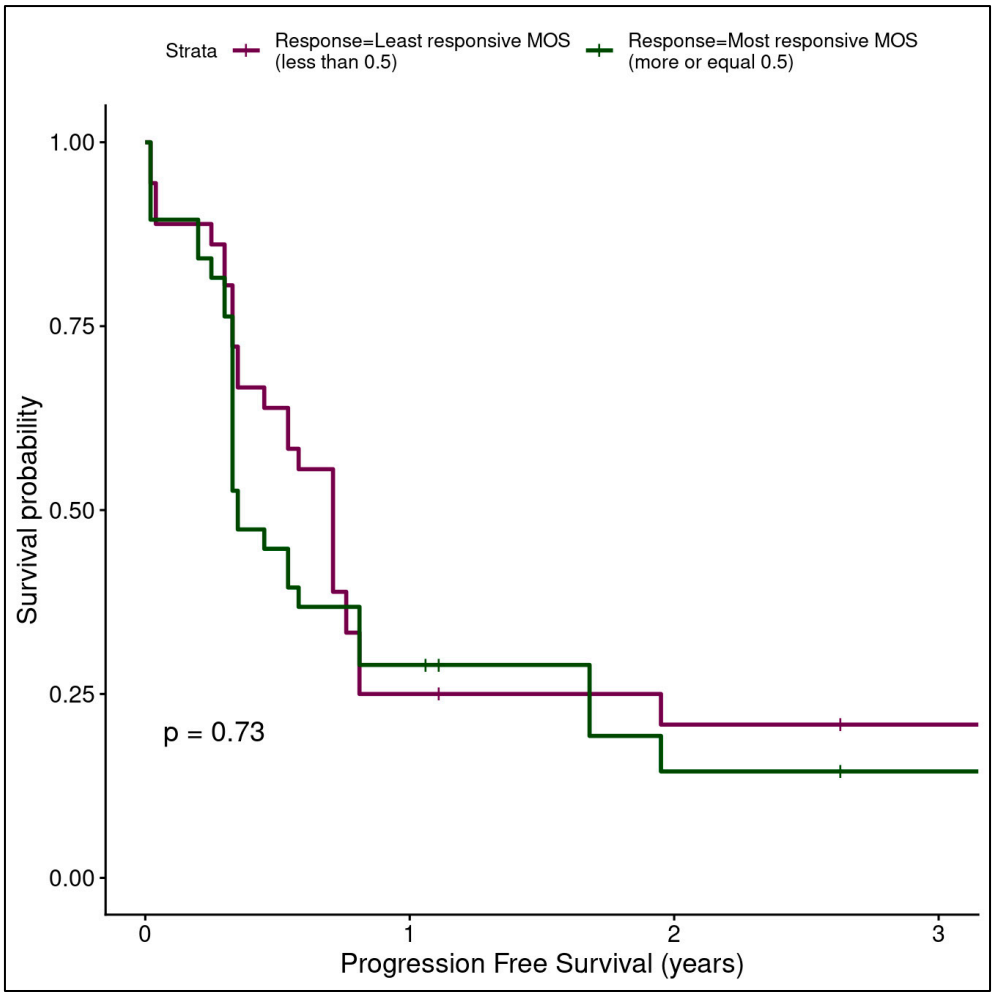
 doi:10.20944/preprints202602.1067.v1

study cohort, indicating the reliance of batch correction on our experimental batches for correction resulting in strong discriminative ability. This outcome provides confidence that batch correction behaves as expected and does not spuriously introduce differences between experimental response classes.

3.8. Disease Free Survival Analysis

Clinical DFS information was available for most patients. While our cohort was modestly sized and relatively underpowered for a time-to-event analysis we generated an exploratory analysis using Kaplan Meier plots to assess the potential for patient-tumor derived MOS to predict longer-term patient outcomes. This analysis was not stratified by study phase and was only conducted for the full study cohort due to the limitations of sample numbers and the high potential for spurious results in underpowered data subsets. Post-batch compensation results (see Figure 3C in Gobits et al.[39]) showed convincing separation (log rank test p=0.18) and higher MOS assay responsiveness appeared to correctly predict longer DFS. While the analysis was exploratory, underpowered for statistical significance, and requires follow-up with an expanded patient cohort, the initial results were an encouraging indication of the potential for our assay to predict time-to-event as well as tumor-level response, and they also provide an orthogonal indication of the ability of batch compensation to correctly attenuate experimental batch effects.

Pre-batch compensated data (Figure 6) showed weak separation (log rank test p=0.73) of the patients with the most responsive (>= 50th percentile responsiveness) and least responsive (< 50th percentile of responsiveness) MOS assay results. Furthermore, the visual separation of the data was weakly suggestive of less responsive MOS assays having higher time to event, which is the opposite of what a predictive assay would be expected to indicate. This provided further indication of the strength, necessity and technical grounding of batch-compensation as a first step in the analytical pipeline.



 doi:10.20944/preprints202602.1067.v1

Figure 6. Kaplan Meier plot for all samples (n=37) using where lines represent patients with the most responsive (>= 50th percentile responsiveness) and least responsive (< 50th percentile of responsiveness) MOS assay pIC50s pre-batch compensation. Separation is degraded without batch compensation (log rank test p=0.73) and less responsive MOS appear to show marginally increased survival probability based on visual inspection alone, in opposition to what would be expected if the assay were discriminative.

4. Discussion

We have described an approach to monitoring, detecting and compensating for batch effects in a real-world organoid-based retrospective clinical correlation study. By conducting longitudinal monitoring of drug sensitivities using a moving average [42–47] to compensate for natural observed variability across our project timeline, and by investigating clinical correlation at the post and pre-batch compensation stage, we were able to identify drift within an initially discriminative assay. Elements of good study design including metadata collection and semi-random order of sample processing subsequently enabled us to investigate and compensate for the issues observed. Application of a traditional method of batch compensation [48] with documented wide-ranging use cases [49–53], parameterized with recorded batch metadata was capable of attenuating observed batch effects and improving discriminative ability in latter study phases, which matched the observations of the initial study phase, prior to manifestation of overt batch effects. The post-batch compensation results were subjected to extended permutation analysis that demonstrated the robustness of the results and reinforced the basis of batch compensation outcomes in the compensatory importance of the experimental batches themselves, and the improbability of undesired behavior in the batch correction methodology [58].

It is clear from the observations we have described that even a carefully conducted study can experience batch-related issues, and diligence is required at all stages of the study, through initial design, execution and analysis. Many good experimental practices as described in our manuscript and in greater detail elsewhere [15,19] were implemented and paramount in compensating for batch effects in our study. Nonetheless, occasional shortcomings in design and implementation present learning opportunities for the future. Examples include the fact that media batch information was not recorded in the initial phase of our study and that drug batch information went uncaptured. A further opportunity for future design robustification will be the comprehensive, study-wide use of technical control organoids, or bridging samples [59]. These were only introduced in the latter phases of our study and could not cover the entire repertoire of agents being tested simultaneously, therefore they provided an incomplete longitudinal control. The information provided by a complete set of control lines would undoubtedly be of high value [60,61] and we encourage others to run these alongside study samples with at least one fully representative complement of drugs and doses per experimental batch.

It is likely relevant that our study, which was performed in a recently established laboratory, unavoidably encountered elements of protocol adjustments, equipment and supply chain events that were naturally in flux, increasing the potential for variability to occur even in the presence of operator diligence. Despite trained staff, a high degree of automation in organoid generation, handling, imaging and drug treatment, and widespread use of standard operating procedures, batch effects nonetheless manifested and demonstrated the capacity to compromise the signal of interest across two of our study phases. Notably, within an operations environment we have processed and treated control organoid lines across many months without notable drift in drug sensitivity being detected (data not shown), demonstrating the robustness of the technology when combined with fully established protocols.

As stated, the issues observed in our study occurred despite good practices, state-of-the art technology, and automation being in-place. Inevitably others engaged in organoid studies will face increased challenges introduced by operating within less controlled, well-equipped or automated environments. Here, the potential for issues to arise will only increase. Furthermore, as industry and academia migrate increasingly toward the use of NAMs [62], organoids will grow in ubiquity, and

 doi:10.20944/preprints202602.1067.v1

the number of studies, publications and confounded results will inevitably expand alongside them. It is vital that investigators recognize the potential for confounded results and adapt suitable best practices within their studies [13,14,19,55,57].

The lack of consideration given to batch control in organoid studies is notable in the face of their recognized potential for variability [4,11,28–30,32] and this challenge is evidenced by the effects we have reported in a single study of modest scale. The literature appears effectively devoid of studies dedicated to characterizing batch effects in organoid studies, or organoid studies that at least methodologically detail if and how they consider the potential for batch effects. As noted by others [57] there appears to be a tendency to condense description of batch effect treatment in published works to the point of uninformativeness, and we believe it is necessary for mandatory minimum documentation frameworks to be formulated and adopted, particularly in light of increasing clinical or clinically adjacent use-cases. The discussion of batches and their treatment within these studies should be brought to the fore as key requirements of study documentation and academic publications. Ultimately recognizing these issues and bringing them into the open will facilitate collaborative efforts, improve understanding and by extension lead to dedicated studies and peer-reviewed or consensus-led optimized methods and protocols for their avoidance or correction. At the very least, describing the methods used to process study data is a long-recognized best practice and if conducted, will provide others with the ability to investigate, repeat and potentially refine such approaches [63].

Regardless of best intentions and adherence to good practices, it is well recognized that a real-world clinical study will present challenges that may be difficult to entirely overcome [19,64]. Even in collaboration with a clinical center of excellence, the order and number in which samples are obtained can be unpredictable. This can lead to the potential for imbalance in study design, unintended temporal clustering of experimental groups of interest, or inconsistencies in batch sizes. Collaboration with external partners also has the potential to introduce a disconnect between clinical and research teams. With limitations in control acknowledged, it is important to ensure that cross-team conversations and planning commence early and continue regularly. The potential for experimental confounding to occur should be discussed and efforts should be made to ensure that all sites are following standardized protocols, recognized metadata points are known and recorded, and that all measures possible are taken to ensure some extent of randomization in the order of sample group processing, and the formation of batches of an acceptable size. Changes to protocols or deviations from standard practice should be communicated and noted centrally alongside sample metadata to ensure its availability at analysis time. Superfluous information can ultimately be filtered or ignored, but missing data can rarely be restored.

Within the course of our study, it was not possible to identify with certainty the cause of the observed batch effects. While batch metadata in combination with appropriate software was capable of attenuating the effects, the batch itself is likely only correlative with the underlying cause. Our observed effects point toward a longitudinal increase or decrease in drug potency or sensitivity depending on the phase and drug, but whether this was caused by differences in the treatments themselves or another factor remains undetermined. While drug batch was not recorded, standard practice dictated that drugs be discarded following three freeze-thaw cycles and follow up diagnostic experiments were unable to produce reduced drug potency in control line organoids following up to ten cycles (data not shown). The inability to fully diagnose a cause versus identify a surrogate in instances of batch correction is a known and accepted limitation [19] of scientific experimentation, since it is generally not possible to consider or record every variable of potential interest. As an example, a study of microarray assays was able to show that something as theoretically benign as atmospheric ozone levels in fact had an insidious ability to affect experimental variability [65]. This goes to further enforce the need to plan and capture information as widely as possible throughout a study timeline, so that causes or surrogates are captured as completely as possible and follow-up investigations and analyses are empowered. While there is always room for improvement, our record keeping was sufficiently comprehensive to enable retrospective investigation of our data, exploration

 doi:10.20944/preprints202602.1067.v1

of univariate linear trends, assessment of a set of possible confounders, and ultimately the parameterization of batch compensation software with the processing batch IDs that enabled attenuation of the batch effects.

We have intentionally described our experience in the form of a narrative, rather than attempting to frame it as a protocol. Others' studies will likely possess their own nuances and require study-specific customization of approaches to deal with batch effects. Furthermore, the underlying cause of an issue may be unique enough to necessitate customized analysis or corrective action. Our approach is not intended to be exhaustive, but we believe it represents an initial blueprint that others might study and optimize for their purposes. We hope that the description of our approach will both raise awareness and bestow others with a solid foundation to base their analytical frameworks upon, particularly when such a dearth of literature exists on the topic. We believe that this discussion can provide a valuable and overdue consideration of the topic and seed further literature discussion that will lead to novel and robust methodologies and packages specific to organoid studies in a similar manner to which these have been developed for other areas of science.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. The original study received approval by the local medical ethics committee (S-136/2021 and S-708/2019).

Informed Consent Statement: Not applicable. For the original study all patients provided written informed consent for data and tissue collection and analysis.

Data Availability Statement: The datasets presented in this article are not readily available because they are part of an ongoing study. Requests to access the datasets should be directed to gavin.oliver@xilis.net.

Acknowledgments: We would like to acknowledge all the individuals involved with the original clinical study conducted prior to this work.

Conflicts of Interest: GRO and CCB are employees of Xilis, Inc. AMD was an employee of Xilis, BV at the time of the study.

Abbreviations

The following abbreviations are used in this manuscript:

NAM	New approach methodology
PDTO	Patient derived tumor organoid
SOM	Standardized organoid modeling
DKFZ	German Cancer Research Center
MOS	MicroOrganosphere
pIC50	Negative logarithm of the half maximal inhibitory concentration
DFS	Disease-free survival
ROC	Receiver operating characteristic
AUC	Area under the curve

References

1. Kim, J.; Koo, B.-K.; Knoblich, J.A. Human Organoids: Model Systems for Human Biology and Medicine. *Nat. Rev. Mol. Cell Biol.* **2020**, *21*, 571–584.

 doi:10.20944/preprints202602.1067.v1

2. Taurin, S.; Alzahrani, R.; Aloraibi, S.; Ashi, L.; Alharmi, R.; Hassani, N. Patient-Derived Tumor Organoids: A Preclinical Platform for Personalized Cancer Therapy. *Transl. Oncol.* **2025**, *51*, 102226.
3. Sato, T.; Vries, R.G.; Snippert, H.J.; van de Wetering, M.; Barker, N.; Stange, D.E.; van Es, J.H.; Abo, A.; Kujala, P.; Peters, P.J.; et al. Single Lgr5 Stem Cells Build Crypt-Villus Structures in Vitro without a Mesenchymal Niche. *Nature* **2009**, *459*, 262–265.
4. Zhao, Z.; Chen, X.; Dowbaj, A.M.; Sljukic, A.; Bratlie, K.; Lin, L.; Fong, E.L.S.; Balachander, G.M.; Chen, Z.; Soragni, A.; et al. Organoids. *Nat. Rev. Methods Primers* **2022**, *2*, doi:10.1038/s43586-022-00174-y.
5. Han, X.; Cai, C.; Deng, W.; Shi, Y.; Li, L.; Wang, C.; Zhang, J.; Rong, M.; Liu, J.; Fang, B.; et al. Landscape of Human Organoids: Ideal Model in Clinics and Research. *Innovation (Camb.)* **2024**, *5*, 100620.
6. Wensink, G.E.; Elias, S.G.; Mullenders, J.; Koopman, M.; Boj, S.F.; Kranenburg, O.W.; Roodhart, J.M.L. Patient-Derived Organoids as a Predictive Biomarker for Treatment Response in Cancer Patients. *NPJ Precis. Oncol.* **2021**, *5*, 30.
7. Tong, L.; Cui, W.; Zhang, B.; Fonseca, P.; Zhao, Q.; Zhang, P.; Xu, B.; Zhang, Q.; Li, Z.; Seashore-Ludlow, B.; et al. Patient-Derived Organoids in Precision Cancer Medicine. *Med (N. Y.)* **2024**, *5*, 1351–1377.
8. United States Congress. (2022). *FDA Modernization Act 2.0*, Pub. L. No. 117-286, 136 Stat. 6103 Available online: https://www.congress.gov/bill/117th-congress/senate-bill/5002/ (accessed on 30 October 2025).
9. National Institutes of Health. (2025, April 29). *NIH to Prioritize Human-Based Research Technologies* (News Release) Available online: https://www.nih.gov/news-events/news-releases/nih-prioritize-human-based-research-technologies (accessed on 30 October 2025).
10. Yang, H.; Li, J.; Wang, Z.; Khutsishvili, D.; Tang, J.; Zhu, Y.; Cai, Y.; Dai, X.; Ma, S. Bridging the Organoid Translational Gap: Integrating Standardization and Micropatterning for Drug Screening in Clinical and Pharmaceutical Medicine. *Life Med.* **2024**, *3*, lnae016.
11. Jiang, S.; Zhao, H.; Zhang, W.; Wang, J.; Liu, Y.; Cao, Y.; Zheng, H.; Hu, Z.; Wang, S.; Zhu, Y.; et al. An Automated Organoid Platform with Inter-Organoid Homogeneity and Inter-Patient Heterogeneity. *Cell Rep. Med.* **2020**, *1*, 100161.
12. Yang, C.; Yang, L.; Feng, Y.; Song, X.; Bai, S.; Zhang, S.; Sun, M. Modeling Methods of Different Tumor Organoids and Their Application in Tumor Drug Resistance Research. *Canc. Drug Resist.* **2025**, *8*, 32.
13. Goh, W.W.B.; Yong, C.H.; Wong, L. Are Batch Effects Still Relevant in the Age of Big Data? *Trends Biotechnol.* **2022**, *40*, 1029–1040.
14. Yu, Y.; Mai, Y.; Zheng, Y.; Shi, L. Assessing and Mitigating Batch Effects in Large-Scale Omics Studies. *Genome Biol.* **2024**, *25*, 254.
15. Wagner, M.R.; Kleiner, M. How Thoughtful Experimental Design Can Empower Biologists in the Omics Era. *Nat. Commun.* **2025**, *16*, 7263.
16. Liljeholm, M. How Multiple Causes Combine: Independence Constraints on Causal Inference. *Front. Psychol.* **2015**, *6*, 1135.
17. Youden, W.J. Enduring Values. *Technometrics* **1972**, *14*, 1.
18. Edwards, A.W.F. RA Fischer, Statistical Methods for Research Workers, (1925). *Landmark writings in western mathematics 1640-1940* **2005**.
19. Leek, J.T.; Scharpf, R.B.; Bravo, H.C.; Simcha, D.; Langmead, B.; Johnson, W.E.; Geman, D.; Baggerly, K.; Irizarry, R.A. Tackling the Widespread and Critical Impact of Batch Effects in High-Throughput Data. *Nat. Rev. Genet.* **2010**, *11*, 733–739.
20. Parker, H.S.; Corrada Bravo, H.; Leek, J.T. Removing Batch Effects for Prediction Problems with Frozen Surrogate Variable Analysis. *PeerJ* **2014**, *2*, e561.
21. Akey, J.M.; Biswas, S.; Leek, J.T.; Storey, J.D. On the Design and Analysis of Gene Expression Studies in Human Populations. *Nat. Genet.* 2007, *39*, 807–808; author reply 808-9.
22. Spielman, R.S.; Bastone, L.A.; Burdick, J.T.; Morley, M.; Ewens, W.J.; Cheung, V.G. Common Genetic Variants Account for Differences in Gene Expression among Ethnic Groups. *Nat. Genet.* **2007**, *39*, 226–231.
23. Alberts, B. Editorial Expression of Concern. *Science* **2010**, *330*, 912.
24. Sebastiani, P.; Solovieff, N.; Puca, A.; Hartley, S.W.; Melista, E.; Andersen, S.; Dworkis, D.A.; Wilk, J.B.; Myers, R.H.; Steinberg, M.H.; et al. Genetic Signatures of Exceptional Longevity in Humans. *Science* **2010**, *2010*, doi:10.1126/science.1190532.

 doi:10.20944/preprints202602.1067.v1

25. Biotechnology, N.; 2006 The MicroArray Quality Control (MAQC) Project Shows Inter-and Intraplatform Reproducibility of Gene Expression Measurements. *News@nat.,Com*.
26. Biotechnology, N.; 2010 The MicroArray Quality Control (MAQC)-II Study of Common Practices for the Development and Validation of Microarray-Based Predictive Models. *News@nat.,Com*.
27. Luo, J.; Schumacher, M.; Scherer, A.; Sanoudou, D.; Megherbi, D.; Davison, T.; Shi, T.; Tong, W.; Shi, L.; Hong, H.; et al. A Comparison of Batch Effect Removal Methods for Enhancement of Prediction Performance Using MAQC-II Microarray Gene Expression Data. *Pharmacogenomics J.* **2010**, *10*, 278–291.
28. Aisenbrey, E.A.; Murphy, W.L. Synthetic Alternatives to Matrigel. *Nat. Rev. Mater.* **2020**, *5*, 539–551.
29. Li, K.; He, Y.; Jin, X.; Jin, K.; Qian, J. Reproducible Extracellular Matrices for Tumor Organoid Culture: Challenges and Opportunities. *J. Transl. Med.* **2025**, *23*, 497.
30. Lumibao, J.C.; Okhovat, S.R.; Peck, K.L.; Lin, X.; Lande, K.; Yomtoubian, S.; Ng, I.; Tiriac, H.; Lowy, A.M.; Zou, J.; et al. The Effect of Extracellular Matrix on the Precision Medicine Utility of Pancreatic Cancer Patient-Derived Organoids. *JCI Insight* **2024**, *9*, doi:10.1172/jci.insight.172419.
31. Driehuis, E.; Kretzschmar, K.; Clevers, H. Establishment of Patient-Derived Cancer Organoids for Drug-Screening Applications. *Nat. Protoc.* **2020**, *15*, 3380–3409.
32. Sandoval, S.O.; Cappuccio, G.; Kruth, K.; Osenberg, S.; Khalil, S.M.; Méndez-Albelo, N.M.; Padmanabhan, K.; Wang, D.; Niciu, M.J.; Bhattacharyya, A.; et al. Rigor and Reproducibility in Human Brain Organoid Research: Where We Are and Where We Need to Go. *Stem Cell Reports* **2024**, *19*, 796–816.
33. Bruun, J.; Kryeziu, K.; Eide, P.W.; Moosavi, S.H.; Eilertsen, I.A.; Langerud, J.; Røsok, B.; Totland, M.Z.; Brunsell, T.H.; Pellinen, T.; et al. Patient-Derived Organoids from Multiple Colorectal Cancer Liver Metastases Reveal Moderate Intra-Patient Pharmacotranscriptomic Heterogeneity. *Clin. Cancer Res.* **2020**, *26*, 4107–4119.
34. Xiang, D.; He, A.; Zhou, R.; Wang, Y.; Xiao, X.; Gong, T.; Kang, W.; Lin, X.; Wang, X.; PDO-based DST Consortium; et al. Building Consensus on the Application of Organoid-Based Drug Sensitivity Testing in Cancer Precision Medicine and Drug Development. *Theranostics* **2024**, *14*, 3300–3316.
35. Tansey, W.; Tosh, C.; Blei, D.M. A Bayesian Model of Dose-Response for Cancer Drug Studies. *arXiv [stat.ML]* 2019.
36. Sakshaug, B.C.; Folkesson, E.; Haukaas, T.H.; Visnes, T.; Flobak, Å. Systematic Review: Predictive Value of Organoids in Colorectal Cancer. *Sci. Rep.* **2023**, *13*, 18124.
37. Han, W.; Li, L. Evaluating and Minimizing Batch Effects in Metabolomics. *Mass Spectrom. Rev.* **2022**, *41*, 421–442.
38. Messner, C.B.; Demichev, V.; Wang, Z.; Hartl, J.; Kustatscher, G.; Mülleder, M.; Ralser, M. Mass Spectrometry-Based High-Throughput Proteomics and Its Role in Biomedical Studies and Systems Biology. *Proteomics* **2023**, *23*, e2200013.
39. Gobits, R.; Schleußner, N.; Oliver, G.R.; Rutenberg Schoenberg, M.; de Jesus Domingues, A.M.; Ramkumar, P.; Suen, S.W.F.; Koomen, M.P.M.; Paolucci, F.; Martens, K.; et al. Functional Precision Medicine Using MicroOrganoSpheres for Treatment Response Prediction in Advanced Colorectal Cancer. *JCO Precis. Oncol.* **2026**, *10*, e2500501.
40. Wang, Z.; Boretto, M.; Millen, R.; Natesh, N.; Reckzeh, E.S.; Hsu, C.; Negrete, M.; Yao, H.; Quayle, W.; Heaton, B.E.; et al. Rapid Tissue Prototyping with Micro-Organospheres. *Stem Cell Reports* **2022**, *17*, 1959–1975.
41. Ding, S.; Hsu, C.; Wang, Z.; Natesh, N.R.; Millen, R.; Negrete, M.; Giroux, N.; Rivera, G.O.; Dohlman, A.; Bose, S.; et al. Patient-Derived Micro-Organospheres Enable Clinical Precision Oncology. *Cell Stem Cell* **2022**, *29*, 905-917.e6.
42. Brock, W.; Lakonishok, J.; LeBARON, B. Simple Technical Trading Rules and the Stochastic Properties of Stock Returns. *J. Finance* **1992**, *47*, 1731–1764.
43. Oppenheim, A.V.; Schafer, R.W. Discrete Time Signal Processing Third Edition. *Pearson Higher Education, Inc.* **2010**.
44. Hansen, J.; Ruedy, R.; Sato, M.; Lo, K. Global Surface Temperature Change. *Reviews of geophysics* **2010**, doi:10.1029/2010RG000345.

 doi:10.20944/preprints202602.1067.v1

45. Cowling, B.J.; Wong, I.O.L.; Ho, L.-M.; Riley, S.; Leung, G.M. Methods for Monitoring Influenza Surveillance Data. *Int. J. Epidemiol.* **2006**, *35*, 1314–1321.
46. Makridakis, S.W.; Wheelwright, S. SC and Hyndman, RJ (1998) Forecasting: Methods and Applications. *John Wiley & Sons Inc. New York*.
47. Panagiotelis, A.; Athanasopoulos, G.; Gamakumara, P.; Hyndman, R.J. Forecast Reconciliation: A Geometric View with New Insights on Bias Correction. *Int. J. Forecast.* **2021**, *37*, 343–359.
48. Leek, J.T.; Johnson, W.E.; Parker, H.S.; Jaffe, A.E.; Storey, J.D. The Sva Package for Removing Batch Effects and Other Unwanted Variation in High-Throughput Experiments. *Bioinformatics* **2012**, *28*, 882–883.
49. Orlhac, F.; Eertink, J.J.; Cottereau, A.-S.; Zijlstra, J.M.; Thieblemont, C.; Meignan, M.; Boellaard, R.; Buvat, I. A Guide to ComBat Harmonization of Imaging Biomarkers in Multicenter Studies. *J. Nucl. Med.* **2022**, *63*, 172–179.
50. Wang, J. ComBat-Met: Adjusting Batch Effects in DNA Methylation Data. *NAR Genomics and Bioinformatics* **2025**, *7*, doi:10.1101/2024.08.13.607838.
51. Jaramillo-Jimenez, A.; Tovar-Rios, D.A.; Mantilla-Ramos, Y.-J.; Ochoa-Gomez, J.-F.; Bonanni, L.; Brønnick, K. ComBat Models for Harmonization of Resting-State EEG Features in Multisite Studies. *Clin. Neurophysiol.* **2024**, *167*, 241–253.
52. Pelletier, S.J.; Leclercq, M.; Roux-Dalvai, F.; de Geus, M.B.; Leslie, S.; Wang, W.; Lam, T.T.; Nairn, A.C.; Arnold, S.E.; Carlyle, B.C.; et al. BERNN: Enhancing Classification of Liquid Chromatography Mass Spectrometry Data with Batch Effect Removal Neural Networks. *Nat. Commun.* **2024**, *15*, 3777.
53. Chen, Y. *DSS-v2.0*; Github;
54. Zhang, Y.; Jenkins, D.F.; Manimaran, S.; Johnson, W.E. Alternative Empirical Bayes Models for Adjusting for Batch Effects in Genomic Studies. *BMC Bioinformatics* **2018**, *19*, doi:10.1186/s12859-018-2263-6.
55. Hui, H.W.H.; Kong, W.; Goh, W.W.B. Thinking Points for Effective Batch Correction on Biomedical Data. *Brief. Bioinform.* **2024**, *25*, bbae515.
56. ComBat: Adjust for Batch Effects Using an Empirical Bayes Framework in Sva: Surrogate Variable Analysis Available online: https://rdrr.io/bioc/sva/man/ComBat.html (accessed on 8 December 2025).
57. Nygaard, V.; Rødland, E.A.; Hovig, E. Methods That Remove Batch Effects While Retaining Group Differences May Lead to Exaggerated Confidence in Downstream Analyses. *Biostatistics* **2016**, *17*, 29–39.
58. Ojala, M.; Garriga, G.C. Permutation Tests for Studying Classifier Performance. *2009 Ninth IEEE International Conference on Data Mining* **2009**, *11*, 908–913.
59. Xia, Q.; Thompson, J.A.; Koestler, D.C. Batch Effect Reduction of Microarray Data with Dependent Samples Using an Empirical Bayes Approach (BRIDGE). *Stat. Appl. Genet. Mol. Biol.* **2021**, *20*, 101–119.
60. Schuyler, R.P.; Jackson, C.; Garcia-Perez, J.E.; Baxter, R.M.; Ogolla, S.; Rochford, R.; Ghosh, D.; Rudra, P.; Hsieh, E.W.Y. Minimizing Batch Effects in Mass Cytometry Data. *Front. Immunol.* **2019**, *10*, 2367.
61. Lawrence, B.E. How to Identify and Prevent Batch Effects in Longitudinal Flow Cytometry Research Studies Available online: https://cytekbio.com/blogs/blog/how-to-identify-and-prevent-batch-effects-in-longitudinal-flow-cytometry-research-studies (accessed on 9 December 2025).
62. Sewell, F.; Alexander-White, C.; Brescia, S.; Currie, R.A.; Roberts, R.; Roper, C.; Vickers, C.; Westmoreland, C.; Kimber, I. New Approach Methodologies (NAMs): Identifying and Overcoming Hurdles to Accelerated Adoption. *Toxicol. Res. (Camb.)* **2024**, *13*, tfae044.
63. Sandve, G.K.; Nekrutenko, A.; Taylor, J.; Hovig, E. Ten Simple Rules for Reproducible Computational Research. *PLoS Comput. Biol.* **2013**, *9*, e1003285.
64. Forshed, J. Experimental Design in Clinical 'omics Biomarker Discovery. *J. Proteome Res.* **2017**, *16*, 3954–3960.
65. Fare, T.L.; Coffey, E.M.; Dai, H.; He, Y.D.; Kessler, D.A.; Kilian, K.A.; Koch, J.E.; LeProust, E.; Marton, M.J.; Meyer, M.R.; et al. Effects of Atmospheric Ozone on Microarray Data Quality. *Anal. Chem.* **2003**, *75*, 4672–4675.
66. Malyutina, A.; Tang, J.; Pessia, A. Drda: An *R* Package for Dose-Response Data Analysis Using Logistic Functions. *J. Stat. Softw.* **2023**, *106*, doi:10.18637/jss.v106.i04.
67. R: Apply Rolling Functions Available online: https://search.r-project.org/CRAN/refmans/zoo/html/rollapply.html (accessed on 5 December 2025).

 doi:10.20944/preprints202602.1067.v1

68. Display and Analyze ROC Curves [R Package PROC Version 1.19.0.1] Available online: https://cran.r-project.org/web/packages/pROC/index.html (accessed on 5 December 2025).
69. Drawing Survival Curves Using "ggplot2" [R Package Survminer Version 0.5.1] Available online: https://cran.r-project.org/web/packages/survminer/index.html (accessed on 5 December 2025).