Cell Systems Transparent Peer Review Record

Comparing the efficacy of cancer therapies between subgroups in basket trials

Adam C. Palmer, Deborah Plana, Peter K. Sorger

Summary

Initial Submission:	Received Jan. 14, 2020 Preprint: https://doi.org/10.1101/401620 *Deposited on bioRxiv, Jan. 14, 2020* Scientific editor: Ernesto Andrianantoandro, Ph.D.
First round of review:	Number of reviewers: Four *Four confidential, zero signed* Revision invited May 22, 2020 *Major changes anticipated* Revision received July 27, 2020
Second round of review:	Number of reviewers: Two *Two original, zero new* *Two confidential, zero signed* Accepted Sept. 12, 2020
Data freely available:	Yes
Code freely available:	Yes

This Transparent Peer Review Record is not systematically proofread, type-set, or edited. Special characters, formatting, and equations may fail to render properly. Standard procedural text within the editor's letters has been deleted for the sake of brevity, but all official correspondence specific to the manuscript has been preserved.

Editorial decision letter with reviewers' comments, first round of review

Dear Dr. Sorger,

I hope this email finds you well. The reviews of your manuscript are back and I've appended them below. On balance, the reviewers appreciate the goals of the work presented here; they've provided constructive comments that are aligned with our hopes for the paper. Accordingly, we're happy to invite a revision. ***We appreciate that the COVID-19 pandemic challenges and limits what you and your lab can do, so to make sure we're absolutely on the same page about the feasibility of revisions, let's schedule a Zoom call at our earliest mutual convenience.***

To help guide this revision, I've highlighted points that seem to warrant special attention. I'd also like to be explicit about an almost philosophical stance that we take at Cell Systems.

I hope you find this feedback helpful. If you have any questions or concerns, I'm always happy to talk, either over email or by phone. More technical information and advice about resubmission can be found below my signature. Please read it carefully, as it can save substantial time and effort later.

I look forward to seeing your revised manuscript.

All the best,

Ernesto Andrianantoandro, Ph.D.
Scientific Editor, Cell Systems

Reviewers' comments:

Reviewer #1: General comments:
This is an interesting study of using alternative statistical methodology to complement the traditional Simon's two stage binominal approach utilising the objective response rates (ORR) in the GO/NO GO decisions in BASKET trials. The authors highlighted an example with SUMMIT study where potential neratinib benefit in HER2 exon 20 mutated non-small cell lung cancer (NSCLC) might have been missed by the binominal approach, especially when there were differential effects with different cohorts of primary tumour origins. Many oncology drugs that eventually improve overall survival in phase III trials had very low ORR. On the other hand, in rare cancers with uncommon molecular subgroups, efficacy benefits could be missed with using ORR alone in basket trials. However, the major limitation of the current approach described by the authors is the lack of real time applicability in the GO/NO GO decision within these trials. Often within these BASKET trials, they include uncommon or rare cancers and with an even rarer incidence with the particular molecular target the trial drug. Therefore different cohorts might be recruiting at different rates and thus the data on progression free survival (PFS) and indeed tumour volume changes might be maturing at different time periods. Therefore this type of complementary analysis could only be carried "post-hoc" when several cohorts have completed recruitment with mature

data. Depending on the early readout from the trial, often based on Simon's two stage binominal approach, sadly specific drug development might already be put on halt and pharmaceutical resources re-directed towards other drug development programmes.

.

Specific Comments:

Introduction

* Page 3 first paragraph

The original BASKET study with vemurafenib should be quoted (Hyman et al N Engl J Med 2015, Subbiah et al Cancer Discov 2020).

* Page 3 first paragraph Please consider amending:

Basket trials are particularly helpful when (i) expanding from an initially successful indication to one or more additional tumor types (ii) searching for a responsive setting in which to perform pivotal trials (iii) studying the predictive value of a biomarker in multiple cancer types and (iv) evaluating rare tumours with even rare molecular subgroups where standalone trials are not feasible to conduct (for example Erdheim-Chester disease/ Langerhans cell histiocytosis with vemurafenib in the BASTKET trial (Diamond et al JAMA Oncol 2018).

* Page 4 first paragraph

It is important to note that individual trial drug might be relevant rather than the concept. For BRAF-mutated colorectal cancer, the ORR was only 4% with vemurafenib + cetuximab in the BASKET trial (Hyman et al N Engl J Med 2015) whereas ORR was 20% with encorafenib + cetuximab in the BEACON trial (Kopetz et al N Engl J Med 2019).

Methods

* Whereas PFS and tumour volume changes might correlate with overall survival in some solid tumours, this has not been demonstrated for all tumour types. This particular limitation should be acknowledged in choosing these endpoints as the primary analysis.

Results

* Page 10/11

In relation to the pembrolizumab basket trial for patients with mismatch repair deficiencies, the authors should validate their findings in the more recently published KEYNOTE 158 study (Marabelle et al J Clin ONcol 2019) for non-colorectal MMR-deficient tumours and KEYNOTE 164 study (Le et al J Clin Oncol 2019) for MMR-deficient colorectal cancers. These studies had much larger sample size in each disease cohorts and indeed were the basis for the FDA approval for pembrolizumab in this indication. In particular MMR-deficient brain, pancreatic and ovarian cancers seemed to have lower ORR compared to the rest and therefore it would be useful for the authors to validate their findings with the permutation testing of PFS and tumour volume change to conclude whether the no difference null hypothesis is still true.

* Page 11 first paragraph

For the pembrolizumab and larotrectinib trials, the authors did not test"no difference in efficacy by class of mutation" null hypothesis. Please indicate the reasons why this selective approach was adopted. I assumed that data are not available in publication according to MLH1, MSH6, MSH2 and PMS2 status for

MMR deficiency or different TRK fusions.

Discussion
* Page 15 second paragraph
As mentioned before, whereas PFS and tumour volume changes might correlate with overall survival in some solid tumours, this has not been demonstrated for all tumour types. This particular limitation should be acknowledged.

Reviewer #2:

The authors proposed an analysis of progression free survival and tumor volume changes across tumor types in basket trials using permutation testing. This analysis can be perform evaluating the effect of the drug in patients according to their tumor type or according to the specific mutation of their tumors.

I have some minor comments and clarification requests:

Comment 1: I would suggest reviewing the sentence "use of master protocol is intended to expedite the development of drugs by reducing the time and number of patients required to find an efficacious therapy for specific subgroup" in page 3. Although is true that master protocols are studies intended to improve drug development, I consider their main aim is to evaluate the activity in of a drug in a small patient population that share a common molecular alteration. For some alterations, it would have been more efficient, in terms of study time, to perform a standard study in a single indication. Performing a master protocol, allow to test the drug in patients who otherwise would never participate in a clinical trial due to the rarity of their disease or the alteration their tumors are harboring.
The referenced article by Hirakawa et al. points more in this direction, from my point of view.

Comment 2: I would suggest the authors to provide more information on whether the method can be limited evaluating PFS if there are big difference in the PFS across tumor types. For example, salivary gland cancers (which are tumors of slow growth) vs NSCLC; or in the future in basket studies evaluating combination of agents (such as immune checkpoint inhibitors) if there are differences in PFS across tumor types with one of the agents, as single agent (for example, an anti-PD1 agent which achieve a PFS three times longer in NSCLC than in HNSCC). If feasible, I would provide some comments in the discussion.

Comment 3: Is there any minimum number of patients with certain mutation to drive conclusions or it would depend on the magnitude if the effect? If feasible, provide some insight in the discussion.

Comment 4: In page 7, use ERRB2 and ERBB3 to define the genes encoding HER2 and HER3 proteins.

Comment 6: review the consistency of the italics in the names of genes across de article.

Comment 5: Page 9 or Supplementary Table 2: the number of each hotspot mutation should be included.

Comment 6: I would recommend adding the number of patients per histology in Table S4.

I consider this article provides a potential answer to an important issue in Basket studies. Permutation analysis may allow to analyze tumor growth changes and PFS in Basket Trials, which are relevant assessment to evaluate potential active drugs in Oncology. This method might provide an additional strategy to extract more information in these small patient populations participating in Basket Trials.

Reviewer #3: In this paper Palmer et collegues evaluate the use of permutation to analyze tumor volume changes and Progression Free Survival across subtypes in basket trials . They give the examples for neratinib, larotrectinib, pembrolizumab, and imatinib, and how permutation testing can actually provide different results from the traditional response rates in a traditional Simon two stage design

They describe in detail current problems we are facing in Basket trials and how difficult it is to evaluate the real value of PFS in non randomized trials with multiple tunor types and how applying permitation could give us another way of evaluating response and capture the real value of treatments.

The paper is well written and I believe it to be of a lot of interest to the field.

Some comments:

Could the authors describe any caveat to use this method?
If permutation tests are well stablished, why have they not used before?-

Reviewer #4: OVERVIEW

Palmer et al. developed a new method for the data analysis of basket trials, in which a single drug is tested simultaneously in multiple tumor (sub)types (defined by tissue of origin or biomarker status). It is likely that the new drug is efficacious only in a subset of these tumor (sub)types, and the goal is to find the promising (sub)types for further investigation. The main challenges include the usually small sample size for each tumor (sub)type.

In my opinion, the manuscript is well written, and the results are interesting. The main methodological innovations include:
(1) The use of tumor volume changes and progression free survival (PFS) data for inference, and
(2) A novel "no difference" test across baskets based on permutation testing.
In contrast, existing methods use objective response rate and binomial testing for inference.
The case studies of four basket trials suggest a potentially overlooked opportunity for the use of neratinib in lung cancer patients, which are interesting and can be useful for drug companies. Because of the use

of different data types and a new hypothesis testing approach, slightly different findings are expected, thus the proposed method can be used as a complement to the current standard.

DATA & CODE AVAILABILTY

I have tried to retrieve the data on the GitHub site provided by the authors. The data are available. One minor comment: the imatinib dataset seems to have only 145 outcomes, which is different from what is mentioned in the manuscript (186 patients).

I have a few major and minor comments, mainly regarding the clarity of the presentation:

MAJOR COMMENTS

(1) On many occasions (e.g., in Figure 3), "type 1 error" and "type 2 error" are mentioned. However, it is not clear what they mean when multiple hypothesis tests are conducted simultaneously in a basket trial. I suggest the authors use more precise terms. For example, in Figure 3 and on page 11, "type 1 error" and "1 - type 2 error" should really be "true positive rate" and "true negative rate", respectively.

(2) Similarly, when mentioning "controlling for multiple hypothesis testing", it is more clear to specify what kind of control it is. For example, when a Benjamini-Hochberg procedure is used, it controls the false discovery rate (instead of family-wise error rate).

(3) In the abstract, the authors mentioned "Permutation testing is a complement ... while controlling for multiple hypothesis testing." Arguably, the Benjamini-Hochberg procedure can also be applied to multiple binomial tests and control for the FDR. It is more precise to separately say "Permutation testing is a complement to binomial ..." and "We apply a Benjamini-Hochberg procedure to control for FDR, which is not done in most existing works."

(4) Simulation is a great way to understand the operating characteristics of the proposed method. The authors mentioned on pages 11-12 "simulated basket trials in which a varying proportion of tumor subtypes responded to therapy", but only one simulation study with 3/10 responsive tumor types was reported. It would be helpful if the authors could report the results for a few more simulation scenarios, perhaps in the supplement. For example, 1/10 and 5/10 responsive tumor types. It may also be helpful to consider one scenario with, say, 3/10 responsive tumor types, in which one tumor type has significantly better response compared to the other two (similar to breast cancer in the SUMMIT trial). These simulation studies could help the reader better understand the strength and (potential) weakness of the proposed method.

MINOR COMMENTS

(1) "Tumor subtype" is a very important term throughout the paper. It may be helpful to define it clearly from the start. It was not immediately clear to me whether it refers to a specific tumor type (based on tissue of origin) or different tumor types with the same genetic alteration, until page 7 (and "subgroup" is

used).

(2) On page 12, line 2, "Supplementary Method S6" is mentioned, which seems to be a typo.

Authors’ response to the reviewers’ first round comments

Attached.

Editorial decision letter with reviewers’ comments, second round of review

Dear Dr. Sorger,

I'm very pleased to let you know that the reviews of your revised manuscript are back, the peer-review process is complete, and only a few minor, editorially-guided changes are needed to move forward towards publication.

In addition to the final comments from the reviewers, I’ve made some suggestions about your manuscript within the “Editorial Notes” section, below. Please consider my editorial suggestions carefully, ask any questions of me that you need, make all warranted changes, and then upload your final files into Editorial Manager.

We hope to receive your files within 5 business days, but we recognize that the COVID-19 pandemic may challenge and limit what you can do. Please email me directly if this timing is a problem or you're facing extenuating circumstances.

I'm looking forward to going through these last steps with you. More technical information can be found below my signature, and please let me know if you have any questions.

All the best,

Ernesto Andrianantoandro, Ph.D.
Scientific Editor, Cell Systems

Editorial Notes

General: Please choose "Methods" as the article type when uploading your revision.

STAR Methods:

Starting in August of 2020, Cell Systems papers will need to contain a comprehensive and structured "Data and Code Availability" statement. These statements will exceed standard STAR Methods requirements, so please note that **the instructions below supersede the instructions found here**.

Data and Code Availability statements pertain to the source data and original code reported in the study. In this context, **source data** is defined as the collection of individual, unprocessed observations used to generate the figures reported in the paper. Examples include scRNA-seq and proteomic datasets, but also CSV spreadsheets used to generate graphs, and original micrographs in TIFF format. **Code** is defined as any computationally implemented program, algorithm, or pipeline necessary to reproduce the analysis or conclusions reported in a paper. Smaller **scripts** that have been used to visualize data and generate figures should also be included in the statement, as described below.

Data and Code Availability statements are reported in the first section of the STAR Methods**. They have four parts and each part must be present. Each part should be listed as a bullet point, as indicated above.**

Part 1 pertains to source data. Examples can be used in any number or combination, making sensible modifications as necessary:

- [Data-type] source data have been deposited at [data-type-specific repository] and are publicly available under the accession numbers: [Insert].
- [Data-type] source data have been deposited at [general repository] and are publicly available at [insert DOI].
- [Data-type] source data are available in the paper's Supplemental Information.
- The [data-type] source data reported in this study have not been deposited in a publicly available repository because [reason why data are not public] . They have been archived locally [insert archiving plan]. To request access [insert instructions].
- This paper analyzes existing, publicly available data. These datasets' accession numbers are provided in the Key Resource Table.
- Source data are not provided in this paper but are available from the Lead Contact on request. *(Note: Cell Systems discourages this practice. If you need to make this statement, please discuss it with your editor first.)*

Part 2 pertains to original code. Examples can be used in any number or combination, making sensible modifications as necessary:

- [Adjective] original code is publicly available at [repository name and DOI].
- [Adjective] original code is available in this paper's Supplemental Information.
- The original code reported in this study is not publicly available repository because [reason why data are not public]. Original code has been archived locally [insert archiving plan]. To request access [insert instructions].
- This paper does not report original code.

Part 3 pertains to scripts used to generate figures. Examples to be used in any number or combination:

- The scripts used to generate the figures reported in this paper are available at [repository name and DOI].
- The scripts used to generate the figures reported in this paper are available in this paper's Supplemental Information.
- The scripts used to generate the figures reported in this paper are available in the [name software package, with version, and provide reference or URL] and their use is described in the STAR Methods.
- Scripts were not used to generate the figures reported in this paper.
- Scripts used to generate the figures presented in this paper are not provided in this paper but are available from the Lead Contact on request. *(Note: Cell Systems discourages this practice. If you need to make this statement, please discuss it with your editor first.)*

Part 4 is a statement: "Any additional information required to reproduce this work is available from the Lead Contact."

In addition,

Please list the basket trial data in the Key Resources Table under "Deposited Data".

Thank you!

Reviewer comments:

Reviewer #3: I believe with authors have answered all questions from the reviewers and the manuscript is very interesting
Thanks

Reviewer #4: I thank the authors for carefully addressing my previous comments. I have no further comments.

>

Response to Review
CELL-SYSTEMS-D-20-00014 ***"Comparing the efficacy of cancer therapies between subgroups in basket trials"***

GENERAL COMMENTS

We thank their reviewers for their comments on the manuscript. We have performed additional simulations and analysis and added the findings either to the text and main figures or to the supplementary figures. We have also made numerous small changes in response to the reviewers' specific comments, all of which were found helpful.

We have edited the manuscript to conform to the STAR methods format, and are submitting an eTOC blurb, Highlights, Key Resources table, and graphical abstract along with our updated text and analysis.

REVIEWERS' COMMENTS:

Reviewer #1: General comments:

This is an interesting study of using alternative statistical methodology to complement the traditional Simon's two stage binominal approach utilising the objective response rates (ORR) in the GO/NO GO decisions in BASKET trials. The authors highlighted an example with SUMMIT study where potential neratinib benefit in HER2 exon 20 mutated non-small cell lung cancer (NSCLC) might have been missed by the binominal approach, especially when there were differential effects with different cohorts of primary tumour origins. Many oncology drugs that eventually improve overall survival in phase III trials had very low ORR. On the other hand, in rare cancers with uncommon molecular subgroups, efficacy benefits could be missed with using ORR alone in basket trials. However, the major limitation of the current approach described by the authors is the lack of real time applicability in the GO/NO GO decision within these trials. Often within these BASKET trials, they include uncommon or rare cancers and with an even rarer incidence with the particular molecular target the trial drug. Therefore different cohorts might be recruiting at different rates and thus the data on progression free survival (PFS) and indeed tumour volume changes might be maturing at different time periods. Therefore this type of complementary analysis could only be carried "post-hoc" when several cohorts have completed recruitment with mature data. Depending on the early readout from the trial, often based on Simon's two stage binominal approach, sadly specific drug development might already be put on halt and pharmaceutical resources re-directed towards other drug development programmes.

We thank the reviewer for these concerns. There is no technical reason preventing our analysis (and permutation testing more generally) from being used in patient recruitment or real-time 'go / no go' decisions used to proceed from the first to second stage in a 2-stage basket trial design.

However, in discussions with the FDA and practitioners we have noticed much higher resistance to changing the way a new trial is conducted than analyzing trials underway or already completed. We therefore present our approach as an analytical framework that is used alongside a traditional 2-stage approach to enrollment (this can involve contemporaneous analysis, not necessarily a retrospective analysis as suggested by the reviewer). Used in this way, the value of our approach lies in defining patient-selection criteria for subsequent trials. This could include making a decision about which cancer subtypes should proceed to a larger Phase 2 or a Phase 3 study.

We believe that rigorously exploring how to implement empirically constructed null distributions to make real-time enrollment decisions in ongoing trials will require additional research. In particular, new simulations make clear that permutation testing and the Simon two-stage design differ with respect to false positive rates and power. Specifically, the first stage of two-stage trials is intentionally permissive, being intended to screen out wholly ineffective treatments, with the consequence of a high false positive rate. At this early stage with few patients, our application of permutation testing yields substantially lower false positive rates but this comes at the cost of lower power. At the end of stage two, permutation testing has superior power and similar false positive rates.

Thresholds can be adjusted of course, to make a permutation-based two-step design more closely mimic current practice. The goal of additional research would be to study how differences in accrual rates across subgroups would impact power and type 1 error. We also propose to study how external, historical control arms could act as null distributions for such analysis. Additionally, two-stage designs should be compared to newer designs such as Bayesian adaptive trials. We hypothesize that use of permutation tests in this setting would improve the ability of early-stage trials to detect therapeutic signals from small numbers of patients as compared to traditional trial statistical methodologies, while being less complex to implement than Bayesian designs.

We have edited the discussion to introduce these ideas and describe how permutation testing could be used in trial design. However, we believe that extensions in our approach to enrollment would best be accomplished by pairing our approach in parallel with an established design such as the Simon 2-Stage in a new basket trial, and then rigorously evaluating the impact with a particular patient cohort.

Specific Comments:

Introduction
* Page 3 first paragraph
The original BASKET study with vemurafenib should be quoted (Hyman et al N Engl J Med 2015, Subbiah et al Cancer Discov 2020).

We apologize for missing this reference. We now include the citation in the Introduction when discussing BRAF inhibitors (Page 4).

* Page 3 first paragraph Please consider amending:
Basket trials are particularly helpful when (i) expanding from an initially successful indication to one or more additional tumor types (ii) searching for a responsive setting in which to perform pivotal trials (iii) studying the predictive value of a biomarker in multiple cancer types and (iv) evaluating rare tumours with even rare molecular subgroups where standalone trials are not feasible to conduct (for example Erdheim-Chester disease/ Langerhans cell histiocytosis with vemurafenib in the BASTKET trial (Diamond et al JAMA Oncol 2018).

This is an excellent point – we thank the reviewer for highlighting the use of basket trials in the setting of rare cancer subtypes. We have included the suggested addition to the text.

* Page 4 first paragraph
It is important to note that individual trial drug might be relevant rather than the concept. For

BRAF-mutated colorectal cancer, the ORR was only 4% with vemurafenib + cetuximab in the BASKET trial (Hyman et al N Engl J Med 2015) whereas ORR was 20% with encorafenib + cetuximab in the BEACON trial (Kopetz et al N Engl J Med 2019).

The reviewer is of course correct: there can be substantial differences in response rates, even in the same cancer subtype, across different types of drugs with nominally the same target. We have modified the text by adding a note of caution about comparing response rates in trials using different BRAF inhibitors.

Methods
* Whereas PFS and tumour volume changes might correlate with overall survival in some solid tumours, this has not been demonstrated for all tumour types. This particular limitation should be acknowledged in choosing these endpoints as the primary analysis.

Correlation between surrogate endpoints and overall survival for a given cancer type is indeed an important consideration in the implementation of our method. We have included a sentence in the Methods section of the manuscript (first paragraph) stating that the decision of whether to use tumor volume or PFS or both in any analysis should be informed by clinical experience in a specific disease and treatment setting. Our approach could also be applied to overall survival data, if it were to be reported in basket trials, although normal complications with OS as an endpoint (e.g. possible switching to salvage therapy, other influences on post-progression survival) would be enhanced in a multi-histology trial and could be expected to have systematic differences by tumor type (e.g. progression of a brain tumor is more likely to be rapidly fatal than progression of a lung tumor).

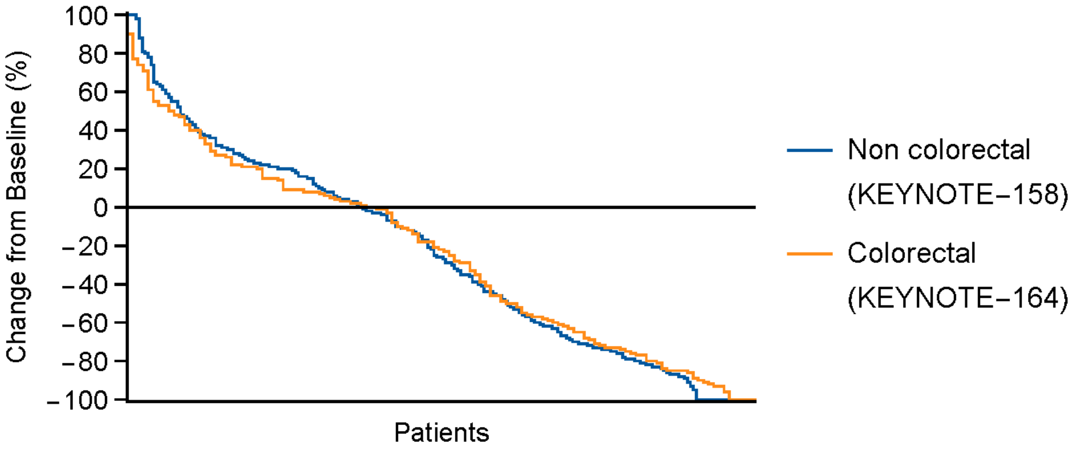
Results
* Page 10/11
In relation to the pembrolizumab basket trial for patients with mismatch repair deficiencies, the authors should validate their findings in the more recently published KEYNOTE 158 study (Marabelle et al J Clin ONcol 2019) for non-colorectal MMR-deficient tumours and KEYNOTE 164 study (Le et al J Clin Oncol 2019) for MMR-deficient colorectal cancers. These studies had much larger sample size in each disease cohorts and indeed were the basis for the FDA approval for pembrolizumab in this indication. In particular MMR-deficient brain, pancreatic and ovarian cancers seemed to have lower ORR compared to the rest and therefore it would be useful for the authors to validate their findings with the permutation testing of PFS and tumour volume change to conclude whether the no difference null hypothesis is still true.

We thank the reviewer for bringing attention to these newer trials with larger cohorts of MMR-deficient cancers. We now include a comparison of colorectal and noncolorectal tumors from KEYNOTE-164 and KEYNOTE-158, described in detail below. Unfortunately the KEYNOTE-158 study of noncolorectal tumors does not distinguish lesion size change data (the ‘waterfall plot’) by specific tumor types. The publication aggregates PFS and tumor volume data across all subtypes. The paper does present summary statistics (ORR) for specific subtypes, in which the only tumor type with ORR significantly different from the whole-population ORR (after adjusting for multiple hypothesis testing) is brain tumors. Other potentially interesting indications of differences (endometrial better; pancreatic and ovarian worse) are precisely what we would like to analyze by permutation tests, but the published summary statistics are insufficient for the application of our method. This is one of many areas in which insufficient data is provided in published papers for a re-analysis of the data; we hope that papers such as ours will encourage trial sponsors to provide additional data, ideally in a digital form, or else apply this analytical method to their own data.

We have contacted the authors of the KEYNOTE 158 study but have not heard back; we would be happy to perform such analysis on this larger cohort if the necessary data is made available, but in the absence of participation from the sponsors we are unable to meet the reviewer's otherwise entirely reasonable request. We have modified our discussion section to note that applying approaches such as ours is contingent upon the availability of patient-level data across tumor subtypes and we appeal to authors and sponsors to make this data available or perform the test themselves.

Even though we are not able to apply our permutation analysis to these trials, as noted, we have been able to use data reported in these papers on tumor volume data to ask the question: are there clear differences in responsive to pembrolizumab between MMR-deficient colorectal and non-colorectal patients?

We find that non-colorectal cancers in the KEYNOTE-158 trial and colorectal cancers in the KEYNOTE-164 trials have similar distributions of tumor volume changes (shown below). This result is consistent with the findings of the formal analysis performed in our manuscript for the 2017 NCT01876511 trial: no subgroup was identified that was significantly less drug-responsive than the average of all tumors. Broadly this supports pembrolizumab's tumor agnostic approval. However, in light of ORR data suggesting possible differences in brain, pancreatic, ovarian, and endometrial cancers, we do think that permutation testing could have value in a post-marketing setting to refine the use of medicines that initially receive a tissue-agnostic approval. We have included this additional evidence as a supplementary figure in our manuscript (Supplementary Figure S1).



* Page 11 first paragraph
For the pembrolizumab and larotrectinib trials, the authors did not test "no difference in efficacy by class of mutation" null hypothesis. Please indicate the reasons why this selective approach was adopted. I assumed that data are not available in publication according to MLH1, MSH6, MSH2 and PMS2 status for MMR deficiency or different TRK fusions.

The reviewer is correct that such mutation-specific data were not available for analysis of the pembrolizumab trial. We added a statement in our Results section to mention that limited data availability limited our ability to perform mutation-specific analysis in this trial. Once again we hope that sponsors or authors will report such data in the future, or will perform their own permutation testing.

In the case of the larotrectinib trial, we have now included an analysis of different NTRK paralogs, and also of different TRK fusion partners (Supplementary Table S2). No significant differences in larotrectinib response are observed by these distinctions. We initially did not perform such analysis because the data seemed ‘intuitively’ to show no differences by TRK fusions, but we are grateful for the suggestion to perform this analysis because it illustrates the applicability of the method in this context. We think this application is valuable because future trials of other targeted therapies may result in activity that is not so consistent by genotype, making this formal method of comparison useful to identify differences where they do exist.

Discussion
* Page 15 second paragraph
As mentioned before, whereas PFS and tumour volume changes might correlate with overall survival in some solid tumours, this has not been demonstrated for all tumour types. This particular limitation should be acknowledged.

The reviewer is correct - the strength of correlation between surrogate endpoints and overall survival for a given cancer type is indeed an important consideration in the implementation of our method. We have edited our Discussion section to highlight this consideration (see also our response above).

Reviewer #2:

The authors proposed an analysis of progression free survival and tumor volume changes across tumor types in basket trials using permutation testing. This analysis can be perform evaluating the effect of the drug in patients according to their tumor type or according to the specific mutation of their tumors.

I have some minor comments and clarification requests:

Comment 1: I would suggest reviewing the sentence "use of master protocol is intended to expedite the development of drugs by reducing the time and number of patients required to find an efficacious therapy for specific subgroup" in page 3. Although is true that master protocols are studies intended to improve drug development, I consider their main aim is to evaluate the activity in of a drug in a small patient population that share a common molecular alteration. For some alterations, it would have been more efficient, in terms of study time, to perform a standard study in a single indication. Performing a master protocol, allow to test the drug in patients who otherwise would never participate in a clinical trial due to the rarity of their disease or the alteration their tumors are harboring. The referenced article by Hirakawa et al. points more in this direction, from my point of view.

We appreciate the reviewer’s thoughtful comment and agree with her/his reading of the referenced article: performing a basket trial is not more efficient as compared to performing a single traditional trial for a common cancer type. We have modified the sentence to conform to the reviewer’s understanding. We still mention that time and resources could be saved as compared to performing multiple traditional trials in parallel, and have added a note about their benefit in rigorously studying drug benefit for rare tumor types.

Comment 2: I would suggest the authors to provide more information on whether the method can be limited evaluating PFS if there are big difference in the PFS across tumor types. For example, salivary gland cancers (which are tumors of slow growth) vs NSCLC; or in the future in basket studies evaluating combination of agents (such as immune checkpoint inhibitors) if there are differences in PFS across tumor types with one of the agents, as single agent (for example, an anti-PD1 agent which achieve a PFS three times longer in NSCLC than in HNSCC). If feasible, I would provide some comments in the discussion.

We broadly agree with the reviewer and have expanded on this point in the discussion. Statistically speaking a permutation test is a valid way to find PFS differences between subgroups, and the subsequent clinical question is whether significant PFS differences relate to drug response or are innate to the tumor type irrespective of therapy. We now point out that the ideal standard of evidence is consistent signals of response according to both tumor shrinkage and PFS, as we have observed for two tumor types treated with neratinib. The revised discussion cautions that a case of longer PFS without shrinkage requires special attention to the natural history of that tumor type. In the given example of salivary gland cancers one would know that a PFS signal without significant tumor shrinkage warrants a skeptical interpretation.

A second consideration is that since ‘progression’ time is typically defined based on a minimum threshold of radiologically detectable growth (often 1.2 or 1.3-fold change on one axis) rather than a longer-term measurement of a more substantial amount of growth, differences in tumor proliferation rate have empirically modest impact on PFS time in patients who are not responding to treatment. For example the salivary gland cancer cohort of KEYNOTE-028 had median PFS of 3.8 months (Cohen *et al*, American Journal of Clinical Oncology: 2018; 41:p1083-1088), which is a relatively small difference from non-responding NSCLC (e.g. median PFS 2.8 months in OAK trial).

On the topic of combination therapies the reviewer is absolutely correct that variation between patients in the magnitude of benefit conferred by one agent can interfere with proper interpretation of the benefit of the second agent. This problem is not specific to basket trials, but affects even classical single-histology trials; addressing this matter in detail is a topic of our ongoing research.

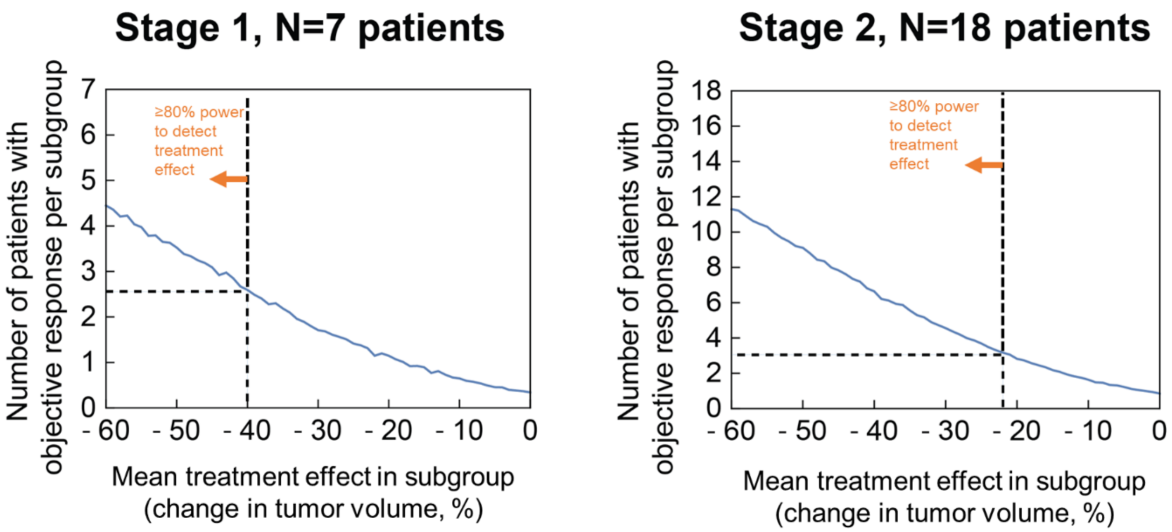
Comment 3: Is there any minimum number of patients with certain mutation to drive conclusions or it would depend on the magnitude if the effect? If feasible, provide some insight in the discussion.

We agree that this is an important question and we have now performed additional analysis to aid in understanding how many patients exceeding a specific response threshold (e.g. more than 30% tumor volume reduction) are required to declare a subgroup as “responsive”. For the simulations shown in Figure 3, we asked: how many patients would exhibit an “objective response” given a specific magnitude of treatment effect, and how does this relates to the ability of the permutation tests to detect such signals?

Traditionally, a power of 0.8 or more (false negative rate ≤ 0.2) is considered a desirable characteristic for clinical trials. Based on our results reported in Figure 5 (formerly Figure 3), we find that 0.8 power is maintained at Stage 1 with an average treatment effect greater than ~ 40%

change in tumor volume with 7 patients enrolled per subgroup and in Stage 2 when average treatment effect is greater than ~ 23% tumor shrinkage with 25 patients enrolled per subgroup. Our newly completed simulations show these effects that can be detected by permutation testing with 80% power correspond to 2-3 responsive patients with ~40% tumor shrinkage in Stage 1 of a basket trial, and ~ 3 patients with 23% tumor shrinkage in Stage 2 of a basket trial.

In summary, simulations show that 3 patient "responders" per subgroup are sufficient to identify responsive tumor subgroups using a permutation testing procedure at 80% power. We have included this additional analysis as a supplementary figure, which is also pasted below (Supplementary Figure S2). Note that we only performed the analysis for 7 Stage 1 and 25 Stage 2 patients to closely match the sample sizes used for permutation analysis in a Simon 2-stage design.



Comment 4: In page 7, use ERRB2 and ERBB3 to define the genes encoding HER2 and HER3 proteins.

We have edited the manuscript to address this comment.

Comment 6: review the consistency of the italics in the names of genes across de article.

We have edited the manuscript to address this comment.

Comment 5: Page 9 or Supplementary Table 2: the number of each hotspot mutation should be included.

We have edited the table to show the sample size for each class of hotspot mutation used in the analysis; note this table has been moved from supplementary material to the main figure.

Comment 6: I would recommend adding the number of patients per histology in Table S4.

We have edited the table to denote the sample size for each histological subgroup used in the analysis; note this table has been moved from supplementary material to the main figure.

I consider this article provides a potential answer to an important issue in Basket studies. Permutation analysis may allow to analyze tumor growth changes and PFS in Basket Trials, which are relevant assessment to evaluate potential active drugs in Oncology. This method might provide an additional strategy to extract more information in these small patient populations participating in Basket Trials.

Reviewer #3: In this paper Palmer et collegues evaluate the use of permutation to analyze tumor volume changes and Progression Free Survival across subtypes in basket trials . They give the examples for neratinib, larotrectinib, pembrolizumab, and imatinib, and how permutation testing can actually provide different results from the traditional response rates in a traditional Simon two stage design

They describe in detail current problems we are facing in Basket trials and how difficult it is to evaluate the real value of PFS in non randomized trials with multiple tunor types and how applying permitation could give us another way of evaluating response and capture the real value of treatments.

The paper is well written and I believe it to be of a lot of interest to the field.

Some comments:

Could the authors describe any caveat to use this method?
If permutation tests are well stablished, why have they not used before?-

We thank the reviewer for these comments. In our revised discussion we have expanded the discussion of limitations and caveats of our approach. One limitation applicable to published trial data is insufficient patient-level, subgroup-specific, data; this would not be an issue for sponsors, but it does limit our ability to test permuation approaches on published data.

Permutation tests are of course well established in statistics at large but we suspect they have not been used in clinical oncology because the binomial methodology was established approximately 40 years ago when computing power was limited. Application of permutation testing to continuous, rather than ordinal data, as is necessary here, requires computing power that was unavailable during the era when much of the biostatistical framework for clinical trials in oncology was established (circa 1960 to 1980). Because trialists and the FDA are appropriately conservative about changes in methodology, relatively recent advances in numerical simulation have not become routine practice. We hope papers such as ours inspire changes in this culture – based on open source software and public domain innovations.

Reviewer #4: OVERVIEW

Palmer et al. developed a new method for the data analysis of basket trials, in which a single drug is tested simultaneously in multiple tumor (sub)types (defined by tissue of origin or biomarker status). It is likely that the new drug is efficacious only in a subset of these tumor (sub)types, and the goal is to find the promising (sub)types for further investigation. The main challenges include the usually small sample size for each tumor (sub)type.

In my opinion, the manuscript is well written, and the results are interesting. The main methodological innovations include:
(1) The use of tumor volume changes and progression free survival (PFS) data for inference, and
(2) A novel "no difference" test across baskets based on permutation testing.
In contrast, existing methods use objective response rate and binomial testing for inference.
The case studies of four basket trials suggest a potentially overlooked opportunity for the use of neratinib in lung cancer patients, which are interesting and can be useful for drug companies. Because of the use of different data types and a new hypothesis testing approach, slightly different findings are expected, thus the proposed method can be used as a complement to the current standard.

DATA & CODE AVAILABILTY

I have tried to retrieve the data on the GitHub site provided by the authors. The data are available. One minor comment: the imatinib dataset seems to have only 145 outcomes, which is different from what is mentioned in the manuscript (186 patients).

Despite the enrollment of 186 total patients in the imatinib basket trial, only 145 outcomes could be extracted and used as part of our analysis. Response data in 41 patients was absent from the original trial publication (Heinrich et al., *Clin. Cancer Res*, 2008 ; Table 1): in 25 patients, response was reported as being unknown / unevaluable, and a further 16 patients belonged to unique indications, that is, the cancer type was represented by a single patient; in such cases the original publication (Table 1) did not report responses. We appreciate the reviewer pointing out this inconsistency and have edited the Results section to clarify the number of responses included in our analysis.

I have a few major and minor comments, mainly regarding the clarity of the presentation:

MAJOR COMMENTS

(1) On many occasions (e.g., in Figure 3), "type 1 error" and "type 2 error" are mentioned. However, it is not clear what they mean when multiple hypothesis tests are conducted simultaneously in a basket trial. I suggest the authors use more precise terms. For example, in Figure 3 and on page 11, "type 1 error" and "1 - type 2 error" should really be "true positive rate" and "true negative rate", respectively.

This is an excellent point! Based on the reviewer’s suggestion, we have edited the Figure 5 (formerly Figure 3) labels to “false positive rate" and "true positive rate" in order to more clearly convey the meaning of our results. We have also edited the corresponding text in the Results section.

(2) Similarly, when mentioning "controlling for multiple hypothesis testing", it is more clear to specify what kind of control it is. For example, when a Benjamini-Hochberg procedure is used, it controls the false discovery rate (instead of family-wise error rate).

We thank the reviewer for this comment and have followed her/his suggestion. We now include specific information on the use of the Benjamini-Hochberg procedure to control for the false discovery rate in the study.

(3) In the abstract, the authors mentioned "Permutation testing is a complement ... while controlling for multiple hypothesis testing." Arguably, the Benjamini-Hochberg procedure can also be applied to multiple binomial tests and control for the FDR. It is more precise to separately say "Permutation testing is a complement to binomial ..." and "We apply a Benjamini-Hochberg procedure to control for FDR, which is not done in most existing works."

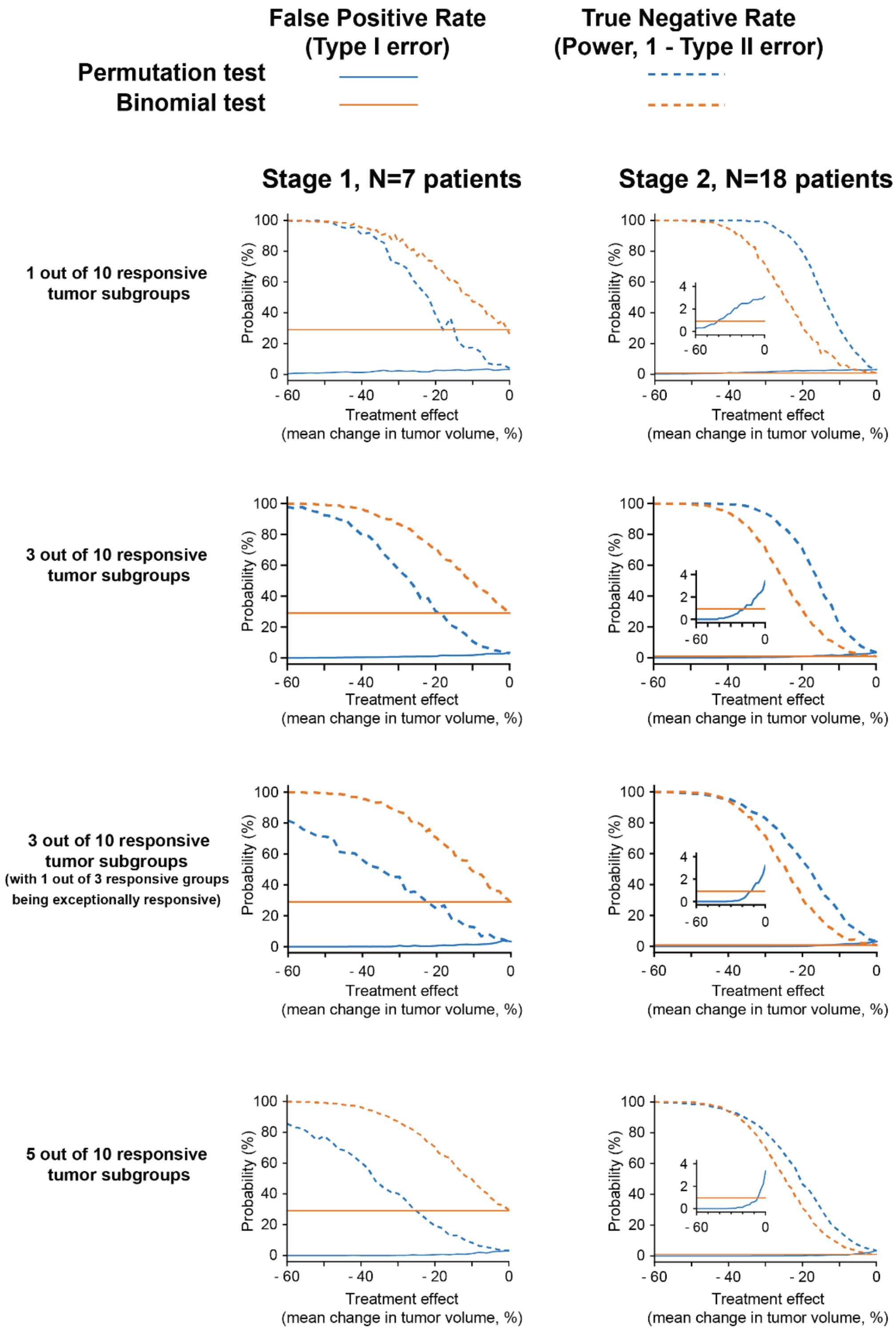
The reviewer is of course correct. We were too casual in our writing and our abstract conflates permutation testing with the Benjamini-Hochberg procedure for multiple hypothesis testing. We have edited the abstract to clarify the distinction between these statistical concepts and their applicability to our analysis.

(4) Simulation is a great way to understand the operating characteristics of the proposed method. The authors mentioned on pages 11-12 "simulated basket trials in which a varying proportion of tumor subtypes responded to therapy", but only one simulation study with 3/10 responsive tumor types was reported. It would be helpful if the authors could report the results for a few more simulation scenarios, perhaps in the supplement. For example, 1/10 and 5/10 responsive tumor types. It may also be helpful to consider one scenario with, say, 3/10 responsive tumor types, in which one tumor type has significantly better response compared to the other two (similar to breast cancer in the SUMMIT trial). These simulation studies could help the reader better understand the strength and (potential) weakness of the proposed method.

We agree with the reviewer and have therefore newly performed the suggested simulations in which False Positive and True Positive rates are calculated for scenarios with 1/10 and 5/10 responsive subgroup. We include the results of this analysis as an updated figure in the main text (Figure 5; formerly Figure 3). In all scenarios, we see similar results: in small cohorts typical of the first stage of a two-stage trial (N=7 patients per tumor type), permutation tests had substantially smaller false positive error rates than binomial tests but also lower true positive rate (less power). In larger cohorts typical of Stage Two (N=25 patients per tumor type), permutation tests have smaller false positive error and greater power for all effect sizes as compared to a binomial tests. These simulations demonstrate that when a greater number of subgroups is responsive to therapy, the permutation test performs more similarly to the Simon's two-stage binomial approach. However, we see no qualitative difference in the results of the simulation based on the number of responsive subgroups used for the analysis.

The new results confirm that our method would be less permissive in Stage 1 as compared to traditional binomial approaches. This is due to the intentionally permissive nature of Stage 1 in the most common implementation of the Simon design: often only 1 response out of 7 patients is required for a subgroup to continue to Stage 2 of a trial. In contrast, permutation based methods have substantially smaller false positive error rates than binomial tests. This is potentially important because a key aim of two stage trial designs is to minimize patients exposed to futile treatments. In larger cohorts typical of Stage Two (N=25 patients per tumor type), permutation tests had greater power for all effect sizes than a binomial test, as well as smaller false positive rate for most treatment effects tested. The lower power of permutation testing can be adjusted, of course, and we suggest in the response to reviewer 1 how this could be achieved.

All three simulations now include data on False Positive and True Positive error rates of tests across a variety of tumor volume changes. Thus, the scenario in which one tumor type has significantly better response as compared to the others is captured by the simulation depicted in Figure 5 (also included below).



MINOR COMMENTS

(1) "Tumor subtype" is a very important term throughout the paper. It may be helpful to define it clearly from the start. It was not immediately clear to me whether it refers to a specific tumor type

(based on tissue of origin) or different tumor types with the same genetic alteration, until page 7 (and "subgroup" is used).

In our manuscript, we aim to use tumor subtype as encompassing classifications using both mutation type and tissue of origin; to our knowledge, this is how it is used in basket trials. In response to the reviewer’s comment, we have added the definition of subtype earlier on in the introduction

(2) On page 12, line 2, "Supplementary Method S6" is mentioned, which seems to be a typo.

We thank the reviewer for bringing this to our attention and have changed the text to reference the Methods section of the manuscript.