bioRxiv preprint doi: https://doi.org/10.1101/401620; this version posted August 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Comparing the efficacy of cancer therapies between subgroups in basket trials

Authors: Adam C. Palmer[*], Deborah Plana*, Peter K. Sorger[†]

*These authors contributed equally

Laboratory of Systems Pharmacology, Department of Systems Biology, Warren Alpert 440, Harvard Medical School, 200 Longwood Avenue, Boston, MA, USA 02115

[†]To whom correspondence should be addressed:
peter_sorger@hms.harvard.edu cc: Christopher_Bird@hms.harvard.edu

The authors declare no potential conflicts of interest.

ORCID ID Numbers

Adam C. Palmer: 0000-0001-5028-7028

Deborah Plana: 0000-0002-4218-1693

Peter K. Sorger: 0000-0002-3364-1838

Keywords: Basket clinical trial, Neratinib, breast cancer, lung cancer, Simon two-stage design

bioRxiv preprint doi: https://doi.org/10.1101/401620; this version posted August 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Statement of Translational Relevance

Basket clinical trials are increasingly common in oncology as a means to identify responsive cohorts in patient populations comprising tumors that arise from many different tissues, particularly when searching for genotypes predictive of outcome. Basket trials typically lack formal control arms and enroll multiple tumor types, each represented by a small number of patients. To overcome the inherent statistical challenges in such patient groups we propose a new biostatistical approach that uses empirical *P* values to test for differences in response across a population. The approach brings added rigor to the interpretation of basket trials, can be applied to both volume changes and PFS duration, and is potentially applicable to any trial in which subdivision of patient populations is desirable as a means to identify outliers or discover and test genetic biomarkers. Our approach is instantiated in open-source code making it simple for others to validate the method and test it on their own data.

bioRxiv preprint doi: https://doi.org/10.1101/401620; this version posted August 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ABSTRACT

Purpose: Basket trials test the activity of drugs in multiple cancer types (tissues of origin or genotypes). Rigorous comparison across types is complicated by small numbers of patients and the lack of a control arm, motivating development of new analytical approaches, particularly as patient stratification becomes more common and refined.

Experimental Design: We reanalyze published basket trials of neratinib in ERBB-mutant cancers and larotrectinib in TRK-fusion cancers, using Monte Carlo permutation tests in which an implicit "*no response to therapy*" null hypothesis is replaced with an empirically derived null of '*no difference in response by tumor type*" (or *class of mutation*). All enrolled patients contribute to null distributions for the analysis of therapy-associated volume change and Progression Free Survival (PFS).

Results: Testing neratinib responses in the SUMMIT trial against a *no difference* null provides insights not obtainable using a conventional dichotomous assessment of volume changes. For example, breast cancers pass the dichotomous standard and exceed the *no difference* test for volume changes but not for PFS. Conversely, lung cancers fail the dichotomous test but exceed *no difference* tests for volume changes and PFS (P= 0.04 and P=0.003) and lung cancers are the sole type for which a specific genotype, ERBB2 Exon 20 mutation (P=0.01), is significantly associated with increased PFS.

Conclusions: Monte Carlo permutation tests enable rigorous determination of tumor types most likely to benefit from therapy in basket trials. Reanalysis of data from SUMMIT identifies an overlooked therapeutic opportunity for neratinib in lung cancers carrying ERBB2 Exon 20 mutations.

bioRxiv preprint doi: https://doi.org/10.1101/401620; this version posted August 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

BACKGROUND

Basket trials in oncology are Phase II clinical trials that explore the efficacy of a drug across multiple types of cancer (typically differing in tissue of origin or genotype) and are informative when: (i) expanding from an initially successful indication to other tumor types (ii) searching for a responsive setting in which to perform pivotal trials (iii) investigating a genomic alteration (e,g oncogenic mutation) in multiple cancer types (1-3). Recent examples of basket trials include the TRK inhibitor larotrectinib in multiple solid tumors with a rearranged *NTRK* gene (4), and immune checkpoint inhibitors for tumors with mismatch repair deficiencies (5). Tissue of origin often exerts a strong influence on response; for example BRAF inhibitors are much less active in BRAF-mutant colorectal carcinomas than melanomas (6). For a single gene, inhibitory, truncating and activating mutations can also have different effects on response (1). Depending on the design, a basket trial can be used to assess the impact of one or more of these variables. For example, the ongoing SUMMIT trial is testing the activity of the ERBB kinase inhibitor neratinib in 21 types of cancer carrying 42 different mutations in the ERBB2 and ERBB3 receptor tyrosine kinases (*HER2* and *HER3*) (7).

In some basket trials activity is detected in almost all cancers, showing that inclusion criteria are effective (e.g. larotrectinib in cancers carrying TRK fusions). In other cases, only a minority of cancers are responsive (e.g. SUMMIT). Current designs for basket trials are based on evaluating drug activity independently for each tumor type and/or analyzing the aggregate of all patients. SUMMIT makes use of the common Simon two-stage optimal design (13), initially enrolling up to seven patients per tumor type in Stage 1, and subsequently expanding enrollment in Stage 2 only if at least one Stage 1 patient exhibits an objective response. The Simon design thereby seeks to detect strong responses while minimizing the number of patients receiving ineffective treatment. Bayesian or frequentist interim analyses have been described that may determine (i) whether aggregate analysis is warranted, and (ii) whether to close enrollment in any subgroups (4,8-12). However, all designs described to date use a fixed threshold to

bioRxiv preprint doi: https://doi.org/10.1101/401620; this version posted August 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

define objective response, most commonly a >30% reduction in tumor volume, assigning a dichotomous response label for each patient.

Almost all basket trials lack a "no drug" control arm involving patients treated with standard of care alone. Instead, the dichotomous threshold is intended to represent a sufficiently high bar that clinically meaningful benefit can be identified (14-16). Hyman *et al* (7) recently reported results for the first 141 SUMMIT patients using a conventional 30% volume reduction. Based on Simon two-stage criteria, breast cancers and some other cancers were selected for Stage 2 of the studies (which are ongoing) but other tumor types were discontinued after Stage 1. In common with other basket trial reports (12), Hyman *et al* (7) commented on but did not formally analyze PFS data or differences in efficacy between subgroups; this reflects the perceived challenge of evaluating 21 tumor types using data from only 141 patients.

Here we propose a complementary approach to analyzing basket trials in which a test against an implicit null hypothesis of '*no efficacy*' (for a Simon two-stage design this is an *ad hoc* threshold) is replaced by a statistically rigorous comparison of tumor types. The analysis is performed using one of two related null hypotheses: '*no difference in efficacy by tumor type*' or '*no difference in efficacy by class of mutation'*. For a given subgroup of patients, defined by tumor tissue of origin or class of mutation, null distributions with an appropriate number of patients are generated to test the alternative hypothesis that response is better than, or worse than, the aggregate of all patients. These formulations of the null hypothesis have the substantial benefit that all patients enrolled in a trial can contribute to the null distribution, and that continuous response variables rather than dichotomous outcomes (e.g. objective response), can be evaluated, including duration of PFS and magnitude of change in tumor volume. When response rates are low as in SUMMIT, the '*no effect*' and '*no difference*' hypotheses are similar; when response rates are high as with larotrectinib, the '*no difference*' hypothesis may identify inferior responses.

bioRxiv preprint doi: https://doi.org/10.1101/401620; this version posted August 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We show here that re-analysis of SUMMIT data against appropriately sampled *no difference* null distributions identifies some tumor types for which conclusions are congruent with criteria in the Simon two-stage design, and others with substantial differences. For example, lung cancers, which fail Simon criteria, significantly exceed the *no-difference* null with respect to volume changes and PFS. In contrast, breast cancers, which greatly exceed Simon criteria based on volume changes, are *no different* from average with respect to PFS.

METHODS

To test the null hypothesis that patient subgroups are equally responsive to neratinib, outcome data from the SUMMIT trial (comprising either change in tumor volume, or duration of PFS) (7) were pooled for all patients who received the drug, regardless of tumor type. We then used a Monte Carlo permutation test (a non-parametric form of bootstrapping) to derive a null distribution for each subgroup of interest. This involved randomly drawing from the pool of all responses, with the number of samples drawn equal to the number of patients found in the cohort being tested (e.g. 26 patients for lung and 5 patients for cervical cancer). A response metric for the sampled set was then calculated and the procedure repeated 10^6 times to compose a null distribution for that cohort. For the analysis of tumor volume changes, the response metric was the average tumor volume change for the cohort, and for the analysis of PFS, the response metric was the hazard ratio (based on the Cox proportional hazards model) of the Kaplan-Meier survival function for a cohort in comparison to that of all patients. An empiric *P* value was then determined by the location of the *observed* response metric, which was the test statistic, on that null distribution. In common with an exact permutation test, the rate of type I error is the significance level, provided that the null distribution is composed from a number of permutations $N >> 1/P$; for $N = 10^6$ and P values $\geq 10^{-3}$ the rate of type I error is accurately determined. The Benjamini-

bioRxiv preprint doi: https://doi.org/10.1101/401620; this version posted August 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Hochberg procedure (17) was used to control the false discovery rate associated with multiple hypothesis testing (FDR = 25%).

In the case of the SUMMIT trial this approach was separately applied to tumor volume changes and to durations of PFS; in the case of the larotrectinib trial (4) it was applied only to tumor volume changes (PFS outcomes by tumor type are not available). For SUMMIT, we also used volume data for all non-breast cancers (n=116) to construct alternative null distributions designed to identify next-best responses (as discussed in a supplementary statistical note; Supplementary Note S8). Because the typical response to neratinib over all tumors was poor (median volume change ≈ 0%; median PFS ≈ 2 months; objective response rate 12%), responses in any one tumor type could not realistically be inferior to the average, and thus we tested only for superiority of each tumor type or mutation class relative to all types. In the case of larotrectinib, the overall response rate was high, and we tested both superiority and inferiority hypotheses.

RESULTS

When neratinib-treated patients in the SUMMIT trial (7) were classified by tissue of origin (Figure 1) and compared to an appropriately resampled "*no difference*" null distribution, breast cancers exhibited significantly greater volume changes than any other tumor type ($p<10^{-6}$; a 45% difference in average volume change from all non-breast tumors). This agrees with the conclusion in Hyman *et al* (7) that breast cancers are the most neratinib-responsive of all tumor types tested. Because this response dominates volume-change data (Supplementary Note S8), we constructed a second set of null distributions for volume changes that included only non-breast tumors (hereafter NB).

When NB distributions were resampled and compared to volume change data for individual tumors, lung, cervical, and biliary cancers were found to significantly exceed the "*no difference by type*" null hypothesis (P=0.04, 0.04 and 0.06; Supplementary Table S1). Whereas cervical and biliary cancers

bioRxiv preprint doi: https://doi.org/10.1101/401620; this version posted August 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

passed the criteria for the first stage of a Simon two-stage design, lung cancer failed at the second stage (Table 1). Thus, we judge as potentially positive a volume change in lung cancers found to be negative by the binary criteria used in a two-stage design. This discordance likely arises because half of lung cancers shrank on therapy but only one exceeded the 30% threshold. The permutation test and Simon criteria therefore provide different insights into the drug responsiveness of small patient populations.

When analyzing duration of PFS rather than tumor shrinkage, the null distribution was drawn from all tumor types (n=141) because no tumor type was so responsive as to dominate the distribution (Supplementary Note S8). Significantly smaller hazard ratios, which is indicative of longer PFS, were identified by a *no difference* test in cervical cancers (P=0.03; average PFS 9.4 months) and lung cancers (P=0.003; average PFS 6.8 months) but - strikingly - not in breast cancers (P=0.36; average PFS 4.6 months, Supplementary Table S2). It is noteworthy that only 5 neratinib-treated cervical cancers are reported, and empirical null distributions are consequently broad (Figure 1), yet the observed responses were sufficiently strong and durable to achieve statistical significance in the small sample (cervical tumors also met the criteria to begin Stage 2 and so additional patients are currently accruing; Table 1). Whereas lung cancers exceed *no difference* tests for both volume changes and hazard ratios, breast cancers differ from the overall population by volume change alone. Lung cancers therefore appear to represent a therapeutic opportunity for neratinib missed by dichotomous criteria.

SUMMIT enrolled patients on the basis of qualifying mutations in ERBB2 or ERBB3, which were classified as 'hotspot' if they occurred in recurrently mutated regions of either gene, or 'non-hotspot' if they lay in rarely mutated regions (7). Tumors with ERBB2 hotspot mutations exceeded the *no-difference* null model as judged by changes in tumor volume or PFS (Figure 2) (P=0.0005 for PFS and P=0.03 for volume changes), which agrees with Hyman's conclusion of ERBB2 hotspot tumors as responsive to therapy. When ERBB2 hotspot mutations were further divided into functional classes (e.g. S310; Exon 20 insertions; V777; PKD; L755; and "other hotspot mutations"), Exon 20 insertions

bioRxiv preprint doi: https://doi.org/10.1101/401620; this version posted August 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

significantly exceeded *no difference* by PFS (P=0.01), which can be attributed almost exclusively to lung tumors (7) (6 lung tumors are among the 7 most durable responses observed for all cancer types having Exon 20 insertions). No other significant signals were detected by mutation class.

A basket trial for the kinase inhibitor larotrecinib (4) provides a contrasting example to SUMMIT. Drilon *et al* (4) recently reported very high rates of larotrectinib response among 54 patients with 17 different types of tumors, all of which expressed a TRK fusion protein. When data for the 9 tumor types represented by 2 or more patients were compared to a *no-difference* null, no significant difference was observed (Figure 3, Supplementary Table S7). This confirms the conclusions of Drilon *et al*. that sensitivity to larotrecinib is dominated by the presence of a TRK fusion and not by tissue of origin.

DISCUSSION

One of the primary motivations for performing a basket trial is to determine whether tumor type or genotype influences drug sensitivity. Because basket trials only rarely have empirical null distributions obtained from a no-treatment control population, contemporary designs for basket trials set a relatively high binary criterion for a meaningful response. For example, in the widely used Simon two-stage design, one or more of the few patients in Stage 1 must experience a change in tumor volume exceeding 30%. In this paper we demonstrate an alternative approach in which volume change or PFS, evaluated as continuous variables for specific tumor or mutation types, can be formally compared to empirical distributions to test the null hypothesis of *no difference by tumor type* (or by *mutation class/genotype*). In the case of trials such as SUMMIT, which are characterized by low average response rates, the *no difference* test approximates a *no effect* tests; in trials with high response rates, such as larotrectinib in cancers carrying TRK fusions, *no difference* tests for both superiority and inferiority.

bioRxiv preprint doi: https://doi.org/10.1101/401620; this version posted August 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Testing for significance in volume changes and PFS involves repeated Monte Carlo resampling of the all-patient distribution to create a null distribution that has the same number of samples as the number of patients of a particular type. The resulting distribution appropriately anticipates the greater variability observed in small cohorts, which adjusts the threshold for identifying a statistically significant deviation in response. For example, the SUMMIT trial reported PFS data for only five cervical cancer patients. In this case, the null distribution of hazard ratios was calculated by repeatedly sampling five response durations from the set of all patients, generating a relatively wide null distribution. Nonetheless, the observed hazard ratio in cervical cancers was significantly smaller than the *no difference* null distribution (P=0.03).

Conclusions drawn from testing for *no difference* in volume changes can differ substantially from those implicit in the Simon two-stage design; in addition, it is possible to evaluate PFS data. We found that the responses of lung cancers to neratinib exceed the *no difference* null with respect to both volume changes (P=0.04; sampling from all non-breast tumors) and PFS (P=0.003, sampling from all tumors) even though this tumor type failed the two-stage criteria. In contrast, breast cancers exhibited highly significant changes in tumor volume by both Simon and *no difference* criteria, but failed the *no difference* test with respect to PFS. Overall, the strongest positive signal for a genetic subtype was for Exon 20 mutations, which exceeded the *no difference* test only for PFS (P=0.005), primarily as a consequence of responses in lung cancers. More generally, our reanalysis of SUMMIT data is consistent with previous literature (18-23) in demonstrating substantial differences between volume changes and PFS data: significant reductions in tumor volume do not necessarily predict durable PFS, and durable PFS can be achieved with modest changes in tumor volume.

In the case of a basket trial for larotrectinib in TRK fusion-positive cancers in which overall response rates were high (4), no subpopulation was identified that was more or less drug-responsive than average. Thus, a formal *no difference* test confirms the results reported by Drilon *et al* that the presence

bioRxiv preprint doi: https://doi.org/10.1101/401620; this version posted August 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of a TRK fusion is an excellent predictor of larotrectinib response. Formally testing for superiority and inferiority is likely to be generally useful for targeted therapies that may have high activity in many but possibly not all tumor subtypes.

Caveats to the approach described here are similar to those for basket trials in general: findings are based on a small number of patients and must be interpreted in the context of known differences between tumor subtypes and patient populations, especially regarding systematic differences in tumor growth rates that may affect duration of PFS (24). Moreover, differences in response between tumor types may be obscured when one or more tumor types are such strong outliers as to render any smaller difference undetectable. In SUMMIT data this is observed for volume changes in neratinib-treated breast cancers ($P<10^{-6}$ relative to the *no difference* null). To enable detection of next-most different volume responses, we removed breast cancers from the null distribution. However, repeated adjustment of the null distribution is strongly discouraged since it can generate null hypotheses that are not representative of average efficacy, and in an extreme case it constitutes "*p-hacking*" (see Supplementary Note S8).

In conclusion, we describe a simple Monte-Carlo permutation test for small patient populations that makes it possible to obtain appropriately scaled null distributions and derive empirical *P* values for drug response as measured by both volume changes and PFS. The methodology is especially valuable in in basket trials and other Phase II studies that lack a control arm and involve multiple patient subgroups generally thought to be too small for formal comparison. As a consequence, we can obtain more data rich insight into therapeutic response than the dichotomous criteria used in Simon two-stage and other basket trial designs. For example, whereas scoring SUMMIT by Simon highlights volume changes in breast cancer as a superior response to neratinib, scoring for *no difference* finds no benefit in breast cancer by PFS. Lung cancer does not meet the Simon threshold for progression to Stage 2 enrollment (only 1 of 26 patients with lung tumors had a response surpassing 30% tumor shrinkage) but it significantly exceeds the *no difference* null with respect to volume changes and PFS and it is the only

bioRxiv preprint doi: https://doi.org/10.1101/401620; this version posted August 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

tumor type for which a specific genotype is associated with stronger response (ERBB2 Exon 20 mutation). We propose that the SUMMIT trial further study neratinib in ERBB-mutant lung tumors, particularly because benefits to overall survival have previously been shown to more strongly correlate with duration PFS than tumor shrinkage (18-22).

The continuing growth of genomic-driven oncology will increasingly enable refined subdivision of patient populations whether in a basket trial or by stratifying patients in conventional Phase II and Phase III studies (7). The promise of such subdivision is better detection of predictive biomarkers but the risk is smaller subsamples and reduced statistical significance. Monte Carlo resampling of response data and reformulation of null hypotheses may be generally useful in these cases.

Acknowledgments: We thank L. Trippa and B. Alexander for helpful discussions and comments on this manuscript.

Funding: This work was supported by NIH grants P50-GM107618 and U54-CA225088 (to PKS). D.P. is supported by NIGMS grant T32GM007753.

Author contributions: Analysis, A.C.P. and D.P.; writing, A.C.P., D.P., and P.K.S.

Competing interests: The authors declare no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/401620; this version posted August 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

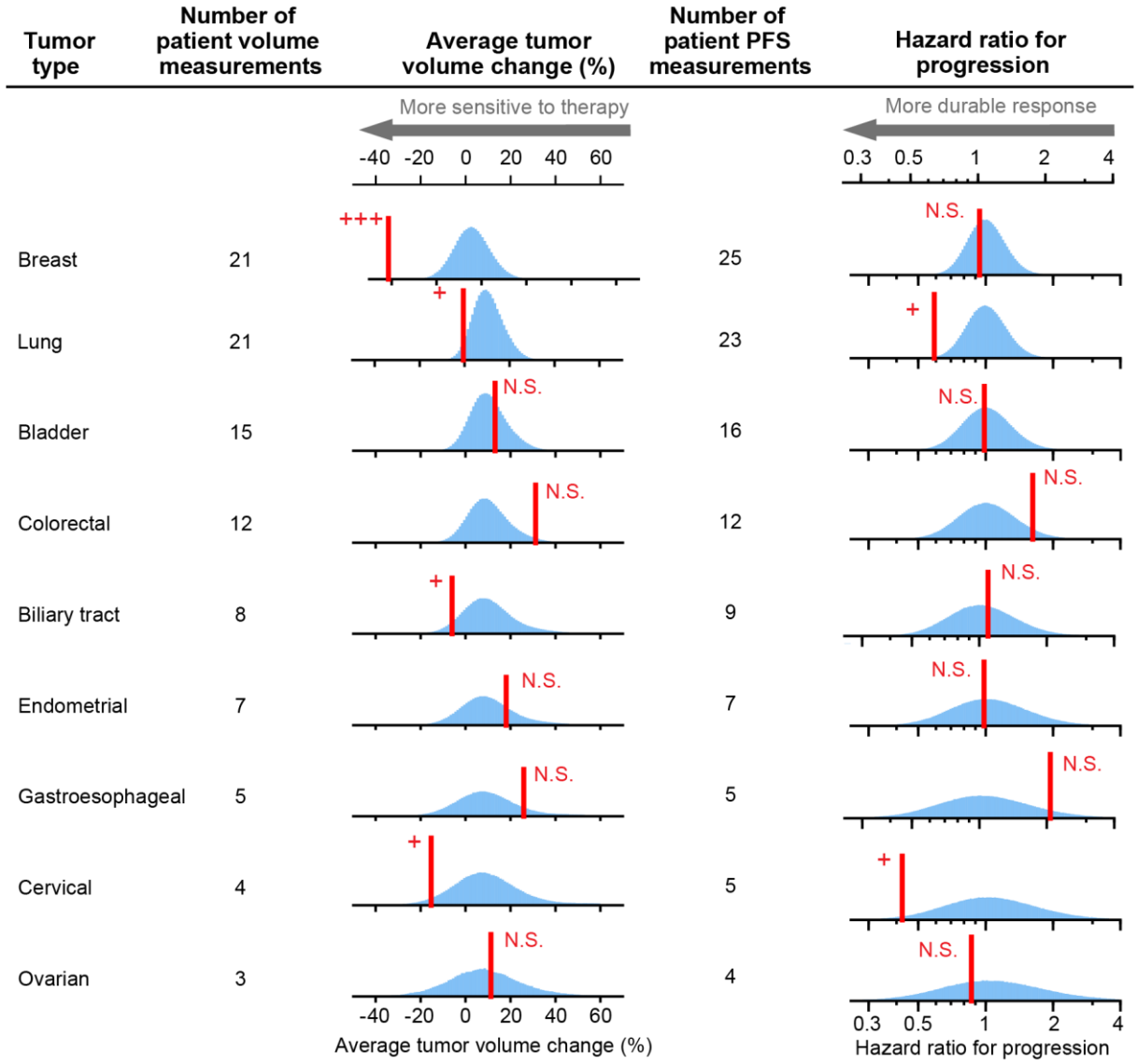

Figure 1. Analysis of neratinib response by tumor tissue of origin.

Red line: observed average response. Blue histogram: responses simulated according to the null hypotheses of *no difference* in response between tumors types. For volume changes, breast tumors are compared with null distributions drawn by Monte Carlo resampling from all tumors; for all other tumors, the null distributions are drawn from all non-breast tumors. Hazard ratio for progression null distributions are drawn from all tumors. Observed responses that significantly exceed the null hypothesis, according to Benjamini-Hochberg procedure for multiple hypothesis testing, are indicated with +; N.S. denotes not significant; +++ denotes $p < 10^{-6}$ (Supplementary Tables S1, S2).

bioRxiv preprint doi: https://doi.org/10.1101/401620; this version posted August 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

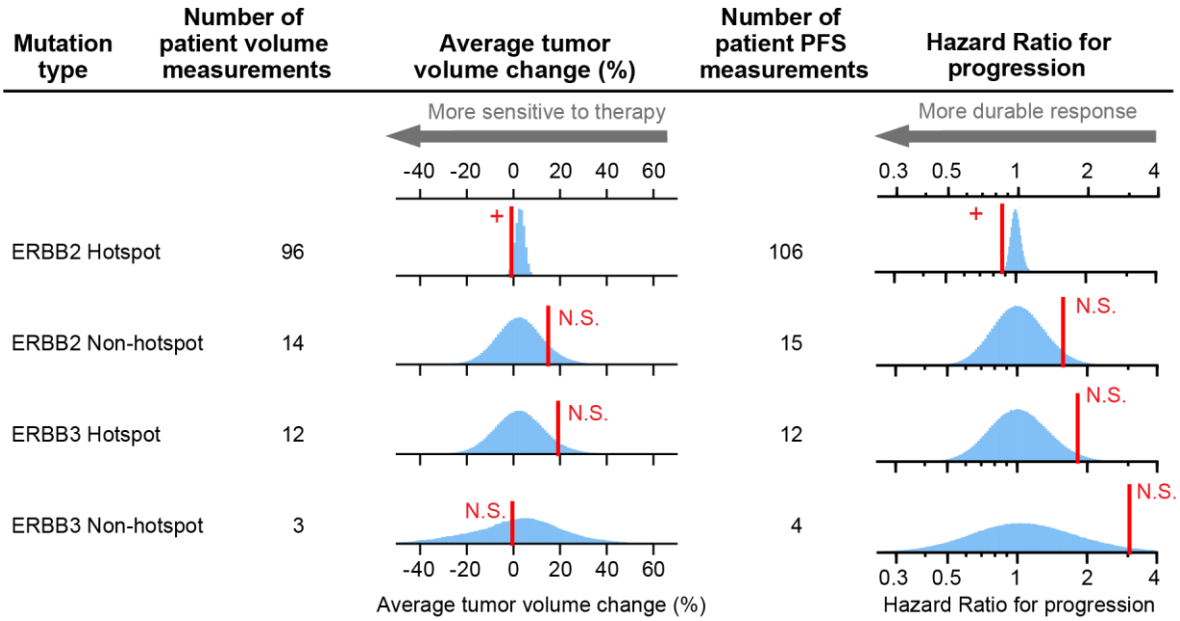

Figure 2. Analysis of neratinib response by mutation type.

Red line: observed average response. Blue histogram: responses simulated according to the null hypothesis of *no difference* in response between mutation types. Observed responses that significantly exceed the null hypothesis, according to Benjamini-Hochberg procedure for multiple hypothesis testing, are indicated with +; N.S. denotes not significant (Supplementary Tables S3, S4).

bioRxiv preprint doi: https://doi.org/10.1101/401620; this version posted August 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

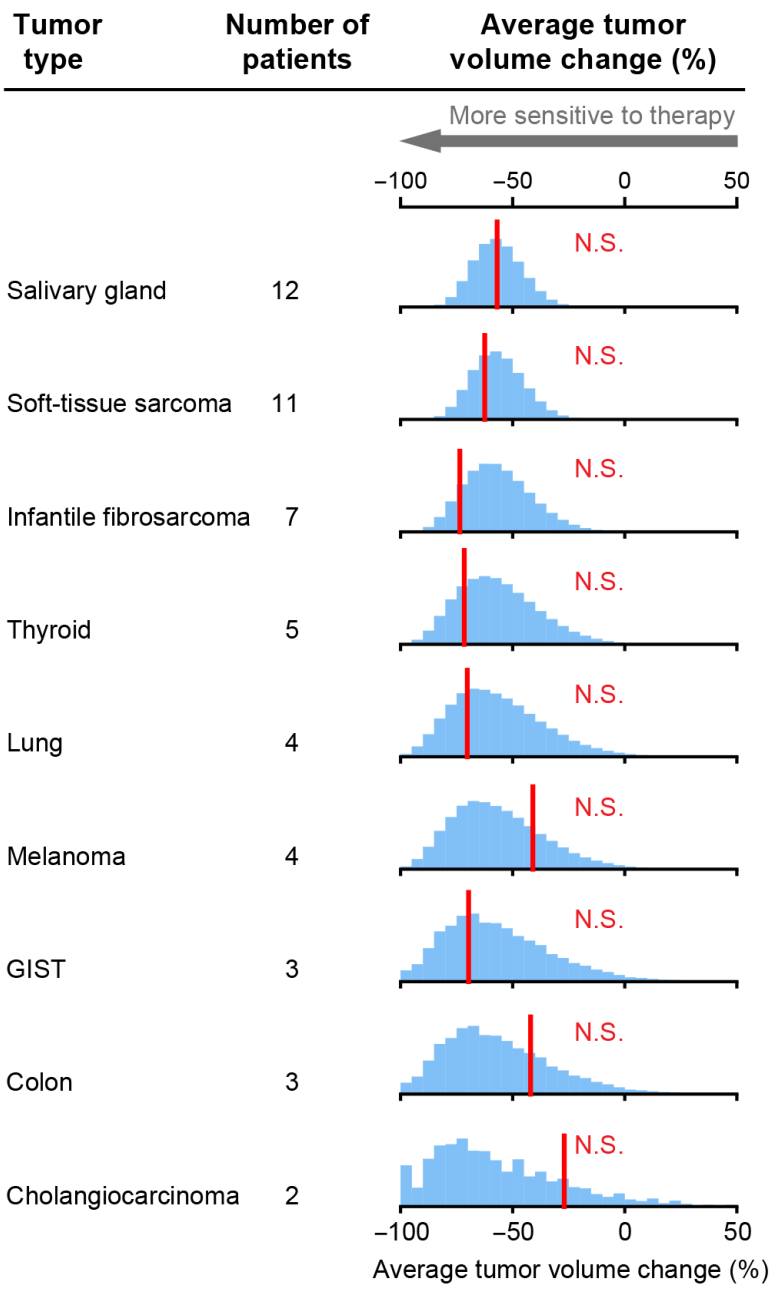

Figure 3. Analysis of larotrectinib response by tumor tissue of origin.

Red line: observed average response. Blue histogram: responses simulated according to the null hypothesis of *no difference* in response between tumors types. N.S. denotes not significant (Supplementary Table S7).

bioRxiv preprint doi: https://doi.org/10.1101/401620; this version posted August 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 1. Conclusions for neratinib in context of SUMMIT trial status.

Tumor type	Number of patients	Status in Simon Optimal 2-stage design		Responses significantly different from other tumors[1]?	
		Stage 1	Stage 2	Volume	PFS
Ovarian	4	Ongoing		-	-
Gastroesophageal	5	Ongoing		-	-
Colorectal	12	Failed		-	-
Bladder	16	Failed		-	-
Endometrial	7	Failed		-	-
Biliary	9	Passed	Ongoing	Superior	-
Cervical	5	Passed	Ongoing	Superior	Superior
Lung	26	Passed	Failed	Superior	Superior
Breast	25	Passed	Passed	Superior	-

[1] Dash denotes no significant difference by Benjamini-Hochberg procedure.

bioRxiv preprint doi: https://doi.org/10.1101/401620; this version posted August 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

REFERENCES

1. Tao JJ, Schram AM, Hyman DM. Basket Studies: Redefining Clinical Trials in the Era of Genome-Driven Oncology. Annual review of medicine **2018**;69:319-31 doi 10.1146/annurev-med-062016-050343.

2. Woodcock J, LaVange LM. Master Protocols to Study Multiple Therapies, Multiple Diseases, or Both. The New England journal of medicine **2017**;377(1):62-70 doi 10.1056/NEJMra1510062.

3. Redig AJ, Janne PA. Basket trials and the evolution of clinical trial design in an era of genomic medicine. Journal of clinical oncology : official journal of the American Society of Clinical Oncology **2015**;33(9):975-7 doi 10.1200/JCO.2014.59.8433.

4. Drilon A, Laetsch TW, Kummar S, DuBois SG, Lassen UN, Demetri GD, *et al.* Efficacy of Larotrectinib in TRK Fusion-Positive Cancers in Adults and Children. The New England journal of medicine **2018**;378(8):731-9 doi 10.1056/NEJMoa1714448.

5. Le DT, Durham JN, Smith KN, Wang H, Bartlett BR, Aulakh LK, *et al.* Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. Science **2017**;357(6349):409-13 doi 10.1126/science.aan6733.

6. Korphaisarn K, Kopetz S. BRAF-Directed Therapy in Metastatic Colorectal Cancer. Cancer journal **2016**;22(3):175-8 doi 10.1097/PPO.0000000000000189.

7. Hyman DM, Piha-Paul SA, Won H, Rodon J, Saura C, Shapiro GI, *et al.* HER kinase inhibition in patients with HER2- and HER3-mutant cancers. Nature **2018**;554(7691):189-94 doi 10.1038/nature25475.

8. Leblanc M, Rankin C, Crowley J. Multiple Histology Phase II Trials. Clinical cancer research : an official journal of the American Association for Cancer Research **2009**;15(13):4256-62 doi 10.1158/1078-0432.CCR-08-2069.

9. Hyman DM, Puzanov I, Subbiah V, Faris JE, Chau I, Blay JY, *et al.* Vemurafenib in Multiple Nonmelanoma Cancers with BRAF V600 Mutations. The New England journal of medicine **2015**;373(8):726-36 doi 10.1056/NEJMoa1502309.

10. Simon R, Geyer S, Subramanian J, Roychowdhury S. The Bayesian basket design for genomic variant-driven phase II trials. Seminars in oncology **2016**;43(1):13-8 doi 10.1053/j.seminoncol.2016.01.002.

11. Cunanan KM, Iasonos A, Shen R, Begg CB, Gonen M. An efficient basket trial design. Statistics in medicine **2017**;36(10):1568-79 doi 10.1002/sim.7227.

12. Cunanan KM, Gonen M, Shen R, Hyman DM, Riely GJ, Begg CB, *et al.* Basket Trials in Oncology: A Trade-Off Between Complexity and Efficiency. Journal of clinical oncology : official journal of the American Society of Clinical Oncology **2017**;35(3):271-3 doi 10.1200/JCO.2016.69.9751.

bioRxiv preprint doi: https://doi.org/10.1101/401620; this version posted August 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

13. Simon R. Optimal two-stage designs for phase II clinical trials. Controlled clinical trials **1989**;10(1):1-10.

14. Ratain MJ, Karrison TG. Testing the wrong hypothesis in phase II oncology trials: there is a better alternative. Clinical cancer research : an official journal of the American Association for Cancer Research **2007**;13(3):781-2 doi 10.1158/1078-0432.CCR-06-2533.

15. Stone A, Wheeler C, Barge A. Improving the design of phase II trials of cytostatic anticancer agents. Contemporary clinical trials **2007**;28(2):138-45 doi 10.1016/j.cct.2006.05.009.

16. El-Maraghi RH, Eisenhauer EA. Review of phase II trial designs used in studies of molecular targeted agents: outcomes and predictors of success in phase III. Journal of clinical oncology : official journal of the American Society of Clinical Oncology **2008**;26(8):1346-54 doi 10.1200/JCO.2007.13.5913.

17. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological) **1995**;57(1):289-300.

18. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? Annals of internal medicine **1996**;125(7):605-13.

19. Zabor EC, Heller G, Schwartz LH, Chapman PB. Correlating Surrogate Endpoints with Overall Survival at the Individual Patient Level in BRAFV600E-Mutated Metastatic Melanoma Patients Treated with Vemurafenib. Clinical cancer research : an official journal of the American Association for Cancer Research **2016**;22(6):1341-7 doi 10.1158/1078-0432.CCR-15-1441.

20. Kaiser LD. Tumor burden modeling versus progression-free survival for phase II decision making. Clinical cancer research : an official journal of the American Association for Cancer Research **2013**;19(2):314-9 doi 10.1158/1078-0432.CCR-12-2161.

21. Fridlyand J, Kaiser LD, Fyfe G. Analysis of tumor burden versus progression-free survival for Phase II decision making. Contemporary clinical trials **2011**;32(3):446-52 doi 10.1016/j.cct.2011.01.010.

22. Seymour L, Ivy SP, Sargent D, Spriggs D, Baker L, Rubinstein L, *et al.* The design of phase II clinical trials testing cancer therapeutics: consensus recommendations from the clinical trial design task force of the national cancer institute investigational drug steering committee. Clinical cancer research : an official journal of the American Association for Cancer Research **2010**;16(6):1764-9 doi 10.1158/1078-0432.CCR-09-3287.

23. Valachis A, Polyzos NP, Patsopoulos NA, Georgoulias V, Mavroudis D, Mauri D. Bevacizumab in metastatic breast cancer: a meta-analysis of randomized controlled trials. Breast cancer research and treatment **2010**;122(1):1-7 doi 10.1007/s10549-009-0727-0.

24. Friberg S, Mattson S. On the growth rates of human malignant tumors: implications for medical decision making. Journal of surgical oncology **1997**;65(4):284-97.

bioRxiv preprint doi: https://doi.org/10.1101/401620; this version posted August 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary data

Tumor Type	Test for larger benefit in tumor volume		Conclusion based on 25% False Discovery Rate
	p-value	Benjamini-Hochberg critical value for 1-sided test	
Cervical	**0.039**	**0.031**	**Larger Benefit**
Lung	**0.040**	**0.063**	**Larger Benefit**
Biliary Tract	**0.059**	**0.094**	**Larger Benefit**
Ovarian	0.569	0.125	N.S.
Bladder	0.659	0.156	N.S.
Endometrial	0.768	0.188	N.S.
Gastroesophageal	0.872	0.219	N.S.
Colorectal	0.977	0.250	N.S.

Supplementary Table S1. Benjamini-Hochberg critical values for analysis of neratinib tumor volume responses by tumor tissue of origin.

Tumor Type	Test for larger benefit in hazard ratio for progression		Conclusion based on 25% False Discovery Rate
	p-value	Benjamini-Hochberg critical value for 1-sided test	
Lung	**0.003**	**0.028**	**Larger Benefit**
Cervical	**0.027**	**0.056**	**Larger Benefit**
Ovarian	0.347	0.083	N.S.
Breast	0.363	0.111	N.S.
Endometrial	0.454	0.139	N.S.
Bladder	0.467	0.167	N.S.
Biliary tract	0.579	0.194	N.S.
Gastroesophageal	0.912	0.222	N.S.
Colorectal	0.938	0.250	N.S.

Supplementary Table S2. Benjamini-Hochberg critical values for analysis of neratinib hazard ratios for progression by tumor tissue of origin.

bioRxiv preprint doi: https://doi.org/10.1101/401620; this version posted August 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Mutation Type	Test for larger benefit in tumor volume		Conclusion based on 25% False Discovery Rate
	p-value	Benjamini-Hochberg critical value for 1-sided test	
ERBB2 Hotspot	**0.030**	**0.063**	**Larger benefit**
ERBB3 Nonhotspot	0.424	0.125	N.S.
ERBB2 Nonhotspot	0.889	0.188	N.S.
ERBB3 Hotspot	0.931	0.250	N.S.

Supplementary Table S3. Benjamini-Hochberg critical values for analysis of neratinib tumor volume responses by general mutation type.

Mutation Type	Test for larger benefit in hazard ratio for progression		Conclusion based on 25% False Discovery Rate
	p-value	Benjamini-Hochberg critical value for 1-sided test	
ERBB2 Hotspot	**0.0005**	**0.063**	**Larger benefit**
ERBB2 Nonhotspot	0.950	0.125	N.S.
ERBB3 Nonhotspot	0.962	0.188	N.S.
ERBB3 Hotspot	0.970	0.250	N.S.

Supplementary Table S4. Benjamini-Hochberg critical values for analysis of neratinib hazard ratios for progression by general mutation type.

bioRxiv preprint doi: https://doi.org/10.1101/401620; this version posted August 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Mutation Type	Test for larger benefit in tumor volume		Conclusion based on 25% False Discovery Rate
	p-value	Benjamini-Hochberg critical value for 1-sided test	
L755 Hotspot	0.031	0.025	N.S.
Exon20 Insertion Hotspot	0.162	0.050	N.S.
S310 Hotspot	0.267	0.075	N.S.
V777 Hotspot	0.276	0.100	N.S.
ERBB3 Nonhotspot	0.421	0.125	N.S.
Other Hotspot	0.529	0.150	N.S.
PKD Nonhotspot	0.651	0.175	N.S.
ERBB3 Hotspot	0.931	0.200	N.S.
PKD Hotspot	0.950	0.225	N.S.
Other Nonhotspot	0.954	0.250	N.S.

Supplementary Table S5. Benjamini-Hochberg critical values for analysis of neratinib tumor volume responses by specific mutation type.

Mutation Type	Test for larger benefit in hazard ratio for progression		Conclusion based on 25% False Discovery Rate
	p-value	Benjamini-Hochberg critical value for 1-sided test	
Exon20 Insertion Hotspot	**0.011**	**0.025**	**Larger Benefit**
S310 Hotspot	0.094	0.050	N.S.
Other Hotspot	0.274	0.075	N.S.
L755 Hotspot	0.456	0.100	N.S.
PKD Hotspot	0.499	0.125	N.S.
PKD Nonhotspot	0.759	0.150	N.S.
V777 Hotspot	0.944	0.175	N.S.
ERBB3 Nonhotspot	0.962	0.200	N.S.
ERBB3 Hotspot	0.970	0.225	N.S.
Other Nonhotspot	0.985	0.250	N.S.

Supplementary Table S6. Benjamini-Hochberg critical values for analysis of neratinib hazard ratios for progression by specific mutation type.

bioRxiv preprint doi: https://doi.org/10.1101/401620; this version posted August 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Tumor Type	Test for larger benefit in tumor volume		Test for smaller benefit in tumor volume		Conclusion based on 25% False Discovery Rate
	p-value	Benjamini-Hochberg critical value for 2-sided test	p-value	Benjamini-Hochberg critical value for 2-sided test	
Infantile fibrosarcoma	0.120	0.014	0.877	0.125	N.S.
Thyroid tumor	0.212	0.028	0.784	0.111	N.S.
Lung tumor	0.275	0.042	0.720	0.097	N.S.
Soft-tissue sarcoma	0.317	0.055	0.680	0.083	N.S.
GIST	0.330	0.069	0.662	0.069	N.S.
Salivary-gland tumor	0.503	0.083	0.493	0.055	N.S.
Colon tumor	0.749	0.097	0.248	0.042	N.S.
Melanoma	0.785	0.111	0.211	0.028	N.S.
Cholangiocarcinoma	0.839	0.125	0.159	0.014	N.S.

Supplementary Table S7. Benjamini-Hochberg critical values for analysis of larotrectinib tumor volume responses by tumor type.

bioRxiv preprint doi: https://doi.org/10.1101/401620; this version posted August 28, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Note S8. On the removal of outlier subgroups from null distributions.

We use Monte Carlo resampling to derive empirical *P* values for volume changes and PFS differences among tumor subgroups; this involves the generation of test distributions by resampling from all observed responses, from all tumor subgroups, in the trial. The resampling draws a number of samples from the all-patient distribution that equals the number of patient responses measured for that cancer type in the test cohort. In the case of the SUMMIT trial, volume (but not PFS) changes in breast tumors were far stronger than for any other tumor type: none of 10^7 simulations of the null hypothesis matched the observed average tumor volume change of breast tumors (we report this as $P < 10^{-6}$). The magnitude of difference between breast tumors and all tumors (45% difference in average volume change) is so large that the inclusion of breast tumors in the null distribution makes it impossible to detect any difference among other tumor types. Because breast tumors represent an outlier with regard to volume changes in response to neratinib treatment, we considered it inappropriate to include breast tumor volume changes in the between-tumor comparison of all other tumor types. We therefore constructed a "no breast tumor" (NB) null distribution for these tests. This reformulation of the null distribution was applied only for this case of a $P<10^{-6}$ outlier, and we advocate for a similarly stringent approach to any future application that may remove subtypes from the null distribution.

There was no other subtype in our analysis for which reformulation of the null hypothesis was appropriate. For example, for PFS data in SUMMIT, lung tumors showed the most durable response (P=0.003). This is significant under the Benjamini-Hochberg procedure in the context of a False Discovery Rate of 25% but it does not hinder the identification of the next-most durable responders, cervical tumors (P=0.03). Thus, a reformulation of the PFS null distribution is not justified. For volume changes reported in SUMMIT, after removing breast tumors from the null distribution, lung and cervical are the next most significant, each with P=0.04. Because their significance is associated with a 25% FDR in this case, reformulating the null distribution is also not justified.