Matters arising

Reanalysis of in vivo drug synergy validation study rules out synergy in most cases

Olaf van Tellingen [ID][1]✉ & Renee X. de Menezes [ID][2]

Narayan et al.[1] described the development of an in silico platform ("Drug Atlas") to predict synergistic drug combinations and claimed successful in vivo validation in five models. Demonstrating the impact of a given drug in in vivo models is challenging because effect sizes are often small[2]. Because of the inherent variability in tumor outgrowth in vivo, achieving adequate statistical power requires sufficiently sized groups. As the experimental complexity increases, for example, when assessing combination effects, the minimum required group size also increases. A statistically sound distinction between additive and synergistic drug effects requires even larger datasets, as do comparisons across multiple drug combinations of drugs.

The formula used by Narayan et al. to assess synergy in vivo was derived from the work by Chou and Talalay[3]. The theoretical foundation of estimating drug interaction in vitro by the methods as described by Chou and Talalay is well established[4]. The basis of this method is the Combination Index (CI), where a CI < 1 implicates synergy, CI = 1 additivity, and CI > 1 antagonism. Importantly, the calculation of the CI requires testing of each individual drug and their combination in a fixed dose ratio at multiple dose levels to create dose-effect curves and assess IC_{50} values[5]. Chou has also applied this methodology to demonstrate synergy in vivo using 65 mice divided over 13 groups of 5 mice each[6]. In this setup, each drug alone and its pairwise combinations were tested at 4 different dose levels, and compared with one untreated control group.

Overall, Narayan et al.[1] claim the successful in vivo validation of synergy in 5 separate experiments, as presented in Fig. 5 of their paper. We will refer to these separate experiments by their corresponding subfigure labels 5A–5E. Experiments depicted in 5A and 5B concern a U87 glioblastoma model, 5C a triple negative breast cancer model, 5D a melanoma model, and 5E a chronic myeloid leukemia model. Growth kinetics of the tumors was assessed by weekly bioluminescence imaging (BLI) measurements, which were used to calculate CI values at each measurement event. Next to that, recorded overall survival times were reported as Kaplan–Meier curves. The paper by Narayan et al.[1] drew our attention due to its claim of demonstrating synergy using only a few and relatively small-sized animal cohorts. Hence, we requested the data to conduct an independent re-analysis.

Findings

Instead of testing each drug alone and in their combinations at multiple dose levels, Narayan et al.[1] employed a shortcut by testing the single agents and their combinations at only one dose level, thus ignoring the requirement to generate full dose-response curves. Moreover, to calculate the CI, they designed a formula that appeared to be biased towards predicting synergy, especially when the individual drugs already exhibit substantial efficacy (i.e., low treatment versus control (T/C) ratios) at the tested dose level (see Fig. 1). To demonstrate how the drug atlas enables synergy predictions, the authors presented their findings on the combination of a microtubule-stabilizing agent, an EGFR inhibitor and an mTOR1/2 inhibitor across glioblastoma (GBM) cell lines. Using this combination as a backbone, they evaluated two triple-drug regimens in vivo in the U87 GBM model. In this case, however, the selected dose level of docetaxel was already at the extreme end of the dose-response curve. For example, in one study, docetaxel alone led to a 98.4% reduction in tumor volume (T/C value: 1.6%). The addition of osimertinib (EGFR inhibitor) and AZD2014 (PI3K/mTOR inhibitor) had only a limited impact relative to the reduction in tumor volume achieved by docetaxel alone: i.e., an additional 0.9%, resulting in a final reduction of 99.3% (T/C = 0.7%). Using their formula, this marginal increase resulted in a calculated CI of 0.55, sufficiently below the theoretical threshold of 1 for claiming synergy. However, osimertinib and AZD2014 each alone also considerably reduced tumor growth (T/C is 29 and 12%, respectively). Assuming that docetaxel, osimertinib, and AZ2016 only have additive effects, their formula yields a value of 0.0402 as the theoretical threshold, which is considerably below both the commonly used threshold of 1 as well as the calculated CI of 0.55. Thus, according to their formula, the combination of docetaxel and osimertinib + AZ2014 at these dose levels is actually antagonistic compared to docetaxel alone.

Experiments 5A, 5B, and 5E

We first conducted a gross examination of the data. Apart from the abovementioned methodological concerns, we uncovered major issues in data handling and reporting that are inconsistent with established standards of good scientific practice. We invite the readers

[1]Division of Pharmacology and Mouse Cancer Clinic, The Netherlands Cancer Institute, Amsterdam, The Netherlands. [2]Biostatistics Centre and Psychosocial Research and Epidemiology Department, The Netherlands Cancer Institute, Amsterdam, The Netherlands. ✉e-mail: o.v.tellingen@nki.nl
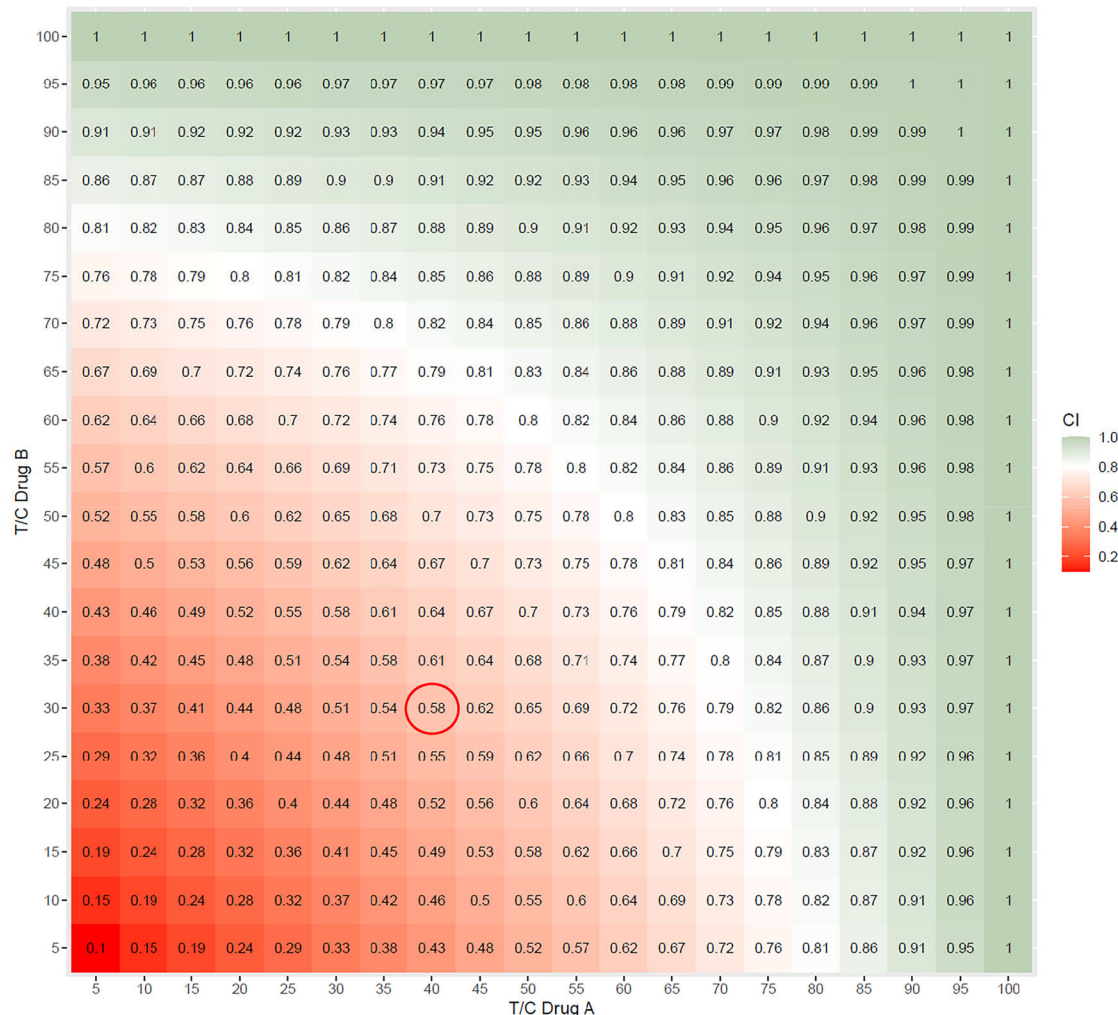
Fig. 1 | Matrix of calculated CI values of additive combinations. Values of the Combination Index (CI) calculated using the same formula as Narayan et al. for 2-drug combinations, assuming only additive effects. For example, when drug A has a T/C of 40% and B has a T/C of 30%, the combination will have a T/C of 12% when additive (40% × 30% = 12%). When we apply these numbers to the formula of Narayan et al.: $CI_{(n, \text{drugs})} = \dfrac{\sum_{k=0}^{n}\left(\frac{1}{v_n}\right)-\left(\frac{n-1}{100}\right)}{\left(\frac{1}{v_{1.n}}\right)}(10)$ the calculated value of the CI yields

$((1/40) + (1/30))-((2-1)/100))/(1/12) = 0.58$. A correct formula should have returned

a CI of 1, as this is an example of a purely additive combination. The color pattern indicates with red combinations which are automatically considered as synergistic, while they are actually just additive. As shown, the calculated CI threshold value declines when single-agent efficacy increases. Hence, the calculated value is not a bonafide CI and cannot be used to assess synergy. We now refer to the values in the figure as the threshold values for calling synergy. For details and calculations, see SI File 1: Part 2 Combination Index in in vivo models.

to consult the Supplementary Information File (Part 1. Detailed Analysis). To summarize the most striking examples:

- Survival data from two separate animal experiments (5A and 5B) were pooled, although these experiments had different study designs and were conducted at different locations. Moreover, this was not disclosed in the "Methods."
- The survival curve of Experiment 5A was not based on overall survival, but on an arbitrary date of progression assessment. This was also not disclosed in the "Methods."
- The p values associated with survival analyses were artificially inflated via the use of a one-sided t-test, rather than a log-rank test, as is the norm in the field. The actual calculations contained many flaws, such as comparing mean versus median values, and the inclusion of censored animals to increase the sample size and which further inflated p values.
- The reported p values for survival were based solely on comparisons between the combination treatment and the control group, omitting comparisons with the respective single-agent

treatments. This limitation is not clearly disclosed in the "Methods" section, and such an analysis does not constitute evidence of synergy.

In addition to these data integrity issues, further concerns arise regarding the experimental protocol and mistakes that affected the results. Concerning Experiment 5A:

- There were many faulty injections of luciferin, causing very low BLI values, which in some cases also affected measurements at day 0, which is the reference for all subsequent BLI measurements.
- Rather than computing tumor growth by normalizing each mouse's BLI signal to its own day 0 value, the analysis was performed using the ratio of group mean BLI values at each time point relative to the group mean on day 0.
- The last recorded BLI data of animals that died during the study were used as input values for subsequent days (carry last value forward). As a result, low BLI values from animals in the triple combination group that died due to treatment-related toxicity
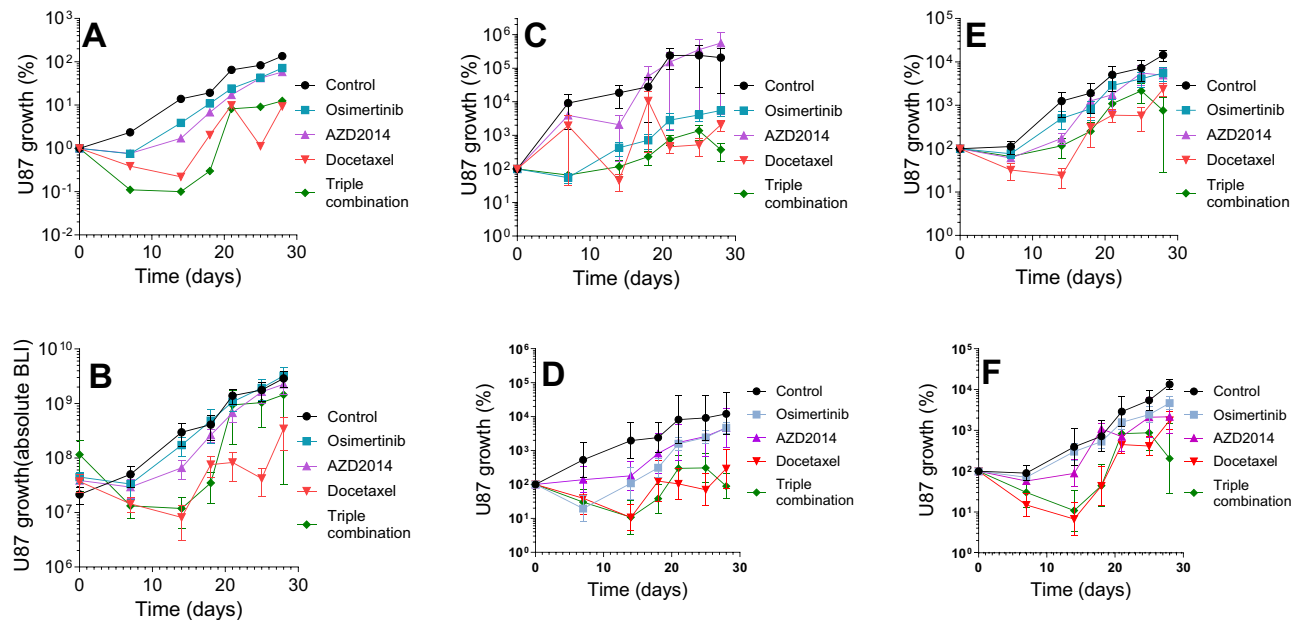
Fig. 2 | Re-analyzed data of Experiment 5A. A Data as used in the paper. Growth(%) was calculated from group means BLI per day, normalized to the group means at day 0 (Start of treatment). **B** Same original data depicted as mean absolute BLI values (±SE) without normalization. **C** Tumor growth(%) calculated as BLI values of each animal relative to its own BLI value at the start day. **E** The same data after rejection of incorrect injections and carry last value forward data points. **D**, **F** are similar to (**C**, **E**), respectively, but after log conversion of the BLI data. In (**B**–**F**), standard errors were computed at each time point separately, ignoring other time points. The control group contained $n = 6$ animals. All other groups were $n = 7$.

continued to influence and lower the group mean values on later days. This approach is misleading, as it assumes tumor stasis in animals that would likely have shown tumor progression had they survived.

- All calculations were performed on linear data, whereas a log-transformation of BLI data should have been applied.

We reanalyzed the data from Experiment 5A (Fig. 2). When we applied a log-transformation to the individual BLI data, normalized to baseline values at the start of treatment, and excluded both carried-forward values and outliers resulting from faulty luciferin injections, the response curve for docetaxel-monotherapy closely resembled that of the triple-drug combination (Fig. 2F). In fact, docetaxel was already highly efficacious when given alone; adding the other drugs did not result in further improvement. Hence, this combination is not synergistic.

Similarly, the drug combination in Experiment 5B did not demonstrate synergy. The study was underpowered, with only four animals per group remaining at $t = 12$ days, dropping to two animals in both Control and GNE-317 groups by the next time point ($t = 21$ days). Beyond this point, the control group had no surviving animals, and only one remained in the GNE-317 group. Consequently, any conclusions drawn from these data are unreliable as they are based on too small sample sizes.

In the case of Experiment 5E, the original data set contained an error (reference to the wrong cells) in the Excel sheet. After correction, the CI values calculated by the flawed formula are close to 1, where the threshold value would be 0.31 (see Fig. 1: Imatinib T/C 0.20 and dasatinib T/C 0.13), and synergy is not shown. The survival plots support the conclusion that the effect is additive. Both drugs are about equally active, increasing the median survival from 35.5 (control) to 58 (imatinib; +23 days) and 60.5 (dasatinib + 25 days), and to 75 days (+40 days) with the combination.

For these reasons, we did not analyze the data for 5A, 5B, and 5E further, considering only the remaining two experiments in the remainder of this document.

Experiments 5C and 5D

Data from the two remaining experiments appeared to be reliable, and processing was done to acceptable standards. We applied a log-transformation to the individual BLI data and normalized it to baseline values at the start of treatment. In Experiment 5C, AZ628 alone has an almost negligible effect, but when given with gemcitabine, it considerably augments the efficacy relative to gemcitabine alone (Fig. 3A). Even considering that the result of the comparison is at one dose level only, this result does suggest a synergistic interaction. Indeed, experimental variation is relatively low, and 95% bootstrap confidence intervals for the CIs comfortably exclude the offset value of CI for synergy, suggesting to reject the hypothesis of no synergy (Fig. 3B, left panel). In contrast, the outcome is much less clear for Experiment 5D (Fig. 3A, right panel). Both CGP-082996 and gemcitabine alone reduce tumor growth to 57% (T/C = 43%) and 32% (T/C = 68%) at day 28 relative to untreated controls, respectively. Their combination further reduces growth to 16%, resulting in a CI of 0.61 for that day. However, due to large experimental variation, the confidence intervals for the CI include values close to 1 (Fig. 3B, right panel) and certainly include 0.70 (the offset value for CI calculated from Fig. 1), suggesting that the reduction in tumor growth is not sufficient to exclude purely additive effects.

Experiments described in Narayan et al.'s Supplementary Fig. 5

Besides the in vivo results presented in Narayan et al.'s Fig. 5, there were also three additional models presented in their Supplementary Fig. 5. These comprised: MDA-MB-231 breast cancer treated with thapsigargin (inhibitor of sarco/endoplasmatic reticulum Ca²⁺ ATPase (SERCA)) and AZ628 (BRAF), HT29 colorectal cancer treated with vemurafenib (BRAF) and gemcitabine (nucleoside analog), and NCI H460 non-small lung cancer treated with A443654 (pan AKT) and mitomycin C (DNA cross-linker). While the latter two showed no survival benefit of the combination, synergy was called with the thapsigargin and AZD628 combination. However, we noted that neither the combination of these drugs, nor drugs interfering with the same targets, was listed as predicted by the Drug Atlas (see: Source Data File
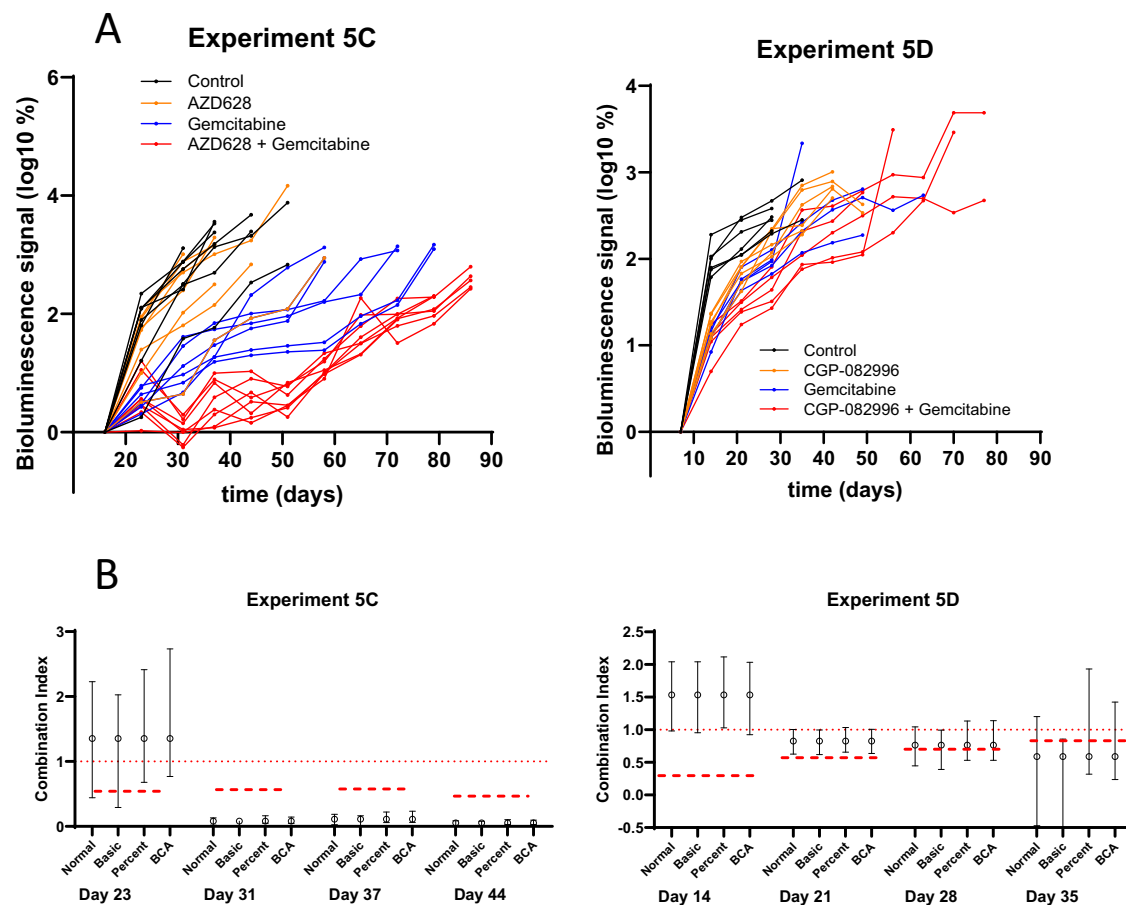
Fig. 3 | Reanalysis of Combination Index in in vivo models. A Growth curves of individual mice, depicted as the ratio of BLI values relative to BLI on the start day of treatment for Experiment 5C (left) and 5D (right). **B** Bootstrap analyses to determine the 95% confidence interval of the calculated combination indices. We make use of four different assumptions to build confidence intervals and display all four (see SI File 1: Part 2 Combination Index in in vivo models). The red dotted lines indicate the threshold CI as computed according to the formula of Nayaran et al. (see Fig. 1).

\Original data\Supplementary info\41467_2020_16735_MOESM6_ESM). Thus, although potentially synergistic, this combination was not predicted and therefore does not validate the Drug Atlas.

Final remarks

Synergy claims predicted by the Drug Atlas of Narayan et al.[1] cannot be confirmed in four out of five reported validation experiments, after re-analysis of their data. The original incorrect conclusions followed flawed data processing and analyses. While the concept of in silico synergy prediction may be scientifically valid in principle, in vivo validation requires far more rigorous and carefully designed animal studies than those presented in that paper. Critically, the dose levels selected for such studies should avoid inducing strong single-agent efficacy, as this can confound the interpretation of combinatorial effects. In addition, proper longitudinal data is essential−not only to more accurately capture treatment dynamics over time, but also to better account for intra- and inter-individual variation, which should improve statistical power, key to minimizing the number of animals required.

Methods

Upon request, we received the raw data files from the last senior author (Bart Westerman), consisting of an Excel file containing the BLI data and 5 Graphpad Prism files containing the information on survival. Later during this investigation, we received an updated version of the Excel file. All files are in the Source Data File (Original Data).

Confidence intervals for the Combination Index

Chou and Talalay's CI was originally proposed for use with in vitro-produced data for drug sensitivity at various doses for the drug combinations in fixed ratios. With a similar experimental design using multiple dose levels and sufficiently sized groups, it can also be applied to in vivo data[3]. Narayan et al.[1] used relatively small sample sizes (group sizes per experiment: 5A $n = 7$; 5B $n = 4$; 5 C $n = 8$; 5D $n = 6$ and 5E $n = 6$), without multiple dosages for drugs and/or their combinations. This means that, in this case, the CI cannot adequately model drug interaction, as the physicochemical mass-action law is not fully measured. The CI, as calculated using the formula proposed by Narayan et al., can only be applied simplistically and is highly sensitive to variation between individuals. Inter-individual variation can be considerable due to the nature of the animal models used. Moreover, as outlined above, the proposed formula is biased, so that 0.8 cannot be used as a general cut-off value to call synergy (Fig. 1).

To take these issues into account, we recalculated the CI per time point as in Narayan et al. using stratified bootstrapping, to better reflect experimental variability[7]. The stratification is used per treatment (control, drug 1, drug 2, or both drugs), and ensures that resampling considers the group structure. For details, see Supplementary Information File (Part 2. Combination Index in in vivo models).

Software

For all analyses, we used R v 3.6.3 and packages boot v 1.3-25 (Angelo Canty and Brian Ripley; 2020; boot: Bootstrap R (S-Plus) Functions). R

package version 1.3-25 and survival 3.2-7 (Therneau T; 2020; A Package for Survival Analysis in R. R package version 3.2-7, URL: https://CRAN.R-project.org/package=survival).

Data availability

The authors declare that the data supporting the findings of this study are available within the paper and the accompanying supplementary information files. Source data are provided with this paper.

References

1. Narayan, R. S. et al. A cancer drug atlas enables synergistic targeting of independent drug vulnerabilities. *Nat. Commun.* **11**, 2935 (2020).
2. Zavrakidis, I., Jozwiak, K. & Hauptmann, M. Statistical analysis of longitudinal data on tumour growth in mice experiments. *Sci. Rep.* **10**, 9143 (2020).
3. Chou, T. C. & Talalay, P. A simple generalized equation for the analysis of multiple inhibitions of Michaelis-Menten kinetic systems. *J. Biol. Chem.* **252**, 6438–6442 (1977).
4. Chou, T. C. Drug combination studies and their synergy quantification using the Chou-Talalay method. *Cancer Res.* **70**, 440–446 (2010).
5. Chou, T. C. & Talalay, P. Quantitative analysis of dose-effect relationships: the combined effects of multiple drugs or enzyme inhibitors. *Adv. Enzym. Regul.* **22**, 27–55 (1984).
6. Chou, T. C. Preclinical versus clinical drug combination studies. *Leuk. Lymphoma* **49**, 2059–2080 (2008).
7. Davison, A. C. & Hinkley, D. V. *Bootstrap Methods and their Application* (Cambridge University Press, 1997).

Author contributions

O.v.T.: conception, re-analyses of the data, writing of the paper. R.M.: re-analyses of the data, model development, writing of the paper.

Competing interests

O.v.T.: none. R.M. was recruited as group leader of the Biostatistics Centre of the Netherlands Cancer Institute. Before, she was a biostatistician at the VUmc, which is the institute of Narayan and colleagues. R.M. is a coauthor of the paper under discussion, but she was not involved in the analyses of the in vivo data.

Additional information

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.