CANCER



# Assessing therapy response in patient-derived xenografts

**Janosch Ortmann[1,2], Ladislav Rampášek[3,4,5], Elijah Tai[3], Arvind Singh Mer[6,7], Ruoshi Shi[6], Erin L. Stewart[6], Celine Mascaux[6,8,9], Aline Fares[6], Nhu-An Pham[6], Gangesh Beri[6], Christopher Eeles[6], Denis Tkachuk[6], Chantal Ho[6], Shingo Sakashita[6], Jessica Weiss[6], Xiaoqian Jiang[10], Geoffrey Liu[6], David W. Cescon[6], Catherine A. O'Brien[6,7,11,12,13], Sheng Guo[10], Ming-Sound Tsao[6], Benjamin Haibe-Kains[3,4,6,7,14]*, Anna Goldenberg[3,4,5,15]***

**Quantifying response to drug treatment in mouse models of human cancer is important for treatment development and assignment, yet remains a challenging task. To be able to translate the results of the experiments more readily, a preferred measure to quantify this response should take into account more of the available experimental data, including both tumor size over time and the variation among replicates. We propose a theoretically grounded measure, KuLGaP, to compute the difference between the treatment and control arms. We test and compare KuLGaP to four widely used response measures using 329 patient-derived xenograft (PDX) models. Our results show that KuLGaP is more selective than currently existing measures, reduces the risk of false-positive calls, and improves translation of the laboratory results to clinical practice. We also show that outcomes of human treatment better align with the results of the KuLGaP measure than other response measures. KuLGaP has the potential to become a measure of choice for quantifying drug treatment in mouse models as it can be easily used via the kulgap.ca website.**

## INTRODUCTION

Despite advances in pharmaceutical research, many patients with cancer do not respond to the first line of therapy. In oncology, researchers rely on preclinical models to investigate drug response to assess whether a drug works against a given cancer type. Of the available preclinical models, in vivo models tend to capture response to the drugs more faithfully than in vitro models (*1*). A standard readout from in vivo models is the size of the tumor growth over time across multiple experimental replicates compared to a set of untreated controls. As with most biological systems, tumor growth can vary within and between host mice, creating substantial variance among biological replicates. Determining whether the in vivo model is actually responsive to the given drug from the set of biological experiments is thus a complex task. An accurate determination of response is, however, essential as it has a direct impact on translation to the clinic.

Many measures have been proposed to quantify response to a treatment for in vivo models (*1*–*3*). Commonly used measures include modified response evaluation criteria in solid tumors (mRECIST), area under the curve (AUC) (*2*), angle of response (Angle) (*2*), and tumor growth inhibition (TGI) (*3*–*5*). Depending on which measure a researcher selects, the assessment of response may yield different, often opposite, conclusions, as none of the existing measures take full advantage of the data collected across replicates. For example, mRECIST (*1*) is easy to compute but does not take controls into account and is thus unable to distinguish true disease control (stable disease) from a naturally slow-growing tumor. The angle of response and TGI only take into account the last measurement rather than the full trajectory of treatment. AUC, angle of response, and TGI measures ignore variance in the replicate experiments, depending only on the mean across replicates. These limitations often lead to an overoptimistic assessment of response.

Heterogeneity among PDX replicates can hamper the accurate evaluation of drug efficacy. Previous work (*1*, *5*) shows that substantial heterogeneity exists within in vivo PDX studies, with up to 33% of individual replicates' mRECIST classifications not matching the majority response classification in (*1*) and up to 50% in (*5*). To have a reliable estimate of treatment response, this heterogeneity needs to be taken into account by the response measure.

In this work, we sought to develop an approach to reliably estimate treatment response in in vivo drug screening experiments. We first show how multiple sources of variation lead to erroneous response calls. We then propose a new response measure, KuLGaP [based on Kullback-Leibler (KL) divergence between Gaussian processes (GPs)] to account for both experimental controls and variation among replicates. We tested and compared KuLGaP to four widely used response measures using 329 patient-derived xenograft (PDX) models and demonstrate that the robustness of KuLGaP allows for experimental designs with fewer animal replicates without noticeable loss in the accuracy of response quantification.

[1]Département AOTI, Université du Québec à Montréal, Montréal, QC H2X3X2, Canada. [2]Group for Research in Decision Analysis (GERAD), Montreal, QC H3T1J4, Canada. [3]Department of Computer Science, University of Toronto, Toronto, ON M5S2E4, Canada. [4]Vector Institute for Artificial Intelligence, Toronto, ON M5G1M1, Canada. [5]Hospital for Sick Children, Toronto, ON M5G1X8, Canada. [6]Princess Margaret Cancer Centre, University Health Network, Toronto, ON M5G1L7, Canada. [7]Department of Medical Biophysics, University of Toronto, Toronto, ON M5G1L7, Canada. [8]Pulmonology Department, Hôpitaux Universitaires de Strasbourg, 67200 Strasbourg, France. [9]Laboratory of Molecular Mechanisms of the Stress Response and Pathologies, INSERM U1113, 3 Avenue Molière, 67200 Strasbourg, France. [10]Crown Bioscience Taicang Inc., No.6 Beijing West Road, Taicang, Jiangsu 215400, P. R. China. [11]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON M5S1A8, Canada. [12]Department of Physiology, University of Toronto, Toronto, ON M5G1L7, Canada. [13]Department of Surgery, Toronto General Hospital, Toronto, ON M5G2C4, Canada. [14]Ontario Institute for Cancer Research, Toronto, ON M5G1L7, Canada. [15]CIFAR, Toronto, ON M5G1M1, Canada.
*Corresponding author. Email: bhaibeka@uhnresearch.ca (B.H.-K.); anna.goldenberg@utoronto.ca (A.G.)

## RESULTS

### KuLGaP, a measure for in vivo therapy response

To illustrate the benefits and pitfalls of various measures assessing drug responses using PDXs, we collected tumor growth curves from 329 PDX models across non–small cell lung carcinoma (NSCLC), colorectal, and breast cancers (table S1) and compared our KuLGaP growth response measure to other commonly used response measures: mRECIST (*1*), AUC (*6*), angle of response (Angle) (*2*), and TGI (*3*–*5*) (Fig. 1). There are two steps to computing KuLGaP. First, two GP models (*7*) are fitted to the PDX tumor growth curves, one for treated PDXs and another for controls. Second, the distance between these two GP models using KL divergence is computed (*8*). An overview of this process is captured in Fig. 1. The benefit of using the GP models is that they model not only the covariance of measurements across time (*7*) but also the variance within a group of replicates over time. The KL divergence that we use to compare treated replicates and controls is often used in machine learning and mathematical fields to measure the difference between two distributions, as KL has a strong theoretical foundation in information theory (*9*) and can be quickly computed for many distributions (*10*), including the Gaussian distribution.

We assessed the statistical significance of the distance between treatment and control arms by computing an empirical null distribution of distances between all pairs of controls in our dataset. Using this empirical distribution, we computed the significance (*P* value) of treatment response for each PDX model. Models with a *P* value less than 0.05 were considered to have a statistically significant distance and classified as responders. The critical value according to our empirical distribution was 7.97. We also calculated the critical values for the 0.1 and 0.01 confidence thresholds to be 5.61 and 13.9, respectively. The computation of KuLGaP is illustrated in Fig. 1 and is described in depth in Materials and Methods.

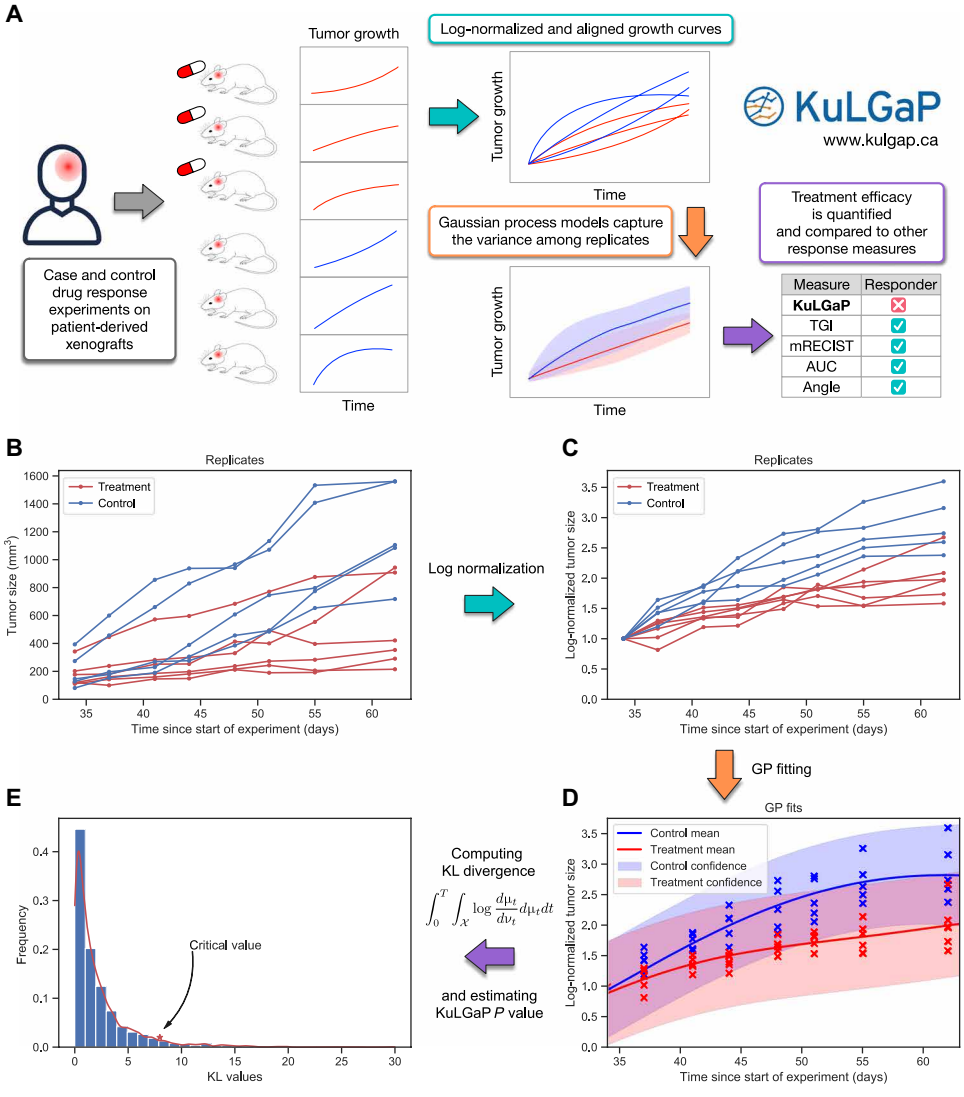


**Fig. 1. KuLGaP pipeline overview.** (**A**) The human tumor is implanted in a set of mouse replicates [patient-derived xenografts (PDX)]; some of which are treated with a given drug (cases; in red), and some are not (controls; in blue). In addition to KuLGaP, we also evaluated four different measures commonly used to assess response status in PDXs. (**B** and **C**) Tumor volumes of each replicate are normalized to the volume at the starting day of the treatment (here, it is day 34) and log-transformed. (**D**) A Gaussian process (GP) is fitted to the treatment and the control group of replicates. (**E**) KL divergence between the two GPs serves as a numerical estimator for how different the treatment and control groups are. We arrived at KuLGaP by computing a *P* value against the estimated null distribution of KL values.

### Comparison of measures for therapy response

We performed a comparative analysis of KuLGaP with mRECIST, AUC, Angle, and TGI. For each pair of measures, we computed the agreement between them as the percentage of experiments for which both measures gave the same classification (responder or nonresponder) (Fig. 2A), the false discovery rate of each measure with respect to each other (Fig. 2B), and the number of measures that classified an experiment as a responder compared to the KuLGaP statistic (Fig. 2C).

Out of the 329 experiments (with a total of 1437 treatment and 1946 control arms), KuLGaP classified 48 as responders (14.6%), compared to 133 (40.4%) for TGI, 187 (56.8%) for mRECIST, 199 (60.4%) for AUC, and 211 (64.1%) for Angle (data file S1). Briefly, TGI and Angle depend on the ratio of the difference between the first and last growth measurement of treatment and control; AUC calculates the cumulative difference at each measurement point between control and treatment groups; mRECIST categorizes observation of growth in the treatment arm into complete response (mCR), partial response (mPR), stable disease (mSD), and progressive disease (mPD). Following established practice (*11*–*13*), we considered all PDX

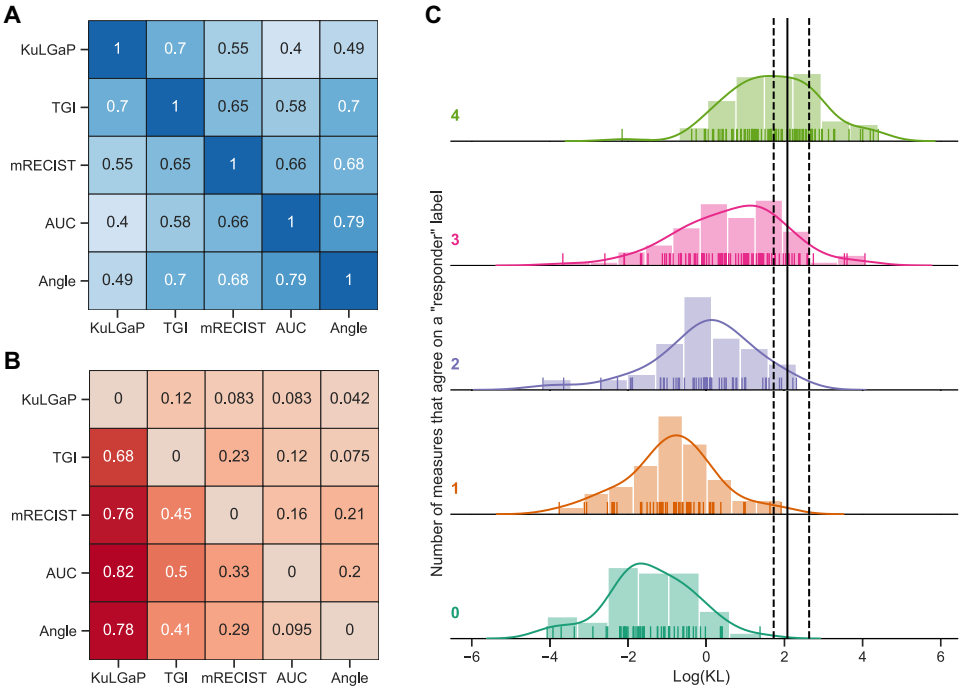


**Fig. 2. Comparison of classifications according to all five response measures.** (**A**) Heatmap of agreement (fraction of experiments where two measures agree) between the different measures across all models. (**B**) Proportion of responders according to the measure in the row that were not considered responders by the measure listed in the column. (**C**) Each row shows a histogram (distribution) of KuLGaP KL values across a group of experiments for which (top to bottom) (i) all baseline measures (TGI, mRECIST, AUC, and Angle) agreed on responder classification for each experiment in this group (top, green), (ii) three of four baseline measures agreed on responder classification for experiments in this group (magenta), and (iii to v) 2-0 (purple, orange, and topaz) baseline measures agreed on a responder classification, respectively. The solid vertical line indicates the KuLGaP's threshold for significance (calling an experiment a responder) at the 0.05 level, whereas the dashed lines indicate the 0.1 and 0.001 significance thresholds, respectively. All experiments to the right of the vertical line are responders according to KuLGaP, and all experiments to the left are nonresponders according to KuLGaP.

with a TGI value of more than 0.6 to be responders. The measures that gave the most similar results were AUC and Angle. Our KuLGaP measure yielded results that were most similar to TGI: The two classifications agreed on responders and nonresponders in 70% of all cases. Overall, KuLGaP was more conservative than all other response measures (Fig. 2B), as indicated by fewer responders called compared to other methods. For example, all but 4 experiments that were classified as responders by KuLGaP (92%) were also responders according to mRECIST, whereas only 31 of 147 mRECIST responders were also responders according to KuLGaP. Similarly, all but two of the KuLGaP responders were called as responders by Angle and AUC, but each of these measures called many KuLGaP-nonresponders as responders (145 for AUC and 165 for Angle). Figure 2C shows that there was substantial disagreement between the different measures. However, KuLGaP captured the majority of experiments for which there was a consensus among the other measures.

Further, we compared the continuous measures underlying TGI and KL classifications. We found that the Spearman rank correlation coefficient between TGI and the logarithm of the KL divergence was 0.69.

## Importance of the control group

We found that the information contained in the control replicates is crucial for an accurate response classification. A downside of the mRECIST classification is that it does not consider the control group but makes a classification based on the treatment group alone. The mRECIST criterion rates each treatment replicate as either mCR, mPR, mSD, or mPD, and then classifies the experiment by a majority vote of all the replicates where all but mPD ratings are considered a responder (*1*). We show an extreme example of how over-optimistic mRECIST would be if mSD were also considered a responder (fig. S1).

Consider the following two NSCLC PDX models (*14*): model 1 (Fig. 3, A to C) treated with afatinib and model 2 (Fig. 3, D to F) treated with erlotinib. The mRECIST framework reports mSD for all replicates of both models, resulting in a "responder" call for both models, whereas KuLGaP called a significant response for model 2 but not model 1. Because mRECIST does not take into account the control group, it missed the fact that the cancer in the untreated arm grew as fast as in the treated arm in model 1 but not in model 2. Therefore, the mRECIST classification was the same for both models, despite the clear differences in response. Both the AUC and Angle classifications agreed with the KuLGaP classification in both cases, whereas TGI classified both models as nonresponders. Because TGI does not consider the length of time for which the treated sample is not growing, it failed to detect the tumor arrest in model 2. In model 2, the treatment arm of this model took about 50 days longer to reach the maximum tumor size of its control replicates, and this effect was detected by our KuLGaP approach.

For model 1 (Fig. 3, A to C), we observed a particularly over-optimistic responder call by mRECIST. An intuitive way to alter the mRECIST classification to be more conservative is to consider only the mCR and mPR ratings as a positive response. However this leads to considerable loss of sensitivity, as demonstrated in fig. S2. The simple alteration cannot fix a fundamental mRECIST flaw.

Furthermore, we evaluated a colorectal cancer PDX with eight control and eight treatment replicates treated with evofosfamide (fig. S3). All measures apart from KuLGaP classified this model as a responder. The mRECIST measure failed to take into account that the treatment and control groups grew at a similar pace, whereas Angle and AUC only consider the last day of measurement and therefore missed the greater similarity of the treatment and control growth curves throughout the experiment. We provide an additional example using empirical data to support our claims in fig. S3 (D to F).

## Accounting for variance among replicates

Accounting for the variance among replicates leads to greater selectivity in declaring a response. An illustration of this scenario is given by the breast cancer PDX experiment with 15 paclitaxel-treated and

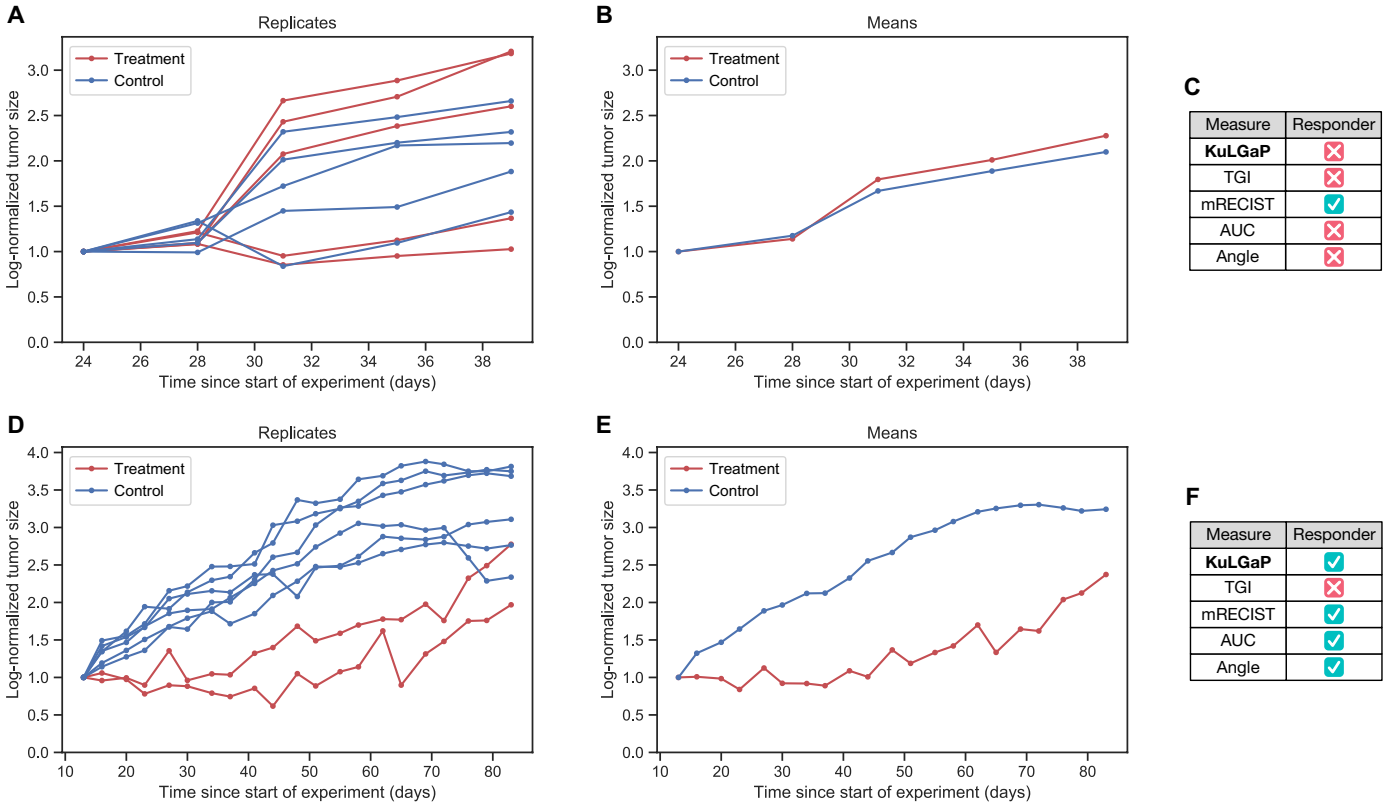


**Fig. 3. Importance of the control group.** (**A**) Log-normalized growth curves (afatinib treatment arm in red and control arm in blue) of an NSCLC PDX model (model 1) with five replicates in each arm (*14*). (**B**) Means across treatment and control replicates of model 1 from (A). (**C**) Classification of model 1 response to the treatment. (**D**) Log-normalized growth curves of another NSCLC PDX model (model 2) with two erlotinib treatment replicates and six controls (*14*). (**E** and **F**) Analogous to (B) and (C), respectively, but for NSCLC PDX model 2. The mRECIST measure identifies both models as responders, particularly as stable disease (mSD); KuLGaP identifies model 1 as a nonresponder and model 2 as a responder.

12 control replicates shown in Fig. 4 (A to C). Although there was a substantial difference in the means between the control and treatment groups (Fig. 4B), there was also substantial variance among replicates in each group (Fig. 4A). TGI, similarly to mRECIST, AUC, and Angle measures, classified this model as a responder. KuLGaP takes into account the variance among replicates and shows that the variance within control and treatment arms is big enough to remove the significance of the mean difference, thus classifying this model as a nonresponder.

Next, consider the following experiment, where 10 replicates of an NSCLC PDX model were treated with dacomitinib (Fig. 4, D to F). The Angle and AUC measures, which do not take into account variance, identified this PDX model as a responder. Our KuLGaP measure picked up on the fact that the variance among replicates in the treatment and control groups was larger than the mean difference between the two groups and therefore declared the experiment a nonresponder. In other words, incorporation of variance led to greater selectivity in declaring response. The TGI measure concurs with the KuLGaP assessment of a nonresponder. The mRECIST classification (which does not consider the control group) is stable disease (mSD), and thus, the model is erroneously considered responsive.

An additional example shows an experiment where even a large difference between the mean growth of the treatment and control arms can be deceptive (fig. S4). Upon closer inspection of the individual replicates, it is clear that any difference in the mean behavior is dwarfed by the large variance, leading to a false-positive call by all measures but KuLGaP.

## Implications of not considering multiple replicates in the study design

The experimental design of xenograft experiments usually requires the researcher to collect responses from multiple replicates of the model treated with the drug comparing them to those that are treatment naive (controls). Because PDX experiments are laborious, a 1 × 1 × 1 experimental design was proposed (*1*), where only a single replicate is used per drug and model. By testing and publishing a dataset on 1000 PDXs, the Novartis Institutes for Biomedical Research Patient-Derived Xenograft Encyclopedia (NIBR PDXE) study greatly contributed to research in this area. Unfortunately, this experimental design has its limitations. In this setup, the researchers were able to gain insight into the population-level response for a given drug. However, this design is not sufficient to draw conclusions for an individual patient/PDX level due to the absence of the variability that can only be derived from the replicates of the same PDX (*15*).

The lack of accounting for the variance in the 1 × 1 × 1 design is particularly detrimental for the mRECIST classification used in the study (*1*). It is common for different replicates to have different mRECIST classifications. An extreme case is given by an experiment

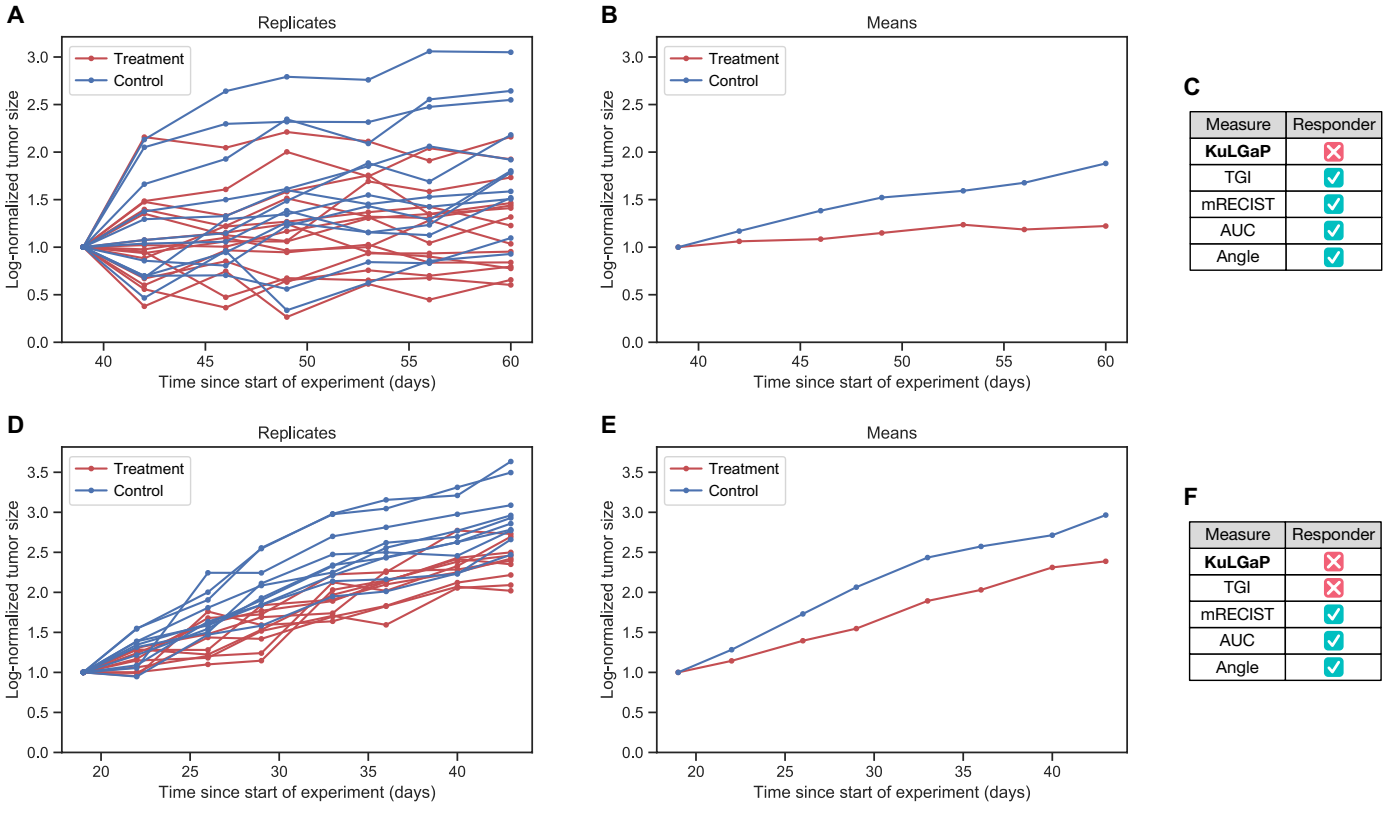


**Fig. 4. Importance of accounting for variance.** (**A**) Log-normalized tumor growth curves of a breast cancer PDX model (*31*) treated with paclitaxel; 15 treatment (in red) and 12 control (in blue) replicates. (**D**) Log-normalized growth curves of an NSCLC PDX model with 10 replicates treated with dacomitinib. (**B** and **E**) Mean treatment and control arm growth curves for each model (A and D), respectively. (**C** and **F**) Computed response classifications by all compared response measures for each model (A and D), respectively.

with five treatment replicates (fig. S5). Two of the five replicates were classified as mPR, two as mSD, and one as mPD. Depending on the one randomly chosen replicate in the $n = 1$ design, the classifications would have been different. This scenario is common because the mRECIST classification is often decided early on in the experiment when tumors are smaller and therefore more susceptible to measurement errors and noise. In our dataset, we found that fewer than 30% (97 of 329) of the models had the same mRECIST classification across replicates. Almost 60% (197 of 329) of the models had two different mRECIST classifications, of which 39 models (11% of the total) had mRECIST classifications that were not adjacent (such as mCR and mSD). In 10% (32 of 329) of the models, treatment replicates were assigned three different mRECIST classifications such that almost half (160 of 329) of the models had a majority decision supported by fewer than 75% of that model's replicates. Consequently, we postulate that the NIBR PDXE study using the $1 \times 1 \times 1$ design with mRECIST criterion is likely to be unreliable for personalized treatment prediction in many clinical scenarios.

## Assessing a study design with fewer replicates

There is a substantial downside to having only a single replicate per experiment. However, a large number of replicates increase both the cost and the use of research animals. We performed an experiment to see whether a smaller number of replicates would achieve reliable results. For each experiment, we randomly sampled without replacement three treatment and three control replicates and computed KuLGaP, mRECIST, Angle, AUC, and TGI classifications based on this subsample. This was repeated three times. Thus, for each model, we obtained three sets of experiments with three replicates each. By comparing the responses using only three replicates to those obtained using the full set of replicates, we were able to estimate how robust each response measure is to a reduced number of replicates. We found that KuLGaP and TGI measures were particularly robust to this form of subsampling, reaching agreements of 95.9 and 94.1% between reduced and original sets. The other measures were less robust, reaching 87.9% (mRECIST), 86.6% (Angle), and 79.9% (AUC). This suggests that it may be possible to reduce the number of replicates to 3 when studying drug response if necessary. We further found that one replicate was insufficient to estimate variance across the data (fig. S6). We have seen that good estimates for the inter-replicate variability are important. Obtaining such estimates can be done better with six or more replicates, and we therefore encourage the experimenters to continue PDX experiments with more replicates to maintain higher accuracy when possible.

## Clinical relevance of KuLGaP

We compared the cisplatin-vinorelbine combination treatment response in PDXs to data from 13 corresponding patients with NSCLC receiving adjuvant platinum-based chemotherapy. For each

of these patients, we considered both the time to recurrence and the growth curves of the corresponding PDX. The time to recurrence was measured from the time of starting adjuvant chemotherapy to either recurrence or last follow-up.

We found that among patients whose corresponding PDX models were classified as responders by KuLGaP, the mean time to recurrence was 4.13 years, compared to 1.02 years in the group of nonresponders according to KuLGaP. The difference (3.11 years) was the highest compared to all other methods (Table 1). We found substantial disagreement between the measures, which showed unanimity between responders in only four cases (three responders and one nonresponder).

Because of the small sample size, it was difficult to assess statistical significance of our clinical validation. However, the fact that there was a substantial difference in survival of the patients KuLGaP predicted as PDX responders compared to other methods is encouraging in terms of clinical relevance of our measure compared to all other currently used approaches.

## DISCUSSION

The problem of drug response prediction is incredibly important for the field of precision medicine, and yet, it is far from being solved, still fraught with many obstacles. PDXs are an appealing paradigm for drug response studies due to their ability to potentially act as a realistic simulation of a given patient and model the spectrum of clinical disease. Among their many applications, PDXs can be used both for predicting response for individual patients through empiric drug treatment and for identifying biomarker-response relationships across heterogeneous collections representing the patient populations. In each use case, efficient testing of many individual PDX models and drugs and accurate drug response quantification are of critical importance. In the former, false-positive or false-negative predictions have a major impact, as patients have a limited opportunity for treatment, and avoidance of ineffective toxic therapy is crucial. In the latter, accurate response calls are necessary to identify or validate predictive biomarkers that can be used to guide patient selection or companion diagnostic development in clinical trials. Our work shows that none of the currently widely used response quantification measures take into account the full extent of the available experimental data, some ignore controls, and others ignore variation among replicates. Grounded in real data and making the fewest assumptions about regularity of measurements and the number and variability of replicates, we derived a measure, KuLGaP, that provides a theoretically sound solution to this problem. We have shown KuLGaP to be more selective on a large set of PDXs and more concordant with patient outcomes than four commonly used measures in a small study.

Our exploration of real-world examples provides an insight into how we could improve other existing measures as well. For example, one way to make the mRECIST measure more selective would be to include mSD in the nonresponder category. Unfortunately, this leads to false-negative classifications: An extreme example of this is illustrated in fig. S2. The TGI measure is one of the widely used measures in the biomedical literature. Like the most commonly used measures, TGI is computed on the basis of the mean value of the replicates and then thresholded, especially in cases when the number of replicates is small and therefore fails to take into account the variation between replicates. As discussed above, this can have a substantial impact on the resulting classification. Moreover, the TGI criterion only takes the first and last measurements into account and is therefore highly susceptible to measurement errors and fluctuations in the tumor size at the specific time points. One way to introduce at least some impact of the variance would be to calculate TGI individually for each control-treatment replicate pair and apply a suitable statistical test. However, this approach would not work well in models with relatively few replicates per model, because this would lead to a low power in the statistical testing. To reduce the impact of a measurement error at the end of the experiment, one could calculate the TGI criterion based on a few of the measurement points and then take a consensus measure. Although it may result in an improvement, this solution will still suffer from not considering the variance across time points. Note that we have not compared KuLGaP to event-free survival or other metrics of durability of response. This is because KuLGaP, like AUC and TGI, is a classification measure of response to treatment and not comparable to survival metrics.

There are limitations to this study. In every analysis presented in this paper, the tumor growth curves were truncated at the time point the first mouse was sacrificed. Although this approach may lead to some loss of information, it is needed for consistent comparison of the response-calling methods. Overall, a median of 7 days were removed from the experiments and more than 40% of the growth curves required no truncation at all. If a mouse is sacrificed for reasons unrelated to the experiment and effect of the treatment, the experimenter may choose to remove it before applying the response-calling pipeline and avoid premature truncation. A second potential limitation is that unlike the methods that do not take variability among replicates into account and can thus be used in scenarios where there is only one replicate available, KuLGaP



**Table 1. Patient stratification by corresponding PDX response.** Mean time to relapse (in years) in the group of responders and nonresponders according to each measure. The number of patients is indicated in parentheses.

| Measure | Mean time to relapse in responders | Mean time to relapse in nonresponders | Difference (years) |
|---|---|---|---|
| KuLGaP | 4.13 (3) | 1.02 (10) | 3.11 |
| mRECIST | 2.62 (7) | 0.97 (6) | 1.65 |
| Angle | 2.14 (10) | 1.53 (3) | 0.61 |
| AUC | 2.02 (11) | 0.32 (2) | 1.70 |
| TGI | 2.22 (7) | 1.25 (6) | 0.97 |

computation requires at least three replicates per experiment. On the plus side, we have shown that we achieve similar performance with three replicates as with the standard currently widely used design (5 to 10), and thus, using KuLGaP may lead to more cost-effective experimentation in the future. Last, we note that KuLGaP measure is not directly comparable to survival metrics.

As our results on a reduced number of replicates show, KuLGaP still performs well when there are only three replicates in the treatment and control arms. Although we certainly recommend a larger number of replicates, we would not advise a smaller number of replicates than three so that the inter-replicate variation can still be reliably estimated. Overall, in our experience, there is no substitute for a measure that models all of the available data simultaneously, taking advantage of the multiple replicates for cases and controls; KuLGaP fulfills these criteria. In addition, KuLGaP can be used to compare the difference between any two treatment groups. For example, it could be applied to comparing combination and single-agent treatments to detect potential additive or synergistic effects in combination therapies. We expect that introducing such a measure will lead to more faithful predictions of clinical outcomes and biomarker-response relationships. We have thus created a simple-to-use web interface to assess the response for any PDX clinical experiments, kulgap.ca, that is equally easy to use for both clinicians, technicians, and biostatisticians, which we hope will result in wide uptake and reproducible results across drug response research.

## MATERIALS AND METHODS

### Study design

In this study, we introduced a measure of tumor response, KuLGaP, to account for. We tested and compared KuLGaP to four widely used response measures using 329 PDX models obtained from previously published studies of lung, breast, and colon cancers in Canada and a variety of cancers in China. Each PDX model experiment consisted of several replicates of mice that were assigned randomly to treatment and control groups. The number of replicates varied between 3 and 16 mice per group in each model. For the purposes of this study, each mouse is represented by its tumor growth curve.

### Data preparation

At each measured time point of an experiment, we took tumor volume estimated from the tumor dimensions as the observed treatment response. The first day of drug administration was designated as the initial point of the experiment, and we studied the growth curves from that point onward. Curves were truncated to end at the time point where the first mouse (control or treatment) was sacrificed, allowing for consistent comparison. Next, the growth curve of each PDX replicate in both treatment and control arms was log-normalized to tumor size at the starting day of the treatment. For all measures considered in this paper, the treatment response was then assessed from these truncated log-normalized curves.

Although truncation at the time where the first mouse was sacrificed can be a limiting factor, we note that our methodology can be extended to allow for a different truncation. For example, if the research team knows that a mouse was sacrificed for a reason that is unrelated to the tumor treatment, this mouse can be removed from the response assessment (as is commonly done in practice already). Furthermore, KuLGaP can be applied multiple times within the same experiment: at the time of the first mouse sacrifice, at the time of the second mouse sacrifice (without the first mouse), and so on. However, this latter approach should be performed with care for three reasons. First, if KuLGaP is applied repetitively to the same experiment, the researchers will have to manually correct for multiple hypotheses testing. This is possible to do even using our online implementation, as we provide a table with $P$ values in addition to the response/nonresponse calls. Second, we have shown that a too small number of replicates lead to an unsafe estimation of the variance, so we do not recommend applying it when the number of replicates drops below three. Third, there is a variance-bias trade-off that is in effect here: The fewer replicates used, the less confident the response call will be.

### KuLGaP

There are two steps to computing KuLGaP. In the first, two GP models (*7*) are fitted: one for tumor treated PDX and one for controls. In the second step, we compute a symmetrized integrated version of the KL divergence between the two GP models called KL divergence. KL is frequently used to compute the distance between two distributions. We assessed the significance of divergence between two models by computing KL divergences between all pairs of controls. Using this empirical distribution of divergences, we computed $P$ values of significance of response for each PDX model. Models with a $P$ value less than 0.05 that were considered to have a statistically significant KL divergence were classified as responders.

GPs provide two major benefits in tumor growth modeling over a simpler statistical model, such as the multivariate normal distribution. First, the tumor size measurements are not evenly spread out in time, that is, the time that has passed between two measurements is not constant. Second, the tumor sizes may not always be measured on the same set of days between various experiments. Although the measurements were mostly aligned between cases and controls for each experiment, it does not hold for many control-to-control comparisons we need to evaluate to compute the empirical null distribution of KL values, that is, to establish what differences between experiments are not statistically significant. A further advantage of using GPs is that our model can handle missing measurements and estimate the tumor size between measurements in a statistically sound manner. When adding these desiderata to a multivariate normal model, we naturally arrive at a GP. Before our work, GPs have been successfully used in other biomedical contexts, particularly for patient trajectory modeling and forecasting (*16*–*18*).

#### *Gaussian processes*

Recall that a set of random variables $X_1, \ldots, X_k$ is said to be jointly Gaussian with mean vector $\mu \in R^k$ and covariance matrix $\Sigma \in R^{k \times k}$ if the joint density of $X_1, \ldots, X_k$ is given by

$$f_{X_1,\ldots,X_k}(x_1,\ldots,x_k) = (2\pi)^{-k/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

A GP (*7*) on an interval $[0, T]$ with mean process $m : [0, T] \to R$ and covariance kernel $K : [0, T] \times [0, T] \to R$ can be considered as an infinite-dimensional analog of the joint Gaussian distribution and is formally defined as a random function $X : [0, T] \to R$ such that for any $0 < t_1 < \ldots < t_k < T$ the joint distribution of $(X(t_1), \ldots, X(t_1))$ is Gaussian with mean vector $(m(t_1), \ldots, m(t_k))$ and covariance matrix $\Sigma$, where $\Sigma_{i,j} = K(t_i, t_j)$ for all $i, j$.

Given a collection of measurements such as tumor sizes measured for each replicate in a PDX experiment, separately for treatment and control, and a prior GP, one can use Bayes' theorem to find the posterior distribution of the underlying tumor growth given the noisy observed data points [see also chapter 6.4 in (*19*)]. This was implemented using the GPy package (*20*) (http://github.com/SheffieldML/GPy). Because of its universality (*21*) and for theoretical reasons (*7*), the radial basis function was chosen as the prior distribution, with a variance of 1 and a length scale of 10. This choice for a prior kernel leads to good fits of the data for the posterior distribution. Hyperparameter selection was performed by maximizing the likelihood, using the Broyden-Fletcher-Goldfarb-Shannon algorithm provided by the package, with seven restarts for each model. The schematic for our data analysis pipeline is given in Fig. 1.

***KL divergence and KuLGaP***

The KL divergence (*8*) (also called relative entropy) between two probability measures $P$ and $Q$ on a set X is given by

$$D_{\mathrm{KL}}(P \| Q) = \int_X \log \frac{\mathrm{d}P}{\mathrm{d}Q}\,\mathrm{d}P$$

This is not symmetric, and it will be more convenient to work with the symmetrized version

$$D_{\mathrm{SKL}}(P, Q) = D_{\mathrm{KL}}(P \| Q) + D_{\mathrm{KL}}(Q \| P)$$

For two random processes, that is, sequences of probability measures $(\mu_t, \nu_t : t \in [0, T])$ indexed by a time interval, we define the integrated symmetrized KL divergence between them as

$$D_{\mathrm{ISKL}}(\mu, \nu) = \int_0^T D_{\mathrm{SKL}(\mu_t, \nu_t)}\,\mathrm{d}t$$

Consider now a particular PDX experiment with a given drug $D$, lasting a total of $T$ days. We proceed as follows: First, fit a GP each to the treatment and the control replicates and denote their distributions by $\mu^T = (\mu_t{}^T : t \in [0, T])$ and $\mu^C = (\mu_t{}^C : t \in [0, T])$, respectively, and compute the integrated KL divergence $D_{\mathrm{ISKL}}(\mu^T, \mu^C)$ between them. This quantity can be considered as a continuous estimate of the effect of drug $D$: The larger the KL divergence, the further away the treatment and control replicates are to one another, and therefore the larger an effect by drug $D$.

To test whether an observed KL value corresponded to a successful anticancer therapy, we tested the null hypothesis $H_0$ that the treatment and control GPs did not differ significantly against the alternative hypothesis $H_1$ that they did differ. We chose to estimate the distribution of a KL divergence under $H_0$ empirically as follows. Because each control group did not receive any treatment, it is reasonable to assume that there was no effect. Therefore, we estimated the null distribution by computing empirical distribution by calculating the KL divergence between any pair of control groups from the NSCLC and colorectal PDX. This discrete distribution was then smoothed using a Gaussian kernel with bandwidth 0.27, which was selected via leave-one-out cross-validation by the statsmodels Python module (*22*). Last, the KuLGaP measurement is calculated as the probability of obtaining a KL divergence value at least as large as the one obtained in the experiment (right tail probability/one-sided $P$ value). Specifically, we have calculated a critical value $K_{\mathrm{crit}}$ such that the probability of exceeding this value according to the null distribution was higher than a specified confidence level, in our case 0.05. Thus, an experiment was classified as a responder according to KuLGaP if and only if its KL divergence value was higher than $K_{\mathrm{crit}}$. The observed values and our estimate of the probability distribution are illustrated in fig. S7.

**Modified RECIST**

The RECIST (*23*) is a framework of guidelines for evaluation of tumor response to anticancer therapies, based on linear dimensions of tumor lesions. Four classifications are possible, from the best to the worst outcome: complete response (CR), partial response (PR), stable disease (SD), and progressive disease (PD). The modified RECIST (mRECIST) (*1*) allows the classification based on tumor volume growth curves.

For each time $t$, we determined the relative volume change of the tumor with respect to its reference size $V_0$, that is, we calculated $\Delta V_t = (V_t - V_0)/V_0$. The BestResponse is defined (*1*) to be the minimal value of $\Delta V_t$ for all times $t$ after 3 days. Further, the running average of $\Delta V_0$, $\Delta V_1$, …, $\Delta V_t$ is calculated. The minimal value of this running average is called (*1*) BestAvgResponse. The quantities BestResponse and BestAvgResponse are then used to obtain the mRECIST classification, using the following thresholds:

- BestResponse < −95% and BestAvgResponse < −40%: mCR (modified complete response);
- BestResponse < −50% and BestAvgResponse < −20%: mPR (modified partial response);
- BestResponse < 35% and BestAvgResponse < 30%: mSD (modified stable disease);
- BestResponse > 35% or BestAvgResponse > 30%: mPD (modified progressive disease).

Because the mRECIST criterion does not take into account the presence of multiple replicates, an mRECIST value is calculated for each replicate and a majority vote among replicate classifications is taken. Following (*1*), an mRECIST classification of mPD was considered as a nonresponder, while all others were considered as responders. It should be noted that, by definition, mRECIST is not able to take into account the evolution over time (because it only considers the smallest observation) or the variation between replicates.

**Area under the curve**

As in (*2*), the AUC under each replicate in the treatment and control groups was calculated. Then, $P$ values for group comparisons based on AUC were calculated using a one-tailed nonparametric Mann-Whitney test. A significance threshold of $P < 0.05$ was used to classify each PDX model as either a responder (significant difference) or a nonresponder (no significant difference).

**Response angle**

For each replicate in the treatment and control groups, the angle between the best fit according to ordinary least squares (OLS) regression of the normalized tumor curve and the line $y = 1$ was calculated. Then, the same statistical test as described for the AUC was applied to compare pairwise mean angles of response (*2*), yielding a classification of each PDX model as either a responder (significant difference) or a nonresponder (no significant difference).

**Tumor growth inhibition**

The TGI is computed as follows: $\mathrm{TGI} = 1 - \frac{\Delta y^T}{\Delta y^C} = \frac{\Delta y^C - \Delta y^T}{\Delta y^C}$, where $\Delta y^C$ and $\Delta y^T$ denote the mean difference between last and first measurement for the control and treatment groups, respectively (*5*).

Following established practice (*11*–*13*), we consider all PDX with a TGI value of more than 0.6 to be responders.

### Statistical analysis

For all tests, a significance threshold of 0.05 was used. For KuLGaP, a one-sided test against the KL null distribution was calculated. For AUC and response angle, a one-tailed nonparametric Mann-Whitney test was applied.

### Code Ocean

A dockerized capsule reproducing the full software environment is available on Code Ocean (https://doi.org/10.24433/CO.8958866.v2).

## SUPPLEMENTARY MATERIALS

www.science.org/doi/10.1126/scitranslmed.abf4969
Figs. S1 to S7
Table S1
Data file S1
Reference (*32*)

View/request a protocol for this paper from *Bio-protocol*.

## REFERENCES AND NOTES


1. H. Gao, J. M. Korn, S. Ferretti, J. E. Monahan, Y. Wang, M. Singh, C. Zhang, C. Schnell, G. Yang, Y. Zhang, O. A. Balbin, S. Barbe, H. Cai, F. Casey, S. Chatterjee, D. Y. Chiang, S. Chuai, S. M. Cogan, S. D. Collins, E. Dammassa, N. Ebel, M. Embry, J. Green, A. Kauffmann, C. Kowal, R. J. Leary, J. Lehar, Y. Liang, A. Loo, E. Lorenzana, E. R. McDonald III, M. E. McLaughlin, J. Merkin, R. Meyer, T. L. Naylor, M. Patawaran, A. Reddy, C. Röelli, D. A. Ruddy, F. Salangsang, F. Santacroce, A. P. Singh, Y. Tang, W. Tinetto, S. Tobler, R. Velazquez, K. Venkatesan, F. Von Arx, H. Q. Wang, Z. Wang, M. Wiesmann, D. Wyss, F. Xu, H. Bitter, P. Atadja, E. Lees, F. Hofmann, E. Li, N. Keen, R. Cozens, M. R. Jensen, N. K. Pryer, J. A. Williams, W. R. Sellers, High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nat. Med.* **21**, 1318–1325 (2015).
2. F. Duan, S. Simeone, R. Wu, J. Grady, I. Mandoiu, P. K. Srivastava, Area under the curve as a tool to measure kinetics of tumor growth in experimental animals. *J. Immunol. Methods* **382**, 224–228 (2012).
3. A. Bertotti, E. Papp, S. Jones, V. Adleff, V. Anagnostou, B. Lupo, M. Sausen, J. Phallen, C. A. Hruban, C. Tokheim, N. Niknafs, M. Nesselbush, K. Lytle, F. Sassi, F. Cottino, G. Migliardi, E. R. Zanella, D. Ribero, N. Russolillo, A. Mellano, A. Muratore, G. Paraluppi, M. Salizzoni, S. Marsoni, M. Kragh, J. Lantto, A. Cassingena, Q. K. Li, R. Karchin, R. Scharpf, A. Sartore-Bianchi, S. Siena, L. A. Diaz Jr., L. Trusolino, V. E. Velculescu, The genomic landscape of response to EGFR blockade in colorectal cancer. *Nature* **526**, 263–267 (2015).
4. Y.-M. M. Yao, G. P. Donoho, P. W. Iversen, Y. Zhang, R. D. Van Horn, A. Forest, R. D. Novosiadly, Y. W. Webster, P. Ebert, S. Bray, J. C. Ting, A. Aggarwal, J. R. Henry, R. V. Tiu, G. D. Plowman, S.-B. Peng, Mouse PDX trial suggests synergy of concurrent inhibition of RAF and EGFR in colorectal cancer with *BRAF* or *KRAS* mutations. *Clin. Cancer Res.* **23**, 5547–5560 (2017).
5. S. Guo, X. Jiang, B. Mao, Q.-X. Li, The design, analysis and application of mouse clinical trials in oncology drug development. *BMC Cancer* **19**, 718 (2019).
6. T. D. Laajala, M. Jumppanen, R. Huhtaniemi, V. Fey, A. Kaur, M. Knuuttila, E. Aho, R. Oksala, J. Westermarck, S. Mäkelä, M. Poutanen, T. Aittokallio, Optimized design and analysis of preclinical intervention studies in vivo. *Sci. Rep.* **6**, 30723 (2016).
7. C. E. Rasmussen, C. K. I. Williams, *Gaussian Processes for Machine Learning* (The MIT Press, 2005).
8. S. Kullback, R. A. Leibler, On information and sufficiency. *Ann. Math. Stat.* **22**, 79–86 (1951).
9. M. C. Pardo, I. Vajda, About distances of discrete distributions satisfying the data processing theorem of information theory. *IEEE Trans. Inf. Theory* **43**, 1288–1293 (1997).
10. K. P. Burnham, D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (Springer, 2002).
11. B. Hong, Y. Yang, S. Guo, S. Duoerkun, X. Deng, D. Chen, S. Yu, W. Qian, Q. Li, Q. Li, K. Gong, N. Zhang, Intra-tumour molecular heterogeneity of clear cell renal cell carcinoma reveals the diversity of the response to targeted therapies using patient-derived xenograft models. *Oncotarget* **8**, 49839–49850 (2017).
12. S. Guo, D. Chen, X. Huang, J. Cai, J.-P. Wery, Q.-X. Li, Cetuximab response in CRC patient-derived xenografts seems predicted by an expression based RAS pathway signature. *Oncotarget* **7**, 50575–50581 (2016).
13. S. Guo, B. Mao, H. Q. Li, Abstract 4534: Theory and methodology for the design and analysis of PDX mouse clinical trials. *Cancer Res.* **77**, 4534 (2017).
14. E. L. Stewart, C. Mascaux, N.-A. Pham, S. Sakashita, J. Sykes, L. Kim, N. Yanagawa, G. Allo, K. Ishizawa, D. Wang, C.-Q. Zhu, M. Li, C. Ng, N. Liu, M. Pintilie, P. Martin, T. John, I. Jurisica, N. B. Leighl, B. G. Neel, T. K. Waddell, F. A. Shepherd, G. Liu, M.-S. Tsao, Clinical utility of patient-derived xenografts to determine biomarkers of prognosis and map resistance pathways in EGFR-mutant lung adenocarcinoma. *J. Clin. Oncol.* **33**, 2472–2480 (2015).
15. C. Krepler, K. Sproesser, P. Brafford, M. Beqiri, B. Garman, M. Xiao, B. Shannan, A. Watters, M. Perego, G. Zhang, A. Vultur, X. Yin, Q. Liu, I. N. Anastopoulos, B. Wubbenhorst, M. A. Wilson, W. Xu, G. Karakousis, M. Feldman, X. Xu, R. Amaravadi, T. C. Gangadhar, D. E. Elder, L. E. Haydu, J. A. Wargo, M. A. Davies, Y. Lu, G. B. Mills, D. T. Frederick, M. Barzily-Rokni, K. T. Flaherty, D. S. Hoon, M. Guarino, J. J. Bennett, R. W. Ryan, N. J. Petrelli, C. L. Shields, M. Terai, T. Sato, A. E. Aplin, A. Roesch, D. Darr, S. Angus, R. Kumar, E. Halilovic, G. Caponigro, S. Jeay, J. Wuerthner, A. Walter, M. Ocker, M. B. Boxer, L. Schuchter, K. L. Nathanson, M. Herlyn, A comprehensive patient-derived xenograft collection representing the heterogeneity of melanoma. *Cell Rep.* **21**, 1953–1967 (2017).
16. A. M. Alaa, J. Yoon, S. Hu, M. van der Schaar, Personalized risk scoring for critical care prognosis using mixtures of gaussian processes. *IEEE Trans. Biomed. Eng.* **65**, 207–218 (2018).
17. J. Futoma, S. Hariharan, K. Heller, Learning to detect sepsis with a multitask Gaussian process RNN classifier. arXiv:1706.04152 [stat.ML] (13 June 2017).
18. Y. Xu, Y. Xu, S. Saria, A Bayesian nonparametric approach for estimating individualized treatment-response curves, *Mach. Learn. Healthc. Conf.*, 282–300 (2016).
19. C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer New York, 2006).
20. GPy, GPy: A Gaussian process framework in python (GitHub, since 2012); https://github.com/SheffieldML/GPy.
21. C. A. Micchelli, Y. Xu, H. Zhang, On translation invariant operators which preserve the B-spline recurrence. *Adv. Comput. Math.* **28**, 157–169 (2008).
22. S. Seabold, J. Perktold, Statsmodels: Econometric and statistical modeling with Python, *Proc. 9th Python Sci. Conf.* (2010).
23. P. Therasse, S. G. Arbuck, E. A. Eisenhauer, J. Wanders, R. S. Kaplan, L. Rubinstein, J. Verweij, M. Van Glabbeke, A. T. van Oosterom, M. C. Christian, S. G. Gwyther, New guidelines to evaluate the response to treatment in solid tumors. *J. Natl. Cancer Inst.* **92**, 205–216 (2000).
24. G. K. Sandve, A. Nekrutenko, J. Taylor, E. Hovig, Ten simple rules for reproducible computational research. *PLOS Comput. Biol.* **9**, e1003285 (2013).
25. R. Gentleman, Reproducible research: A bioinformatics case study. *Stat. Appl. Genet. Mol. Biol.* **4**, Article2 (2005).
26. D. F. Stroup, Meta-analysis of observational studies in epidemiology: A proposal for reporting. Meta–analysis of observational studies in epidemiology (MOOSE) group. *JAMA* **283**, 2008–2012 (2000).
27. J. Ortmann, C. Eeles, E. Tai, A. Goldenberg, B. Haibe-Kains, pyKuLGaP: A Python package for statistical analysis and plotting of patient derived xenograft (PDX) models of cancer (since 2020; doi: 10.5281/zenodo.5527205).
28. B. Haibe-Kains, J. Ortmann, E. Tai, L. Rampasek, A. S. Mer, R. Shi, E. L. Stewart, C. Mascaux, A. Fares, N.-A. Pham, S. Sakashita, J. Weiss, X. Jiang, G. Liu, D. Cescon, C. O'Brien, S. Guo, M.-S. Tsao, A. Goldenberg, KuLGaP: A selective measure for assessing therapy response in patient-derived xenografts (Code Ocean, 2020); https://codeocean.com/capsule/2817911/tree.
29. J. Ortmann, L. Rampášek, E. Tai, A. S. Mer, R. Shi, E. L. Stewart, C. Mascaux, A. Fares, N.-A. Pham, G. Beri, C. Eeles, D. Tkachuk, C. Ho, S. Sakashita, J. Weiss, X. Jiang, G. Liu, D. W. Cescon, C. O'Brien, S. Guo, M.-S. Tsao, B. Haibe-Kains, A. Goldenberg, Anonymized PDX growth curves from KuLGaP response classification analysis. *Zenodo* 10.5281/zenodo.4091802 , (2020).
30. A. S. Mer, W. Ba-Alawi, P. Smirnov, Y. X. Wang, B. Brew, J. Ortmann, M.-S. Tsao, D. W. Cescon, A. Goldenberg, B. Haibe-Kains, Integrative pharmacogenomics analysis of patient-derived xenografts. *Cancer Res.* **79**, 4539–4550 (2019).
31. I. Brana, N.-A. Pham, L. Kim, S. Sakashita, M. Li, C. Ng, Y. Wang, P. Loparco, R. Sierra, L. Wang, B. A. Clarke, B. G. Neel, L. L. Siu, M.-S. Tsao, Novel combinations of PI3K-mTOR inhibitors with dacomitinib or chemotherapy in PTEN-deficient patient-derived tumor xenografts. *Oncotarget* **8**, 84659–84670 (2017).
32. J. Haynes, T. D. McKee, A. Haller, Y. Wang, C. Leung, D. M. A. Gendoo, E. Lima-Fernandes, A. Kreso, R. Wolman, E. Szentgyorgyi, D. C. Vines, B. Haibe-Kains, B. G. Wouters, U. Metser, D. A. Jaffray, M. Smith, C. A. O'Brien, Administration of hypoxia-activated prodrug evofosfamide after conventional adjuvant therapy enhances therapeutic outcome and targets cancer-initiating cells in preclinical models of colorectal cancer. *Clin. Cancer Res.* **24**, 2116–2127 (2018).



**Acknowledgments**

**Funding:** J.O. was partially supported by a Horizon postdoctoral fellowship from the Concordia University. The lung PDX resource has been supported by grants held by M.-S.T. (Ontario Research Fund-Research Excellence grant 03-020, Canadian Cancer Society Research

Institute grant 701595, and Canadian Institutes of Health Research-Foundation grant 148395). N.-A.P. and the Princess Margaret Living Biobank are funded by The Princess Margaret Cancer Foundation. A.G., B.H.-K., E.T., and J.O. were supported by a CCSRI and CHRP grant to A.G., and B.H.-K. D.W.C., B.H.-K., and A.S.M. were supported by the Stand Up To Cancer Canada–Canadian Breast Cancer Foundation Breast Cancer Dream Team Research Funding, with supplemental support of the Ontario Institute for Cancer Research through funding provided by the Government of Ontario (funding award SU2C-AACR-DT-18-15). Stand Up To Cancer Canada is a program of the Entertainment Industry Foundation Canada. Research funding is administered by the American Association for Cancer Research International-Canada, the Scientific Partner of SU2C Canada. D.W.C. and B.H.-K. were supported by The Terry Fox Research Institute, and B.H.-K. was supported by the Gattuso Slaight Personalized Cancer Medicine Fund at Princess Margaret Cancer Centre, the Canadian Institute of Health Research, and the Natural Sciences and Engineering Research Council. A.G. was supported by the Canadian Institute of Health Research, the Canadian Cancer Society, Terry Fox Foundation, and the Natural Sciences and Engineering Research Council as well as CIFAR and the Amar Varma Family Chair in Biomedical Informatics and Artificial Intelligence. M.-S.T. was supported by CIHR grant FDN-148395. C.M. was supported by research grants from Boehringer Ingelheim Canada and from the Télévie (Fonds de la Recherche Scientifique–Fonds National de la Recherche Scientifique, Belgium). **Author contributions:** A.G. and B.H.-K. conceived the main concept and supervised the project. A.G. and J.O. developed KuLGaP. J.O. and E.T. wrote the source code. J.O. and L.R. performed and visualized the response-calling analyses. J.W. performed the clinical validation portion of the analysis based on the computed response calls. E.L.S., C.M., S.S., N.-A.P., A.F., G.L., and M.-S.T. provided and prepared the UHN dataset of NSCLC-derived PDXs. D.W.C. provided the breast cancer PDX dataset. C.A.O. provided the colorectal PDX dataset. S.G. and X.J. prepared the Crown Bioscience dataset and analyzed it with the response calling pipeline. C.E. packaged the code into the KuLGaP PyPI package. G.B., D.T., and C.H. designed and implemented the kulgap.ca website. J.O., L.R., A.G., and B.H.-K. wrote the manuscript, with contributions from A.S.M., G.L., D.W.C., C.A.O., S.G., and M.-S.T. All authors approved the manuscript. **Competing interests:** A.G. is a member of the 4YouandMe Advisory Board, with no compensation. E.L.S. is an employee of Pentavere Research Group Inc. B.H.-K. is a shareholder and paid consultant of Code Ocean. X.J. and S.G. were employees of Crown Bioscience Inc. at the time the study was performed and declare no other competing interests. D.W.C. reports consultancy and advisory fees from Agendia, AstraZeneca, Dynamo Therapeutics, Exact Sciences, Eisai, Gilead, GlaxoSmithKline, Merck, Novartis, Pfizer, Puma, and Roche; reports research funding to their institution from GlaxoSmithKline, Inivata, Merck, Pfizer, and Roche; is a member of a trial steering committee for Merck and GlaxoSmithKline; and holds a patent (US62/675,228) for methods of treating cancers characterized by a high expression level of spindle and kinetochore associated complex subunit 3 (ska3) gene. M.-S.T. has received honoraria from Amgen, Bayer, BMS, AstraZeneca, Daiichi-Sankyo, Sanofi-Regeneron, Lilly, Pfizer, Merck, Takeda, and Novartis and research grants from AstraZeneca, Bayer, Merck, Pfizer, Northern Biologics, and MedBiogene. G.L. is on the advisory boards of Pfizer, Merck, AbbVie, Takeda, AstraZeneca, Hoffman La Roche, Takeda, Bristol Myers Squibb, Eli Lilly, and Novartis and holds grants from EMD Serono, Boehringer Ingelheim, AstraZeneca, and Takeda. A.F. was a clinical research fellow at Princess Margaret Cancer Centre at the time the study was performed. Since then, A.F. has worked as consulting for Merck Sharp & Dohme pharma company and has also received honoraria from AstraZeneca and Pfizer. The other authors declare that they have no competing interests. **Data and materials availability:** All data associated with this study are present in the paper or the Supplementary Materials. Our study complies with the guidelines outlined in (*24*–*26*). Our code and documentation are open source and publicly available through the pyKulGaP Zenodo repository (*27*). A detailed tutorial describing how to run our pipeline and reproduce our analysis results is available in the Zenodo repository. A dockerized capsule reproducing the full software environment is available on Code Ocean (*28*) (https://doi.org/10.24433/CO.8958866.v2). All data generated during this study are included in this article or the Supplementary Materials. The raw datasets from UHN including NSCLC, breast, and colorectal cancer PDXs and Crown Bioscience Inc. dataset analyzed in this study are available in the Zenodo repository (*29*) (https://doi.org/10.5281/zenodo.4091802). All these data are available in the form of XevaSet objects as part of the Xeva BioConductor package (*30*).

Submitted 11 November 2020
Resubmitted 19 February 2021
Accepted 14 October 2021
Published 17 November 2021
10.1126/scitranslmed.abf4969