Pharmacological Research 188 (2023) 106671

Contents lists available at ScienceDirect

Pharmacological Research

journal homepage: www.elsevier.com/locate/yphrs

Large inherent variability in data derived from highly standardised cell culture experiments

Ian G. Reddin [a,b], Tim R. Fenton [a,b], Mark N. Wass [a,*], Martin Michaelis [a,*]

[a] School of Biosciences, University of Kent, Canterbury, UK
[b] Cancer Sciences, Faculty of Medicine, University of Southampton, Southampton, UK

ARTICLE INFO

Keywords:
Data reproducibility
Replicability
Cancer cell line
Screen
NCI60
Anti-cancer drugs
Chemotherapy
Attrition
Drug discovery
Drug development

ABSTRACT

Cancer drug development is hindered by high clinical attrition rates, which are blamed on weak predictive power by preclinical models and limited replicability of preclinical findings. However, the technically feasible level of replicability remains unknown. To fill this gap, we conducted an analysis of data from the NCI60 cancer cell line screen (2.8 million compound/cell line experiments), which is to our knowledge the largest depository of experiments that have been repeatedly performed over decades. The findings revealed profound intra-laboratory data variability, although all experiments were executed following highly standardised protocols that avoid all known confounders of data quality. All compound/ cell line combinations with > 100 independent biological replicates displayed maximum GI50 (50% growth inhibition) fold changes (highest/ lowest GI50) > 5% and 70.5% displayed maximum fold changes > 1000. The highest maximum fold change was 3.16×10^{10} (lowest GI50: 7.93×10^{-10} μM, highest GI50: 25.0 μM). FDA-approved drugs and experimental agents displayed similar variation. Variability remained high after outlier removal, when only considering experiments that tested drugs at the same concentration range, and when only considering NCI60-provided quality-controlled data. In conclusion, high variability is an intrinsic feature of anti-cancer drug testing, even among standardised experiments in a world-leading research environment. Awareness of this inherent variability will support realistic data interpretation and inspire research to improve data robustness. Further research will have to show whether the inclusion of a wider variety of model systems, such as animal and/ or patient-derived models, may improve data robustness.

1. Introduction

Cancer drug development is affected by large attrition rates. Only about 5% of agents that enter phase I cancer trials are eventually approved as anti-cancer drugs [1–3]. A lack of predictive power by preclinical models for cancer drug development has been blamed for these low success rates, which has been suggested to be at least in part caused by the limited replicability of findings in such systems [2,4–9].

In this context, the 'Reproducibility Project: Cancer Biology' most recently reported its findings on the replication of 50 experiments from 23 highly influential preclinical cancer studies published between 2010 and 2012, which resulted, according to the assessments of the authors, in the successful replication of only five of the investigated studies [3, 10–13].

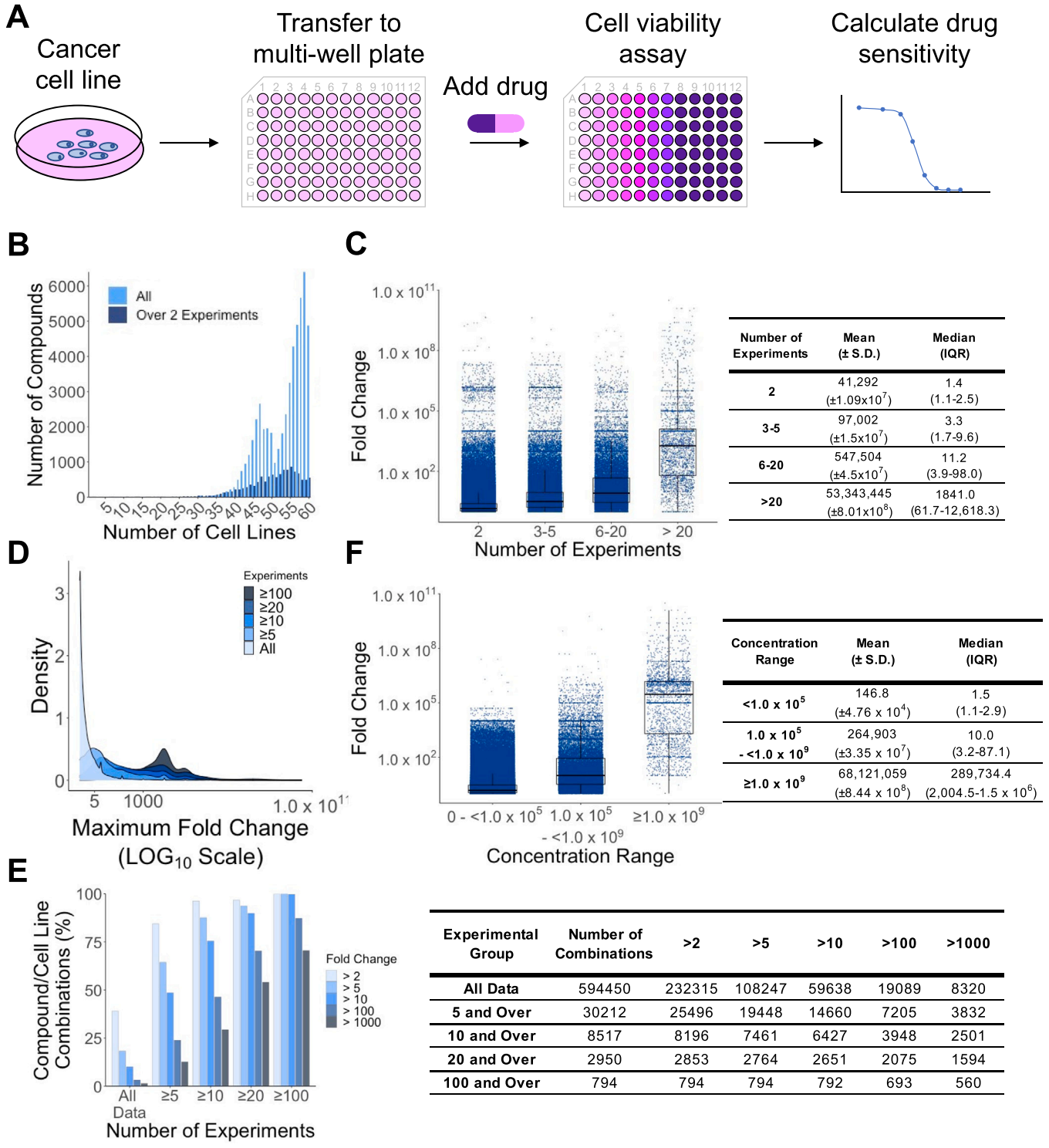
Such findings fit well into the 'reproducibility crisis' narrative in cancer research that commonly considers poor or even questionable research practices such as a lack of thoroughness, poor study design, biased reporting, and insufficient documentation of study detail as drivers of limited replicability and calls for higher research standards and more experimental standardisation [3–6,13–18].

Despite this strong narrative, the actual scale of the reproducibility crisis remains unclear and evidence largely anecdotal [19]. Perceptions are predominantly based on researcher views expressed in survey responses [6,20,21,22] and on findings communicated as Comments or Correspondence without detailed experimental information [4,5]. Moreover, generally accepted criteria defining successful replication attempts are missing [19,22,23,24]. For example, some authors of reports that were considered not successfully replicated by the 'Reproducibility Project: Cancer Biology' claimed that their findings had been independently confirmed by other groups in the meantime and had resulted in clinical drug candidates currently undergoing clinical testing [23].

* Corresponding authors.
E-mail addresses: m.n.wass@kent.ac.uk (M.N. Wass), m.michaelis@kent.ac.uk (M. Michaelis).

https://doi.org/10.1016/j.phrs.2023.106671
Received 23 August 2022; Received in revised form 12 January 2023; Accepted 17 January 2023
Available online 18 January 2023

Some findings have suggested a high level of data variability to be an inherent feature of complex biological systems [25–28]. However, few studies have directly addressed this issue. Replicability datasets are typically small and investigate complex animal models, including behavioural studies, or complex (at the time of the investigation) novel technologies, such as array platforms [26–30]. The complexity of animal studies is naturally high and the factors affecting the outcome are very difficult to disentangle. Moreover, it may not be too much of a surprise that novel technologies investigating highly complex datasets like array-based platforms may at least initially produce data of a



Number of Experiments	Mean (± S.D.)	Median (IQR)
2	41,292 (±1.09x10^7)	1.4 (1.1-2.5)
3-5	97,002 (±1.5x10^7)	3.3 (1.7-9.6)
6-20	547,504 (±4.5x10^7)	11.2 (3.9-98.0)
>20	53,343,445 (±8.01x10^8)	1841.0 (61.7-12,618.3)

Concentration Range	Mean (± S.D.)	Median (IQR)
<1.0 x 10^5	146.8 (±4.76 x 10^4)	1.5 (1.1-2.9)
1.0 x 10^5 - <1.0 x 10^9	264,903 (±3.35 x 10^7)	10.0 (3.2-87.1)
≥1.0 x 10^9	68,121,059 (±8.44 x 10^8)	289,734.4 (2,004.5-1.5 x 10^6)

Experimental Group	Number of Combinations	>2	>5	>10	>100	>1000
All Data	594450	232315	108247	59638	19089	8320
5 and Over	30212	25496	19448	14660	7205	3832
10 and Over	8517	8196	7461	6427	3948	2501
20 and Over	2950	2853	2764	2651	2075	1594
100 and Over	794	794	794	792	693	560

Fig. 1. Variability in NCI60 GI50 data. A) Overview of the principle of the NCI 60 screen. B) Compound/ cell line combinations with two or more experiments in the NCI60 database. C) GI50 fold changes in dependence on the number of experiments per compound/ cell line combination. Numerical data are presented in the adjacent table. D) Distribution of maximum GI50 fold changes illustrated by density plots for experimental compound/cell line combination groups with an increasing minimum number of experiments. E) Percentage of compound cell line combinations with maximum fold changes above the indicated thresholds in dependence of the number of experiments. Numerical data are presented in the adjacent table. F) Distribution of GI50 fold changes in dependence of the concentration ranges in which compounds were tested. Numerical data are presented in the adjacent table.

considerable variability [3,19]. Hence, larger datasets obtained with simpler assays over a long time period are likely to provide a better starting point for determining the inherent variability of biological data.

Therefore, here we analysed the variation in drug response data from the NCI60 screen [31] to develop a realistic understanding of the replicability of standardised assays. The NCI60 screen was selected, because it provides a dataset of unprecedented richness and depth. Since 1989, the NCI60 screen has tested the anti-cancer activity of thousands of compounds multiple times in a 60-cell line panel following strict standard operating procedures (Fig. 1A) [31–37].

Moreover, the lead scientists of the NCI60 screen were well aware of the caveats associated with cell lines and cell line-based experiments including the potential impact of cell line misauthentication, contamination (e.g. mycoplasma), phenotypic drift, cell culture media, culture vessels, and other reagents [31,38–41]. Strict quality measures were in place regarding the sourcing of reagents and materials, cell identity, absence of contamination, cell numbers used for assays, and the avoidance of phenotypic drift by using cells within a defined window of 30 passages [38]. Moreover, internal controls were used, in particular doxorubicin [42], and all data were analysed by exactly the same approach. Thus, the NCI60 database provides an unprecedented wealth of data on the replicability of drug testing in cancer cell lines using highly standardised procedures in a world-leading environment by highly skilled experts and, thus, a unique opportunity to establish an initial understanding of the inevitable intrinsic variability of biological data.

2. Results

2.1. NCI60 drug response data are characterised by a high level of variability

All drug sensitivity data derived from NCI60 testing are made available via Cell Miner [32–34,36,37]. In total, 52,585 compounds were tested in the NCI60 resulting in 2.8×10^6 compound/cell line combinations. Two or more (up to 2286) experiments were carried out for 11,841 compounds and 594,450 compound/cell line combinations (Fig. 1B, Suppl. Table 1; Suppl. Table 2). More than 100 experiments in at least one cell line were performed for 18 compounds and more than 1000 experiments for two compounds (Suppl. Table 3). Concentration ranges covered by the dose-response curves varied from $10^{1.2}$ to $10^{12.1}$. 612 compounds were screened with multiple concentration ranges, and the most common concentration range was 10^4 (11,213/ 94.7% of the compounds), representing the standard testing range using five 10-fold dilution steps (Suppl. Table 4).

The maximum fold change between the lowest and highest GI50 concentration (reduces cell viability by 50%) was detected for cyanomorpholinodoxorubicin in the colorectal cancer cell line COLO 205 (3.16×10^{10}; lowest GI50: 7.93×10^{-10} μM, highest GI50: 25.0 μM) (Suppl. Fig. 1 A, Suppl. Table 5). 232,315 (39.1%) drug/cell line combinations displayed maximum fold changes > 2, 108,247 (18.2%) drug/ cell line combinations fold changes > 5, 59,638 (10%) drug/cell line combinations fold changes > 10, 19,089 (3.2%) drug/cell line combinations > 100, and 8320 (1.4%) drug/cell line combinations > 1000 (Suppl. Table 5). The mean and median maximum GI50 fold changes for all compound/cell line combinations were 318,410 (standard deviation (SD) = 5.71×10^7) and 1.6 (interquartile range (IQR) = 1.1–3.4), respectively (Suppl. Table 6). The low median fold change reflects the large number of experiments that were only performed twice. Only two experiments were performed for 99.9% (361,872) experiments of the 362,135 compound/ cell line combinations (60.9% of the total 594,450 of compound/cell line combinations) that displayed maximum GI50 fold changes of less than two. When we only considered experiments that were repeated more often, the median GI50 values increased considerably, as indicated below.

2.2. Variability increases with the number of experiments

The percentage of compound/cell line combinations with high maximum fold change strongly increased with the number of experiments (Fig. 1 C, Suppl. Fig. 2A, Suppl. Fig. 2B, Suppl. Table 6). The mean and median GI50 fold changes increased from 41,292 and 1.4 for compound/cell line combinations with two experiments to 53,343,445 and 1841 for compound/cell line combinations with > 20 experiments (Fig. 1C, Fig. 1D, Suppl. Table 6).

When we considered compound/cell line combinations with a minimum of five experiments, the mean and median GI50 fold change for all combinations was 5.32×10^6 (SD = 2.48×10^8) and 10 (IQR = 3.1–98.7) (Table 6). 25,496 (84.4%) of 30,212 compound/cell line combinations displayed maximum fold changes > 2 and 3832 (12.7%) compound/cell line combinations > 1000. For compound/ cell line combinations with > 100 experiments, all 794 compound/ cell line combinations displayed a maximum fold change > 5% and 70.5% (560 out of 794) displayed a maximum fold change > 1000 (Fig. 1E, Suppl. Table 7).

Taken together, maximum GI50 fold changes increase with the number of experiments. In agreement, a significant correlation was detected between maximum GI50 fold changes and the number of experiments per compound/cell line combination (Spearman correlation coefficient = 0.34, $p < 2.2 \times 10^{-16}$) (Suppl. Fig. 3A).

2.3. Variability increases with the concentration range covered

The observed fold changes also reflected the tested concentration ranges per compound/ cell line combination in addition to the number of experiments, i.e. the broader the range of concentrations that were tested, the larger was the maximum fold change (Fig. 1D, Suppl. Table 8). A positive correlation was observed between concentration range and maximum fold change for all compound data (Spearman correlation coefficient = 0.31, $p < 2.2 \times 10^{-16}$) (Suppl. Fig. 3B).

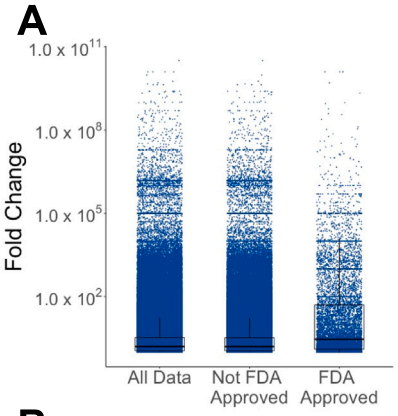
The mean and median GI50 fold changes for compound/ cell line combinations for which a maximum concentration range $< 1.0 \times 10^5$ was covered were 146.8 and 1.5, which increased to 68,121,059 and 289,734 for those with a concentration range of $\geq 1.0 \times 10^9$ (Fig. 1F, Suppl. Table 8).

To further investigate whether a larger concentration range results in larger GI50 fold changes, we considered four FDA-approved drugs (doxorubicin, fluorouracil, cisplatin, vinblastine) with at least 100 experiments performed on more than 20 cell lines. However, the vast majority of experiments were performed using the same concentration range for fluorouracil (99.5% of experiments), doxorubicin (99.3%), and vinblastine (91.2%). Only cisplatin was tested more frequently (21% of experiments) with different concentration ranges. Hence, we used the cisplatin dataset for further analyses.

Cisplatin had been tested more than 100 times in 24 cell lines. We then performed 100 experiments, in which we randomly selected 100 GI50 values for each cisplatin/ cell line concentration in the most commonly used concentration range (0.05–500 μM) and calculated the maximum GI50 fold changes. Then, we performed a further random 100 selections, but this time including all available concentration ranges. This was repeated 1000 times and the median maximum GI50 fold change for a compound/cell line combination was calculated. In 23 of the 24 cisplatin/ cell line combinations, the median GI50 fold change was significantly higher in the 100 random samples across all tested concentration ranges, than across 100 random samples from just one fixed concentration range (Suppl. Fig. 3E). This adds further evidence that the data variability increases when the covered concentration range increases.

2.4. Variability in FDA approved drugs

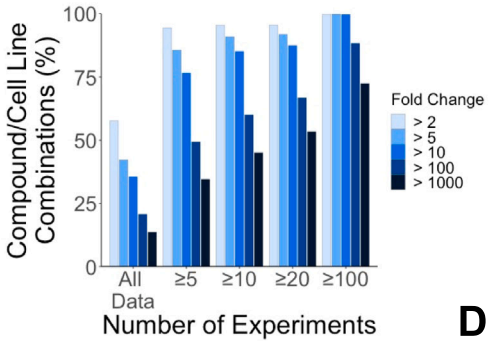
Since reliable clinical therapy outcomes depend on reproducible



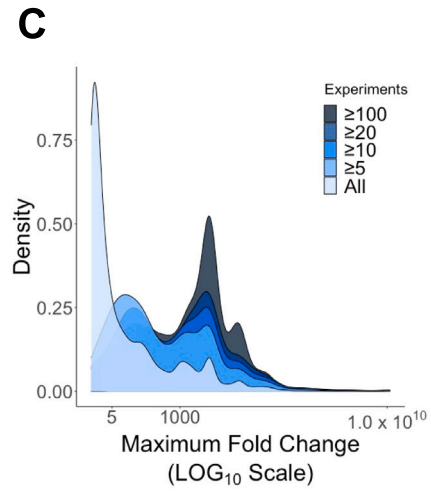
FDA Status	Mean (± S.D.)	Median (IQR)
All Data	318,410 ($\pm 5.71 \times 10^7$)	1.6 (1.1-3.4)
Not FDA Approved	218,154 ($\pm 4.68 \times 10^7$)	1.6 (1.1-3.3)
FDA Approved	6,195,807 ($\pm 2.57 \times 10^8$)	2.9 (1.3-50.0)

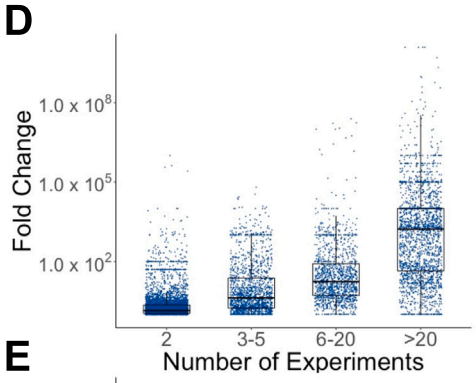
	Percentage of Total Combinations		
Number of Experiments	All Data	Not FDA	FDA
2	81.6	82.1	49.9
3-5	15.0	15.0	12.6
6-20	2.9	2.7	15.8
>20	0.5	0.1	21.7

	Percentage of Total Combinations		
Concentration Range	All Data	Not FDA	FDA
$<1.0 \times 10^5$	94.3	94.6	81.6
1.0×10^5 to $<1.0 \times 10^9$	5.2	5.0	15.5
$>1.0 \times 10^9$	0.4	0.4	2.8

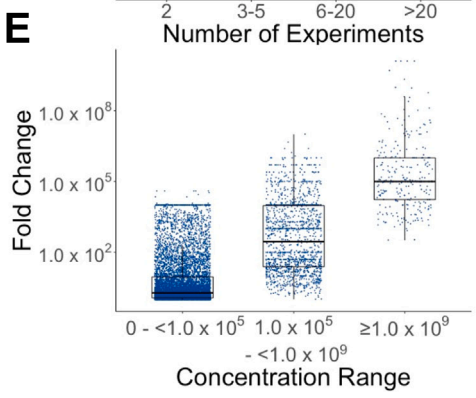


Experimental Group	Number of Combinations	>2	>5	>10	>100	>1000
All Data	9971	5752	4213	3547	2065	1361
5 and Over	3745	3533	3206	2868	1849	1295
10 and Over	2718	2594	2470	2312	1633	1225
20 and Over	2187	2089	2008	1911	1461	1168
100 and Over	751	750	750	749	663	544





Number of Experiments	Mean (± S.D.)	Median (IQR)
2	433.4 (±16,619)	1.4 (1.1-2.3)
3-5	319.8 (±2,337.0)	4.2 (1.7-23.7)
6-20	99,189.1 ($\pm 1.14 \times 10^6$)	17.86 (5.2-93.0)
>20	2.84×10^7 ($\pm 5.50 \times 10^8$)	1,577.6 (43.4-10,000.0)

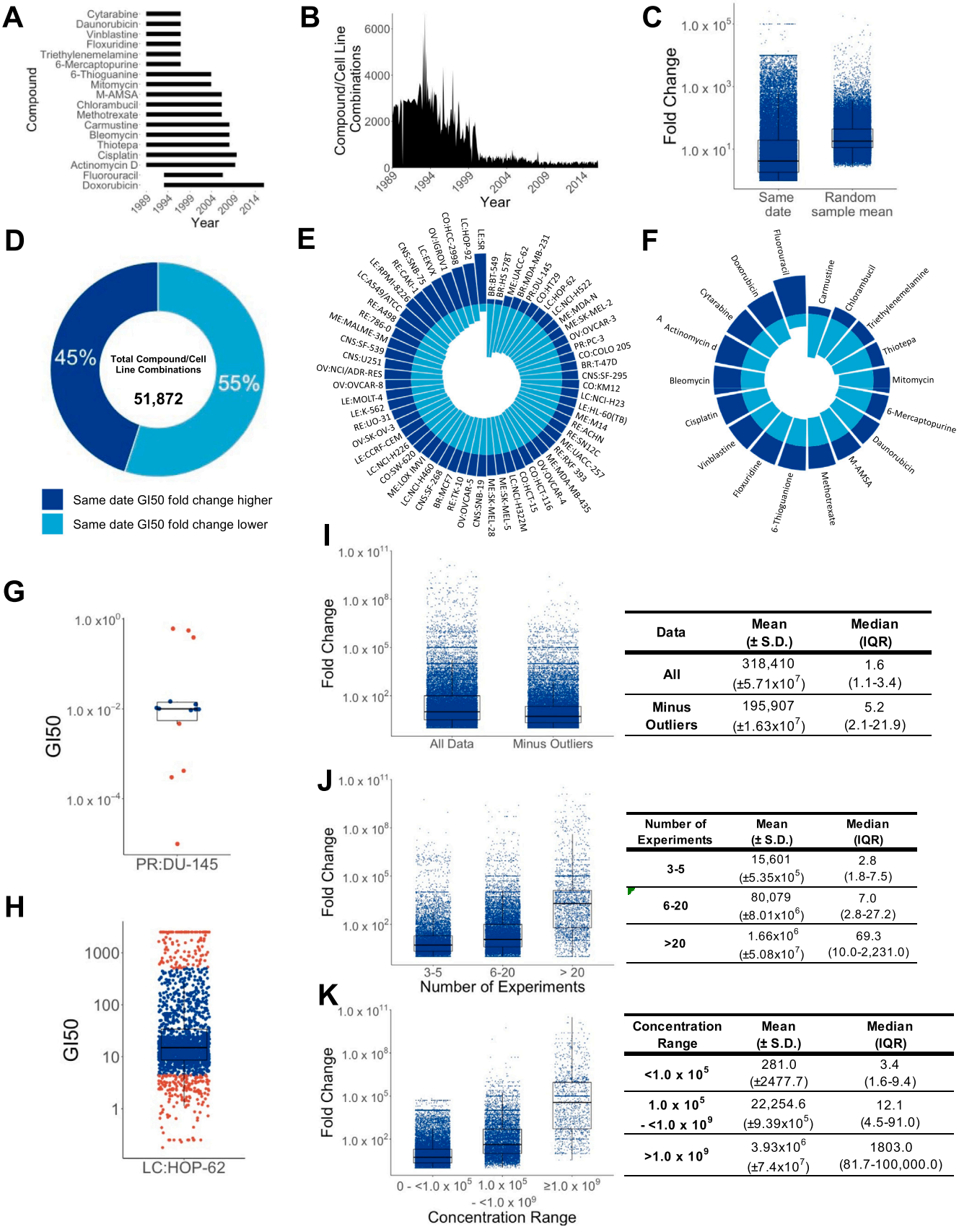


Concentration Range	Mean (± S.D.)	Median (IQR)
$<1.0 \times 10^5$	378.6 (±1,922.2)	1.9 (1.2-9.1)
1.0×10^5 to $<1.0 \times 10^9$	44,413.1 (±321,105.9)	284.5 (24.5-9,418.9)
$\geq 1.0 \times 10^9$	2.19×10^8 ($\pm 1.51 \times 10^9$)	100,000 ($17{,}002$-1.0×10^6)

Fig. 2. GI50 variation for FDA-approved drugs. A) Compound/ cell line combinations with 2 or more experiments in the NCI60 database. Numerical data are presented in the adjacent tables. B) Percentage of FDA-approved drug/ cell line combinations with maximum fold changes above the indicated thresholds in dependence of the number of experiments. Numerical data are presented in the adjacent table. C) Distribution of maximum GI50 fold changes illustrated by density plots for experimental compound/cell line combination groups with increasing minimum numbers of experiments. D) GI50 fold changes in dependence on the number of experiments per compound/ cell line combination. Numerical data are presented in the adjacent table. E) Distribution of GI50 fold changes in dependence of the concentration ranges in which compounds were tested. Numerical data are presented in the adjacent table.

drug effects, it may be speculated that FDA-approved drugs are more robust in their drug response data than experimental agents. However, the drug response data observed for FDA-approved drugs displayed a similar variability to that observed across all tested compounds.

The NCI60 database contained data on 181 FDA-approved drugs, which had been tested at least twice, resulting in 399,686 experiments investigating 9970 individual drug/cell line combinations (Suppl. Table 1). The number of experiments for drug/cell line combinations



Data	Mean (± S.D.)	Median (IQR)
All	318,410 ($\pm 5.71 \times 10^7$)	1.6 (1.1-3.4)
Minus Outliers	195,907 ($\pm 1.63 \times 10^7$)	5.2 (2.1-21.9)

Number of Experiments	Mean (± S.D.)	Median (IQR)
3-5	15,601 ($\pm 5.35 \times 10^5$)	2.8 (1.8-7.5)
6-20	80,079 ($\pm 8.01 \times 10^6$)	7.0 (2.8-27.2)
>20	1.66×10^6 ($\pm 5.08 \times 10^7$)	69.3 (10.0-2,231.0)

Concentration Range	Mean (± S.D.)	Median (IQR)
$<1.0 \times 10^5$	281.0 ($\pm$2477.7)	3.4 (1.6-9.4)
1.0×10^5 - $<1.0 \times 10^9$	22,254.6 ($\pm 9.39 \times 10^5$)	12.1 (4.5-91.0)
$>1.0 \times 10^9$	3.93×10^6 ($\pm 7.4 \times 10^7$)	1803.0 (81.7-100,000.0)

(caption on next page)

Fig. 3. GI50 variability is high between compound/cell line combination experiments on the same date and is not caused by outliers. A) Time periods of drug testing for individual drugs. B) Testing of individual compound/cell line combinations by date. C) Maximum GI50 fold changes in experiments testing compound/ cell line combinations on the same date compared to maximum GI50 fold changes in 1000 random controls of the same sample size. D) Percentage of cases in which same date experiments had a higher fold change than control samples randomly picked across the timeline. E) Proportion of same date GI50 fold changes in compound/ cell line combinations that are higher or lower than random control samples per cell line. F) Proportion of same date GI50 fold changes in compound/ cell line combinations that are higher or lower than random control samples per drug. G) GI50 value distribution for maytansine in the prostate cancer cell line DU-145 (outliers indicated in red). H) GI50 value distribution for 5-fluorouracil in the lung cancer cell line HOP-62 (outliers indicated in red). I) Comparison of maximum GI50 fold changes before and after removal of outliers. Numerical values are presented in the adjacent table. J) Maximum GI50 fold changes increase with experiment number after removal of outliers. Numerical values are presented in the adjacent table. K) Maximum GI50 fold changes increase with the concentration range covered after removal of outliers. Numerical values are presented in the adjacent table.

ranged from 2 to 2286.

The maximum GI50 fold change was 1.25×10^{10} observed for mithramycin in four cell lines, the colorectal cancer cell line COLO-205 (26 experiments; lowest GI50: 1×10^{-7} μM, highest GI50: 1250 μM) (Suppl. Fig. 1B), the CNS cell lines SF-295 (24 experiments; lowest GI50: 1×10^{-7} μM, highest GI50: 1250 μM) and U251MG (27 experiments; lowest GI50: 1×10^{-7} μM, highest GI50: 1250 μM), and the ovarian cancer cell line IGROV1 (26 experiments; lowest GI50: 1×10^{-7} μM, highest GI50: 1250 μM). Mithramycin, a member of the aureolic acid family was approved in 1970 but only temporarily used for testicular carcinoma and other types of cancer due to serious side effects [43]. The second highest GI50 fold change (7.28×10^{7}) was detected for paclitaxel, a stabilising tubulin-binding agent and one of the most commonly used anti-cancer drugs [44], in MDA-MB-435 (lowest GI50: 1×10^{-7} μM, highest GI50: 7.28 μM) (Suppl. Fig. 1 C), which had originally been assumed to be a breast cancer cell line, but was later found to be derived from the melanoma cell line M14 [45]. The mean and median GI50 fold changes for all FDA approved drugs were 6.20×10^{6} (SD = 2.57×10^{8}) and 2.9 (IQR = 1.27–50) (Suppl. Table 9).

The maximum, mean, and median GI50 fold changes were higher among the FDA approved drugs than for the non-FDA approved compounds (Fig. 2A, Suppl. Table 9), probably because they were tested in more experiments and at bigger concentration ranges (Fig. 2A).

When we considered the percentage of FDA-approved drug/ cell line combinations with maximum GI50 fold changes > 2, > 5, > 10, > 100, and > 1000 for combinations with > 5, > 10, > 20, and > 100 experiments (Fig. 2B, Fig. 2C, Suppl. Table 7, Suppl. Table 10), we obtained similar results to those across all compounds (Fig. 1E).

In agreement with the findings across all compound/ cell line combinations, the maximum GI50 fold changes also increased with experiment number when the FDA approved drug/cell line combinations were grouped into combinations with two experiments, 3–5 experiments, 6–20 experiments, and > 20 experiments (Fig. 2D, Suppl. Table 11), and the maximum GI50 fold change was also correlated with the number of experiments performed (Spearman's correlation coefficient = 0.72, $p < 2.2 \times 10^{-16}$) (Suppl. Fig. 3 C).

Moreover, the maximum GI50 fold change increased with the concentration range covered (Fig. 2E, Suppl. Table 12), and there was a significant correlation between the concentration range and the maximum GI50 fold change (Spearman's correlation coefficient = 0.62, $p < 2.2 \times 10^{-16}$) (Suppl. Fig. 3D).

Taken together, there is no indication that FDA-approved drugs would display less variability than experimental compounds.

2.5. GI50 variability in experiments performed by month

The reproducibility of results may be affected by parameters such as changes in the reagents, e.g. use of different lots or batches, different experimenters, and using cell lines at different passages [19,46,47]. Hence, near-contemporaneous experiments may be expected to display greater similarity than experiments performed at more distant points in time during the decades of anti-cancer compound testing by the NCI60.

To investigate the effects of the time of testing on data variability, we compared experiments performed in the same month to control samples of the same size that were randomly selected across the whole testing period. For this analysis, we used the 18 FDA-approved drugs that were tested at least 100 times in at least one cell line over periods of 95–275 months (Fig. 3 A, Suppl. Table 13), resulting in 51,872 drug/cell line combinations and a total of 321,709 experiments (Fig. 3B, Suppl. Table 14).

For every set of experiments performed on the same date, we generated 1000 random control samples across all dates of the same size and compared the value distribution. The variability of GI50 fold changes for same date experiments was indeed lower than for random control samples, but remained high reaching up to 250,035 (Fig. 3 C, Suppl. Table 15) with a mean and median of 298.2 (SD = 2940) and 4.2 (IQR = 17.6). Notably, for 45% of the same date drug/cell line combinations the GI50 fold change was higher than the mean fold change of the corresponding 1000 random samples (Fig. 3D, Suppl. Table 16).

When we looked at the data per cell line, the same-date GI50 fold changes were higher than the mean random sample fold changes for the majority of drugs in ten cell lines, higher for half of the drugs in three cell lines, and lower for the majority of drugs in the remaining 47 cell lines (Fig. 3E, Suppl. Table 17). When we looked at the individual drugs, six displayed a majority of drug/cell line combinations with higher mean same date GI50 fold changes higher than in the random samples and twelve drugs displayed lower ones (Fig. 3F, Suppl. Table 18).

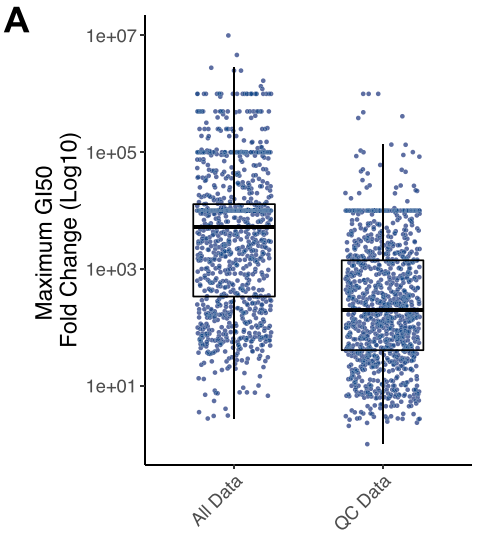
Taken together, experiments performed in close temporal proximity display lower variability than experiments performed over a longer time period, but the variability remains very high, even between experiments conducted on the same date.

2.6. GI50 fold changes remain high after removal of outliers

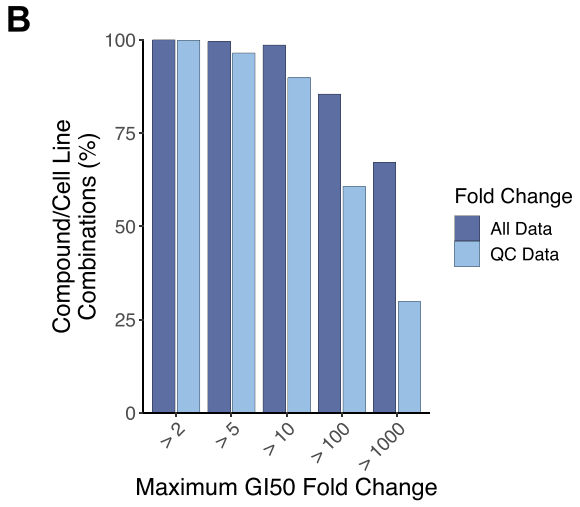
We also determined GI50 outliers for compound/cell line combinations with 5 or more experiments (738 compounds, 30,212 compound/cell line combinations, 598,243 GI50 values) using the adjusted boxplot method [48]. 5.7% (34,216) of GI50 values were outliers and 43.7% (13,208/30,212) of compound/cell line combinations had at least one GI50 outlier (Suppl. Table 19).

The highest percentage of outliers was 50% (7/14 experiments for maytansine in DU-145 prostate cancer cells) (Fig. 3G, Suppl. Table 19). The greatest number of outliers was 291 (16.8%) out of 1731 experiments for 5-fluorouracil in HOP-62 lung cancer cells (Fig. 3H, Suppl. Table 19). Outlier number increased with the number of experiments for a compound/cell line combination with a Spearman correlation coefficient of 0.25 ($p < 2.2 \times 10^{-16}$) (Suppl. Fig. 4). The removal of outliers reduced data variability, but the overall variability remained high with a maximum GI50 fold range of 2.5×10^{9} detected for maytansine in the ovarian cancer cell line OVCAR-5 over 35 experiments (Fig. 3I, Suppl. Table 19).

As detected in the analysis across all experiments, maximum GI50 fold changes increased with the number of experiments and the concentration ranges covered also after the removal of outliers (Fig. 3J, Fig. 3K, Suppl. Table 19, Suppl. Table 20). A significant correlation was observed between experiment number and maximum GI50 fold change with a Spearman correlation of 0.39 ($p < 2.2 \times 10^{-16}$) (Suppl. Fig. 5A) and between concentration range and maximum GI50 fold change with a Spearman correlation of 0.47 ($p < 2.2 \times 10^{-16}$) (Suppl. Fig. 5B).



Data	Mean (± SD)	Median (IQR)
All Data	78,904 (402,545)	5,254 (337.9-12,912)
QC Data	6,734 (57,977)	201.1 (40.7-1,409)



FC Group	QC Data	
	Number of Combinations	% of Combinations
>2	1079	99.91
>5	1042	96.48
>10	970	89.81
>100	655	60.65
>1000	322	29.81

Fig. 4. GI50 variability remains high for drugs that were tested 100 times in at least one cell line when only quality-controlled (QC) [32] data are considered. A) Comparison of maximum GI50 fold changes using all data or quality-controlled (QC) data. **B)** Percentage of compound/ cell line combinations with maximum GI50 fold changes > 2, > 5, > 10, > 100, and > 1000.

2.7. GI50 fold changes remain high when using quality-controlled data

Notably, there is awareness of the data variability within the NCI60 project as indicated by the availability of quality-controlled GI50 values in CellMiner. The quality control procedure includes the removal of experiments, in which a given drug was tested in the NCI60 panel, when the GI50 range was smaller than $\log_{10}$ 1.2 across the investigated cell lines or when the drug was tested in fewer than 35 cell lines. For experiments that pass these criteria, Pearson's correlation coefficients are determined across the individual drug screens. After removal of any experiments with a mean Pearson's correlation coefficient less than 0.334, the experiment with the lowest correlation coefficient (< 0.6) is dropped and mean corelations for experiments recalculated. This is repeated until all experiments have a mean correlation coefficient of 0.6 or higher, or a maximum of 25% or 253 (whichever is lower) experiments remain (minimum 2 experiments) [32]. Among FDA-approved drugs, this resulted in the exclusion of up to 96% of experiments (48 out of 50) for a given compound (i.e. mitotane) (Suppl. Table 21) for the determination of mean GI50 values in COMPARE analyses [32,49].

Quality-controlled data were available for 1080 drug/cell line combinations among the 18 drugs that had been tested at least 100 times in one or more cell lines, and the GI50 variability remained high (Fig. 4, Suppl. Fig. 6, Suppl. Table 22). The highest maximum GI50 fold change (1.0×10^6) was detected for vinblastine in the leukaemia cell line SR (over 41 experiments) and the melanoma cell lines SK-MEL-28 (31 experiments) and UACC-257 (52 experiments). The mean maximum GI50 fold change for a drug ranged from 10.47 (SD = 8.27) for doxorubicin to 87,611 (SD = 231,891) for vinblastine, while the median maximum GI50 fold change ranged from 6.58 (IQR = 4.58–10.1) for carmustine to 10,000 for cytarabine (IQR = 822.7–10,000) (Suppl. Fig. 6, Suppl. Table 22).

Furthermore, when only considering quality-controlled data, all but one (1079 out of 1080, 99.9%) drug/ cell line combination displayed a maximum GI50 fold change of > 2, 96.5% a maximum GI50 fold change of > 5, 89.8% a maximum GI50 fold change of > 10%, and 60.6% a maximum GI50 fold change of > 100 (Fig. 4B, Suppl. Table 22).

2.8. No drift in drug sensitivity over time

Cancer cell lines may display substantial changes in genotype and phenotype over time [46,50]. Hence, part of the variability observed in drug sensitivity may be the consequence of a shift in drug response over time. To investigate this, we established timelines of the GI50 values for the 18 compounds, which had been tested at least 100 times in one or

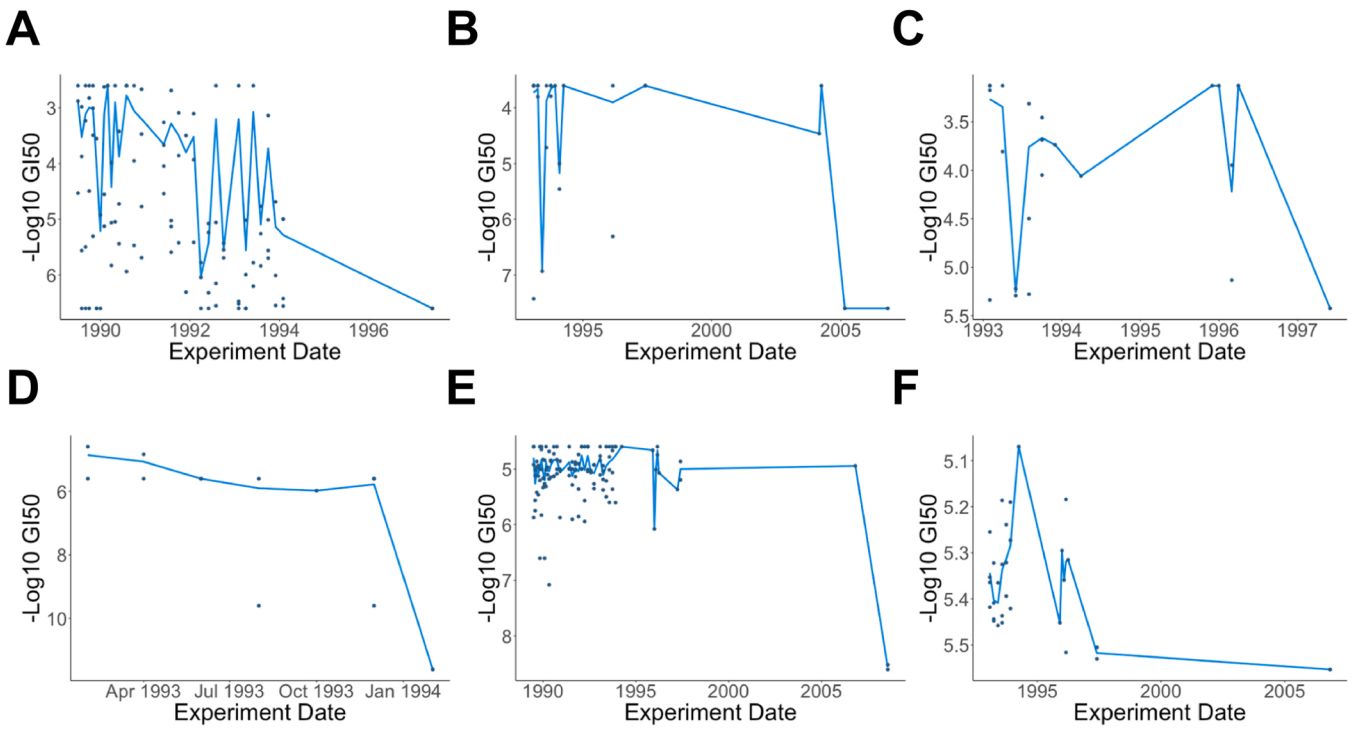
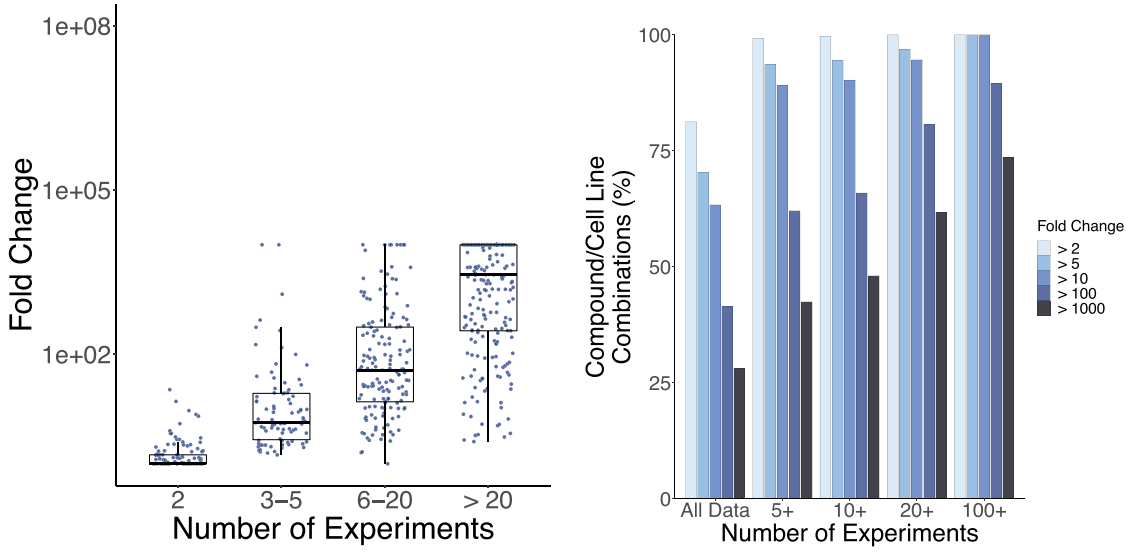


Fig. 5. Experimental time lines for individual compound/cell line combinations. The six experimental timelines for compound/cell line combinations (with more than 100 experiments performed in at least one cell line) with a fold change between the first and last mean GI50 that is 50% greater than the maximum GI50 for the remaining experiments. A) floxuridine in the ovarian cancer cell line SK-OV-3, B) methotrexate in the breast cancer cell line BT-549, C) 6-mercaptopurine in the breast cancer cell line BT-549, D) bleomycin in the leukaemia cell line K-562, E) vinblastine in the breast cancer cell line T-47D, F) M-AMSA in the melanoma cell line MDA-MB-435. Plot lines represent mean GI50s while data points represent individual experiments at an experimental date.



Group	Mean (SD)	Median (IQR Range)
2	1.82 (2.58)	1 (1 - 2.58)
3 to 5	269.7 (1508.3)	5.69 (2.76 - 19.32)
6 to 20	857.58 (2321.4)	51.35 (13.56 - 313.7)
> 20	4701.9 (4463.1)	2831.92 (268.53 - 10000)

Fig. 6. Variability between experiments performed using FDA-approved drugs at the same drug concentration range. GI50 fold changes in dependence of the number of experiments per drug/ cell line combination and percentages of compound/ cell line combinations with maximum fold changes above the indicated thresholds, and mean and median fold changes.

more cell lines. This resulted in time lines for 1080 compound/cell line combinations with time frames ranging from 95 months (vinblastine, floxuridine, cytarabine, daunorubicin, 6-mercaptopurine) to 275 months (Doxorubicin) (Fig. 3A, Suppl. Table 13).

Drug/ cell line combinations, in which the fold change between the mean GI50 on the first experimental date and the mean GI50 on the last experimental date was 50% or greater than the maximum GI50 fold change for the data points in between were considered as candidates for a drift in drug sensitivity. Only six (0.56%) out of 1080 drug/cell line combinations fulfilled these criteria (Fig. 5, Suppl. Table 23).

The distribution of the individual GI50 values for three of the drug/ cell line combinations (floxuridine/ SK-OV-3, methotrexate/ BT-549, 6-mercaptopurine/ BT-549) did not indicate a GI50 shift over time (Fig. 5A-C, Suppl. Table 24). For the other three drug/ cell line combinations (bleomycin/ K-562, vinblastine/ T-47D, M-AMSA/ MDA-MB-435) a drift in sensitivity appears unlikely but cannot be excluded based on the data (Fig. 5D-F, Suppl. Table 24). However, such observations are very rare. Moreover, a phenotypic drift in a cell line would be expected to result in changes in sensitivity to more than one drug over time. Hence, the data provide no evidence suggesting that the drug sensitivity of individual cell lines may have changed over time. These findings may also reflect that the NCI60 uses cell lines within a window of 30 passages [38].

2.9. High GI50 variability when only considering drug-response curves covering the same concentration range

Finally, we analysed the variability among FDA-approved drug/ cell line combinations that were repeatedly tested at the same drug concentration range. Ten drugs were tested at the standard five-point 1:10 dilution range with a starting concentration of 25 μM, resulting in 581 drug/cell line combinations. The variability remained high with a maximum fold change of 10,000, reflecting the fold change between the lowest and the highest tested concentrations (values were not extrapolated, if the GI50 was not reached the highest or lowest tested concentration was used as cut-off), observed for 94 drug/cell line combinations (Fig. 6, Suppl. Table 25). The mean maximum fold change for a drug across the different cell lines ranged from 1.02 (S.D. = 0.14) for tamoxifen (which was tested in two independent experiments per cell line) to 9943 (S.D. = 437.8) for doxorubicin (which was tested in 1894 to 2265 independent experiments per cell line), and the median maximum fold change from 1 (IQR = 1 – 1) for tamoxifen to 10,000 (IQR = 10,000–10,000) for doxorubicin (Supp. Table 26). The maximum fold change increased as the number of experiments increased (Fig. 6). For drug/cell line combinations with two experiments, the mean and median maximum GI50 fold changes were 1.82 (SD = 2.58) and 1 (IQR = 1 – 2.58). For drug/cell line combinations with more than 20 experiments, the mean and median maximum GI50 fold changes were 4702 (SD = 4463) and 2832 (IQR = 269 – 10,000) (Fig. 6).

Among drug/cell line combinations with two or more experiments, 81% (472/581) displayed a maximum GI50 fold change of > 2, 70.4% (409/581) a maximum GI50 fold change of > 5, 63.3% (368/581) a maximum GI50 fold change of $> 10\%$, and 41.5% (241/581) a maximum GI50 fold change of > 100. Among drug/cell line combinations with five or more experiments, 99.2% (374/377) displayed a maximum GI50 fold change of > 2, 93.6% (353/377) a maximum GI50 fold change of > 5, 89.1% (336/377) a maximum GI50 fold change of $> 10\%$, and 62.1% (234/377) a maximum GI50 fold change of > 100 (Fig. 6, Supp. Table 29).

3. Discussion

In this study, we investigated variation among data derived from the NCI60 screen that has tested compounds for anti-cancer activity for decades following highest level quality control procedures [31–42]. Despite this very strict approach in a world-leading research environment, data variability was very high. The largest fold change between the lowest and highest GI50 in a given compound/cell line combination was 3.16×10^{10}. Overall, mean and median GI50 fold changes were 318,410 (SD = 5.71×10^{7}) and 1.6 (IQR = 1.1–3.4). As might have been expected, the fold change between the lowest and the highest GI50 in a specific compound cell line combination increased with the number of experiments and the concentration range tested.

CellMiner contains data on experimental compounds as well as on FDA-approved drugs that are in clinical use [32–37,51]. Although FDA-approved drugs might have been expected to result in more robust data, this was not the case and they displayed a similar data variability as that determined across all compounds. The variability also remained high when we only considered experiments that were performed in the same months, removed outliers, or only considered experiments, in which drugs were tested at the same concentration range.

There is awareness of this variability within the NCI60 project as indicated by the awareness of quality-controlled NCI60 GI50 data in CellMiner, which results in the exclusion of up to 96% of experiments (48 out of 50) for a given compound (Suppl. Table 21) [32]. Such an approach would not be a feasible approach in most research labs. Even when we only considered these quality-controlled data among the 18 compounds that were tested 100 or more times in at least one cell line, all but one (1079 out of 1080, 99.9%) drug/ cell line combination displayed a maximum GI50 fold change of > 2), 96.5% a maximum GI50 fold change of > 5, 89.8% a maximum GI50 fold change of $> 10\%$, and 60.6% a maximum GI50 fold change of > 100.

This large GI50 variation among dose response experiments repeatedly using the same drug in the same cell line is of relevance for the assessment of the potential clinical activity of drug candidates. Cytotoxic anti-cancer drugs are typically used at maximum tolerated doses that cannot be further increased without unacceptable toxicity [52–54]. Moreover, the maximum effects of targeted drugs, e.g. antibodies or kinase inhibitors that interfere with cancer-specific structures or entities, do not further increase beyond the ‘optimal biological dose’, i.e. the dose at which the biological target is completely inhibited [52–55]. Hence, even a two-fold difference in the GI50, which occurred in 25,496 (84.4%) of 30,212 compound/cell line combinations with at least five experiments, is of potential relevance, as a two-fold increase of the clinical dose of an anti-cancer drug is rarely feasible.

Data variability was not driven by changes in the sensitivity of the NCI60 cell lines over time, which indicates that the use of cell lines within 30 passages indeed prevented phenotypic drift [38]. Since the NCI60 applies strict quality measures that control for all known sources data variability in cancer cell line experiments including consistent sourcing of reagents and materials (including cell culture media and foetal calf serum), authentication of cell lines, testing for contamination, consistent cell numbers, and using doxorubicin as internal control [31, 38–52,56], the observed data variability has at least in part to be attributed to the variation that is inherently associated with the use of biological systems. Even among the quality-controlled data for doxorubicin maximum GI50 fold changes ranged from 2.4 (in the prostate cancer cell line DU-145) to 56.9 (in the renal cancer cell line A498) (Suppl. Table 22).

In conclusion, this study analysed data replicability in the largest dataset that has been investigated for this purpose. In contrast to other replication studies that often use complex model systems and novel technologies [26–30], the NCI60 dataset reports data from a comparably simple cancer cell line screen that has been performed over decades under highly standardised conditions in a world-leading environment applying the highest standards to avoid known sources of data variability [31,38–41], i.e. under ideal conditions that the vast majority of research groups will not be able to afford. Hence, a significant part of the observed data variation is likely due to the inherent complexity of biological systems. Strict experimental standardisation as suggested by many commentators to improve data quality and reliability [5,14–19, 57–60] does therefore not appear to be a straightforward way to resolve

issues associated with limited replicability in preclinical cancer research and high attrition rates during cancer drug development. In this context, our data support the notion that the primary value of pharmacological cell line studies lies in the generation and validation of hypotheses rather than the direct prediction of clinical drug activity (which is also likely to be an unrealistic aim, given that cell lines do not adequately represent the tumour environment) [61,62]. Notably, our findings are not only of relevance to the cancer field, but to the whole life science field, in which there is a perception of a "reproducibility crisis" or "replication crisis" [4–6,14,19–21,63–67].

Awareness of the inherent variability of experimental results, will help researchers to develop a realistic understanding of the meaning of their data and inform and inspire further research that will improve the robustness, reliability, and meaningfulness of research data. Experiment heterogenisation, the testing of a hypothesis in many different (experimental) systems and datasets and different laboratories, has been suggested to provide more robust and meaningful data, in particular for more complex experimental systems that are known to be affected by reproducibility issues [2,25–28,68–70]. Such a strategy is also in line with the NCI60 strategy that emphasises that follow-up testing, including animal testing, is needed to identify drug leads [31]. Moreover, the involvement of patient-derived cancer models may improve data robustness [71]. However, further research will need to show whether and, if yes, to which extent such strategies and others that are still to be developed will actually improve data replicability and robustness.

4. Methods

4.1. Data availability

All data were obtained from CellMiner [72] Version 2.2. Dose concentration range data (June 2018 release) were obtained from the National Cancer Institute DTP NCI bulk data for download pages (https://wiki.nci.nih.gov/display/NCIDTPdata/NCI-60 +-Growth+Inhibition+Data). Of the 52,585 NCI codes given to compounds tested on the NCI-60 cell line panel, 42,794 were given to unnamed compounds and 9791 were given to 9027 named compounds. 262 named compounds received two or more individual codes. Some GI50 values represent minimum or maximum drug concentrations where the actual GI50 was not reached [32]. Since such values understate the actual data variation, we did not remove these data. All data generated during this study are included in this published article and its supplementary information files.

4.2. Maximum GI50 fold change calculation

The drug sensitivity data was converted from –log10 GI50, to the GI50 (μM) for all compound, all cell lines and all experiments. Maximum fold changes were calculated for each compound/cell line combination with more than one experiment (594,450) by dividing the maximum GI50 for a cell line by the minimum GI50.

4.3. Number of experiments and experimental groups

The number of experiments for each individual compound/cell line combination was calculated by counting all experiments performed on the same experimental date as well as experiments on different dates. The relationship between number of experiments and maximum fold change was investigated by using Spearman's correlation coefficient as the distribution of maximum GI50 fold change was not normal.

The compound/cell line combinations were then assigned experimental groupings based on the number of experiments performed: all data, 5 or more experiments, 10 or more experiments, 20 or more experiments, and 100 or more experiments. This allowed comparison of "high" maximum fold changes (>2, >5, >10, >100, and >1000) for combinations with varied number of experiments. Additionally, compound/cell line combinations were assigned to experimental groups: 2 experiments, 3–5 experiments, 6–20 experiments and over 20 experiments. These experimental groupings enabled comparison of GI50 fold change statistics (mean, median, minimum, maximum, variance) for compound/cell line combinations with number of experiments ranging from lower to higher.

4.4. Concentration range and experimental groups

Maximum dose concentration range for a compound/cell line combination was determined by using the minimum and maximum dose concentration used in an experiment for an individual compound on an individual cell line. The minimum concentration range was $1.0 \times 10^{1.2}$ and the maximum concentration was $1.0 \times 10^{12.1}$. The relationship between dose concentration range and maximum fold change was investigated by using Spearman's correlation coefficient as the distribution of maximum GI50 fold change was not normal.

Compound/cell line combinations were assigned to groups based on the dose concentration range for that combinations: maximum concentration range less than 1.0×10^5, maximum concentration range 1.0×10^5 to 1.0×10^9 exclusive and maximum concentration range 1.0×10^9 and above. These experimental groupings enabled comparison of GI50 fold change statistics (mean, median, minimum, maximum, variance) for compound/cell line combinations between lower and higher concentration ranges.

4.5. Individual dose concentration range vs all dose concentration ranges

For this analysis, the cisplatin GI50s were used from cell lines, in which cisplatin was examined in at least 100 experiments. The most common individual 5-fold dose concentration range for cisplatin was 0.05–500uM, which was used for 79% of cisplatin/cell line combination experiments. The maximum GI50 fold change was calculated for 100 random experiments using the 0.05–500uM concentration range 1000 times and the median maximum GI50 fold change over all iterations for a drug/cell line combination was calculated. The same method was used on the data considering all dose concentration ranges, and the median maximum GI50 fold changes were compared, including statistical analyses using Wilcoxon rank sum test, and FDR calculated using Benjamini-Hochberg multiple test correction.

4.6. FDA-approved compound analysis

All compounds that were classed as FDA-approved drugs by the NCI-60 in CellMiner Database Version 2.2 and where two or more experiments had been performed were extracted from the complete dataset. This created an FDA-approved dataset of 181 drugs for which 399,686 experiments for 9970 individual drug/cell line combinations were performed. Analysis of relationship between the number of experiments/concentration ranges and maximum GI50 fold change for drug/cell line combinations was performed as for the complete dataset, described above.

4.7. Experiments on the same date

Month and year of each experiment was available so experimental timelines were established for compounds by calculating the time between the first and last experiment date. Multiple experiments were carried out on the same date for many of the compound/cell line combinations, particularly the 18 compounds with at least one cell line with 100 total experiments. The data for these 18 compounds, 17 of which were FDA-approved, was extracted from the complete dataset to create a subset of data deemed suitable to compare GI50 variability on the same date with GI50 variability over an experimental timeline.

The maximum GI50 fold change on each date where there were

multiple experiments for a compound/cell line combination were calculated by dividing maximum GI50 by minimum GI50 value. The number of experiments on a specific date for a compound/cell line combination was used to determine the maximum GI50 fold change over the same number of experiments picked randomly from that combination's experimental timeline. This was performed 1000 times so that for every compound/cell line combination and experimental date with a maximum GI50 fold change over multiple experiments there were 1000 corresponding maximum GI50 fold changes calculated from random samples of the same number of experiments on that compound/cell line combination's timeline. The mean maximum GI50 fold change was calculated for the 1000 random samples for each compound/cell line combination and the number of maximum GI50 fold changes for experiments on the same date higher and lower than the random sample mean were counted. For each compound, significance of the difference between same date maximum GI50 fold change and sample mean GI50 fold change was calculated using Wilcoxon Rank Sum Test. This was performed using all cell line data combined for each compound and for each cell line individually for each compound. Where a significant difference between same date and random sample mean maximum GI50 fold changes were observed, the number of times the same date GI50 fold change was higher or lower than the random sample mean maximum GI50 fold change was counted.

4.8. Analysis of quality-controlled data

The quality controlled-data for the 18 most tested drugs was obtained from CellMiner [72] Version 2.7. For a total of 1080 drug/ cell line combinations, a total of 101,912 individual GI50 values were available for analysis (compared to 326,788 when considering all data). Analyses performed on the quality controlled data were as previously described. Quality control methods are described in [32].

4.9. Drift in drug sensitivity

The mean GI50 fold change was calculated for each experimental date (month) for the 18 compounds with 100 or more experiments for at least one cell line. The GI50 fold change between the first experimental date and the last experimental date was calculated using the mean GI50 on those dates. The first/last GI50 fold change was then compared to the maximum GI50 fold change for each compound/cell line combination and considered a candidate for a drift in sensitivity if it was 50% or more of the maximum fold change.

4.10. Removal of outliers

The adjusted boxplot method was used to identify outlier thresholds. This method was chosen as the data set was highly skewed. To use this method the medcouple (*MC)*, a robust measure of skewness, had to be calculated (where $X_n = \{x_1, x_2, \ldots, x_n\}$ represents data for every compound/cell line combination):

$$MC(x_1, \ldots, x_n) = med\frac{(x_j - med_k) - (med_k - x_i)}{x_j - x_i}$$

Where med_k is the median of X_n, and i and j have to satisfy $x_i \leq med_k \leq x_j$, and $x_i \neq x_j$.

Using the MC the upper (U) and lower (L) thresholds could be determined. If ≥ 0:

$$L = Q_1 - 1.5 \times \exp(-3.5MC) \times IQR$$

$$U = Q_3 + 1.5 \times \exp(4MC) \times IQR$$

If≤ 0:

$$L = Q_1 - 1.5 \times \exp(-4MC)IQR$$

$$U = Q_3 + 1.5 \times \exp(3.5MC) \times IQR$$

If $MC = 0$ the adjusted boxplot method was not used but instead the Tukey method was used:

$$L = Q_1 - 1.5IQR$$

$$U = Q_3 + 1.5IQR$$

Where Q_1 is the lower quartile, Q_3 is the upper quartile and IQR is the interquartile range.

For each compound/cell line combination any GI50 value below L or above U were removed from the dataset. Analyses were performed on this dataset as previously described for the complete dataset.

4.11. Data processing

Data was carried out using perl version 5.26.0, Microsoft Excel (2011) and R statistical packages version 3.4.4. Perl modules Statistics: Descriptive and Statistics::R were used. Packages used in R were robustbase, dplyr, webr, moonBook, tidyverse, reshape2, scales, gplots, ggpubr, ggExtra, RColorBrewer, corrplot, ggplot2, and tidyr.

Significance statement

Only 5% of anti-cancer drug candidates entering clinical trials are eventually approved. This is attributed to a lack of robustness in preclinical research, although the technically achievable replicability level remains unknown. The NCI60 screen has tested compounds in cancer cell lines since 1989 following the strictest quality measures accounting for the sourcing of reagents/ materials, cell identity, contamination, assay parameters, and phenotypic drift. Data variability remains high even under these optimized conditions. 71% of compound/ cell line combinations with $>$ 100 experiments displayed maximum GI50 (50% growth inhibition) fold changes (highest/ lowest GI50) $>$ 1000. The highest maximum fold change was 3.16 $\times$ 1010. Awareness of this inherent variability is crucial for the development of robust approaches and for improving success rates in therapy development.

CRediT authorship contribution statement

IR acquired and investigated the data. MNW and MM conceptualised the study. IR, MNW, and MM developed the study design. All authors analysed and interpreted the findings. IR and MM wrote the initial manuscript. All authors revised the manuscript. All authors read and approved the final manuscript.

Data Availability

Data will be made available on request.

Acknowledgements

The authors thank Dr Robert H Shoemaker for critical reading of our manuscript and helpful discussion.

Declarations of interest

none.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.phrs.2023.106671.

References

[1] C.H. Wong, K.W. Siah, A.W. Lo, Estimation of clinical trial success rates and related parameters, Biostatistics 20 (2) (2019) 273–286.
[2] A. Honkala, S.V. Malhotra, S. Kummar, M.R. Junttila, Harnessing the predictive power of preclinical models for oncology drug development, Nat. Rev. Drug Disco 21 (2022) 99–114.
[3] P.B. Kane, J. Kimmelman, Is preclinical research in cancer biology reproducible enough, Elife 10 (2021), e67527, https://doi.org/10.7554/eLife.67527.
[4] F. Prinz, T. Schlange, K. Asadullah, Believe it or not: how much can we rely on published data on potential drug targets? Nat. Rev. Drug Discov. 10 (2011) 712.
[5] C.G. Begley, L.M. Ellis, Drug development: raise standards for preclinical cancer research, Nature 483 (2012) 531–533.
[6] A. Mobley, S.K. Linder, R. Braeuer, L.M. Ellis, L. Zwelling, A survey on data reproducibility in cancer research provides insights into our limited ability to translate findings from the laboratory to the clinic, PLoS One 8 (2013), e63221.
[7] Z. Liu, B. Delavan, R. Roberts, W. Tong, Lessons learned from two decades of anticancer drugs, Trends Pharm. Sci. 38 (10) (2017) 852–872, https://doi.org/10.1016/j.tips.2017.06.005.
[8] C. Pinto, M.F. Estrada, C. Brito, In vitro and ex vivo models - the tumor microenvironment in a flask, Adv. Exp. Med Biol. 1219 (2020) 431–443, https://doi.org/10.1007/978-3-030-34025-4_23.
[9] T. Xia, W.L. Du, X.Y. Chen, Y.N. Zhang, Organoid models of the tumor microenvironment and their applications, J. Cell Mol. Med 25 (13) (2021) 5829–5841, https://doi.org/10.1111/jcmm.16578.
[10] T.M. Errington, A. Denis, N. Perfito, E. Iorns, B.A. Nosek, Challenges for assessing replicability in preclinical cancer biology, Elife 10 (2021), e67995, https://doi.org/10.7554/eLife.67995.
[11] T.M. Errington, M. Mathur, C.K. Soderberg, A. Denis, N. Perfito, E. Iorns, B. A. Nosek, Investigating the replicability of preclinical cancer biology, Elife 10 (2021 7), e71601, https://doi.org/10.7554/eLife.71601.
[12] T.M. Errington, A. Denis, A.B. Allison, R. Araiza, P. Aza-Blanc, L.R. Bower, J. Campos, H. Chu, S. Denson, C. Donham, K. Harr, B. Haven, E. Iorns, J. Kwok, E. McDonald, S. Pelech, N. Perfito, A. Pike, D. Sampey, M. Settles, D.A. Scott, V. Sharma, T. Tolentino, A. Trinh, R. Tsui, B. Willis, J. Wood, L. Young, Experiments from unfinished registered reports in the reproducibility project: cancer biology, Elife 10 (2021), e73430, https://doi.org/10.7554/eLife.73430.
[13] P. Rodgers, A. Collings, What have we learned, Elife 10 (2021), e75830, https://doi.org/10.7554/eLife.75830.
[14] C.G. Begley, J.P. Ioannidis, Reproducibility in science: improving the standard for basic and preclinical research, Circ. Res 116 (2015) 116–126.
[15] C.G. Begley, Six red flags for suspect work, Nature 497 (2013) 433–434.
[16] J.P. Ioannidis, S. Greenland, M.A. Hlatky, M.J. Khoury, M.R. Macleod, D. Moher, K. F. Schulz, R. Tibshirani, Increasing value and reducing waste in research design, conduct, and analysis, Lancet 383 (2014) 166–175.
[17] C. Hatzis, P.L. Bedard, N.J. Birkbak, A.H. Beck, H.J. Aerts, D.F. Stem, L. Shi, R. Clarke, J. Quackenbush, B. Haibe-Kains, Enhancing reproducibility in cancer drug screening: how do we move forward? Cancer Res 74 (2014) 4016–4023.
[18] W.G. Kaelin Jr., Publish houses of brick, not mansions of straw, Nature 545 (2017) 387.
[19] M.N. Wass, L. Ray, M. Michaelis, Understanding of researcher behavior is required to improve data reliability, Gigascience (2019) 8, giz017.
[20] M. Baker, 1,500 scientists lift the lid on reproducibility, Nature 533 (2016) 452–454.
[21] Nature Editorial, Checklists work to improve science, Nature 556 (2018) 273–274.
[22] D. Fanelli, R. Costas, J.P. Ioannidis, Meta-assessment of bias in science, Proc. Natl. Acad. Sci. USA 114 (2017) 3714–3719.
[23] A. Mullard, Half of top cancer studies fail high-profile reproducibility effort, Nature 600 (7889) (2021) 368–369, https://doi.org/10.1038/d41586-021-03691-0.
[24] T.F. França, J.M. Monserrat, Reproducibility crisis in science or unrealistic expectations? EMBO Rep. 19 (2018), e46008.
[25] N.A. Karp, Reproducible preclinical research-Is embracing variability the answer? PLoS Biol. 16 (2018), e2005413.
[26] B. Voelkl, L. Vogt, E.S. Sena, H. Würbel, Reproducibility of preclinical animal research improves with heterogeneity of study samples, PLoS Biol. 16 (2018), e2003693.
[27] C. Bodden, V.T. von Kortzfleisch, F. Karwinkel, S. Kaiser, N. Sachser, S.H. Richter, Heterogenising study samples across testing time improves reproducibility of behavioural data, Sci. Rep. 9 (2019) 8247.
[28] V.T. von Kortzfleisch, N.A. Karp, R. Palme, S. Kaiser, N. Sachser, S.H. Richter, Improving reproducibility in animal research by splitting the study population into several 'mini-experiments', Sci. Rep. 10 (2020) 16579.
[29] J.C. Crabbe, D. Wahlsten, B.C. Dudek, Genetics of mouse behavior: interactions with laboratory environment, Science 284 (1999) 1670–1672.
[30] D. Pinto, K. Darvishi, X. Shi, D. Rajan, D. Rigler, T. Fitzgerald, A.C. Lionel, B. Thiruvahindrapuram, J.R. Macdonald, R. Mills, A. Prasad, K. Noonan, S. Gribble, E. Prigmore, P.K. Donahoe, R.S. Smith, J.H. Park, M.E. Hurles, N.P. Carter, C. Lee, S.W. Scherer, L. Feuk, Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants, Nat. Biotechnol. 29 (2011) 512–520.
[31] R.H. Shoemaker, The NCI60 human tumour cell line anticancer drug screen, Nat. Rev. Cancer 6 (2006) 813–823.
[32] W.C. Reinhold, M. Sunshine, H. Liu, S. Varma, K. Kohn, J. Morris, J. Doroshow, Y. Pommier, CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set, Cancer Res 72 (2012) 3499–3511.
[33] W.C. Reinhold, S. Varma, F. Sousa, M. Sunshine, O.D. Abaan, S.R. Davis, S. W. Reinhold, K.W. Kohn, J. Morris, P.S. Meltzer, J.H. Doroshow, Y. Pommier, NCI-60 whole exome sequencing and pharmacological CellMiner analyses, PLoS One 9 (2014), e101670.
[34] W.C. Reinhold, M. Sunshine, S. Varma, J.H. Doroshow, Y. Pommier, Using CellMiner 1.6 for systems pharmacology and genomic analysis of the NCI-60, Clin. Cancer Res 21 (2015) 3841–3852.
[35] B.A. Chabner, NCI-60 cell line screening: a radical departure in its time, J. Natl. Cancer Inst. (2016) 108.
[36] W.C. Reinhold, S. Varma, M. Sunshine, V. Rajapakse, A. Luna, K.W. Kohn, H. Stevenson, Y. Wang, H. Heyn, V. Nogales, S. Moran, D.J. Goldstein, J. H. Doroshow, P.S. Meltzer, M. Esteller, Y. Pommier, The NCI-60 methylome and its integration into cellminer, Cancer Res. 77 (2017) 601–612.
[37] W.C. Reinhold, S. Varma, M. Sunshine, F. Elloumi, K. Ofori-Atta, S. Lee, J.B. Trepel, P.S. Meltzer, J.H. Doroshow, Y. Pommier, RNA sequencing of the NCI-60: Integration into CellMiner and CellMiner CDB, Cancer Res. 79 (2019) 3514–3524.
[38] P.L. Lorenzi, W.C. Reinhold, S. Varma, A.A. Hutchinson, Y. Pommier, S.J. Chanock, J.N. Weinstein, DNA fingerprinting of the NCI-60 cell line panel, Mol. Cancer Ther. 8 (2009) 713–724.
[39] M.C. Alley, D.A. Scudiero, A. Monks, M.L. Hursey, M.J. Czerwinski, D.L. Fine, B. J. Abbott, J.G. Mayo, R.H. Shoemaker, M.R. Boyd, Feasibility of drug screening with panels of human tumor cell lines using a microculture tetrazolium assay, Cancer Res. 48 (1988) 589–601.
[40] R.H. Shoemaker, A. Monks, M.C. Alley, D.A. Scudiero, D.L. Fine, T.L. McLemore, B. J. Abbott, K.D. Paull, J.G. Mayo, M.R. Boyd, Development of human tumor cell line panels for use in disease-oriented drug screening, Prog. Clin. Biol. Res. 276 (1988) 265–286.
[41] M.R. Boyd, K.D. Paull, Some practical considerations and applications of the national cancer institute in vitro anticancer drug discovery screen, Drug Dev. Res. 34 (1995) 91–109.
[42] National Cancer Institute, Developmental Therapeutics Program. Standard Operating Procedures for Sample Preparation for NCI60 Screen. https://dtp.cancer.gov/discovery_development/nci-60/handling.htm.
[43] J. Kormanec, R. Novakova, D. Csolleiova, L. Feckova, B. Rezuchova, B. Sevcikova, D. Homerova, The antitumor antibiotic mithramycin: new advanced approaches in modification and production, Appl. Microbiol. Biotechnol. 104 (2020) 7701–7721.
[44] J. Gallego-Jara, G. Lozano-Terol, R.A. Sola-Martínez, M. Cánovas-Díaz, de Diego, T. Puente, A compressive review about Taxol®: history and future challenges, Molecules 25 (2020) 5986.
[45] J.M. Rae, C.J. Creighton, J.M. Meck, B.R. Haddad, M.D. Johnson, MDA-MB-435 cells are derived from M14 melanoma cells–a loss for breast cancer, but a boon for melanoma research, Breast Cancer Res. Treat. 104 (2007) 13–19.
[46] U. Ben-David, B. Siranosian, G. Ha, H. Tang, Y. Oren, K. Hinohara, C.A. Strathdee, J. Dempster, N.J. Lyons, R. Burns, A. Nag, G. Kugener, B. Cimini, P. Tsvetkov, Y. E. Maruvka, R. O'Rourke, A. Garrity, A.A. Tubelli, P. Bandopadhayay, A. Tsherniak, F. Vazquez, B. Wong, C. Birger, M. Ghandi, A.R. Thorner, J.A. Bittker, M. Meyerson, G. Getz, R. Beroukhim, T.R. Golub, Genetic and transcriptional evolution alters cancer cell line drug response, Nature 560 (2018) 325–330.
[47] N. Noronha, G. Ehx, M.C. Meunier, J.P. Laverdure, C. Thériault, C. Perreault, Major multilevel molecular divergence between THP-1 cells from different biorepositories, Int J. Cancer 147 (2020) 2000–2006.
[48] M. Hubert, E. Vandervieren, An adjusted boxplot for skewed distributions, Comput. Stat. Data Anal. 52 (2008) 5186–5201.
[49] K.D. Paull, R.H. Shoemaker, L. Hodes, A. Monks, D.A. Scudiero, L. Rubinstein, J. Plowman, M.R. Boyd, Display and analysis of patterns of differential activity of drugs against human tumor cell lines: development of mean graph and COMPARE algorithm, J. Natl. Cancer Inst. 81 (14) (1989) 1088–1092, https://doi.org/10.1093/jnci/81.14.1088.
[50] Y. Liu, Y. Mi, T. Mueller, S. Kreibich, E.G. Williams, A. Van Drogen, C. Borel, M. Frank, P.L. Germain, I. Bludau, M. Mehnert, M. Seifert, M. Emmenlauer, I. Sorg, F. Bezrukov, F.S. Bena, H. Zhou, C. Dehio, G. Testa, J. Saez-Rodriguez, S. E. Antonarakis, W.D. Hardt, R. Aebersold, Multi-omic measurements of heterogeneity in HeLa cells across laboratories, Nat. Biotechnol. 37 (2019) 314–322.
[51] A. Monks, D.A. Scudiero, G.S. Johnson, K.D. Paull, E.A. Sausville, The NCI anti-cancer drug screen: a smart screen to identify effectors of novel targets, Anticancer Drug Des. 12 (1997) 533–541.
[52] E.A. Eisenhauer, P.J. O'Dwyer, M. Christian, J.S. Humphrey, Phase I clinical trial design in cancer drug development, J. Clin. Oncol. 18 (2000) 684–692.
[53] J.R. Sachs, K. Mayawala, S. Gadamsetty, S.P. Kang, D.P. de Alwis, Optimal dosing for targeted therapies in oncology: drug development cases leading by example, Clin. Cancer Res 22 (2016) 1318–1324.
[54] A. Mansinho, V. Boni, M. Miguel, E. Calvo, New designs in early clinical drug development, Ann. Oncol. 30 (2019) 1460–1465.
[55] P. Corbaux, M. El-Madani, M. Tod, J. Péron, D. Maillet, J. Lopez, G. Freyer, B. You, Clinical efficacy of the optimal biological dose in early-phase trials of anti-cancer targeted therapies, Eur. J. Cancer 120 (2019) 40–46.
[56] M. Pons, G. Nagel, Y. Zeyn, M. Beyer, T. Laguna, T. Brachetti, A. Sellmer, S. Mahboobi, R. Conradi, F. Butter, O.H. Krämer, Human platelet lysate as validated replacement for animal serum to assess chemosensitivity, ALTEX 36 (2) (2019) 277–288.
[57] Z. Safikhani, P. Smirnov, M. Freeman, N. El-Hachem, A. She, Q. Rene, A. Goldenberg, N.J. Birkbak, C. Hatzis, L. Shi, A.H. Beck, H.J.W.L. Aerts, J. Quackenbush, B. Haibe-Kains, Revisiting inconsistency in large pharmacogenomic studies. Version 3, F1000Res. 5 (2016) 2333.

[58] L.P. Freedman, I.M. Cockburn, T.S. Simcoe, The economics of reproducibility in preclinical research, PLoS Biol. 13 (2015), e1002165.
[59] M.F. Jarvis, M. Williams, Irreproducibility in preclinical biomedical research: perceptions, uncertainties, and knowledge gaps, Trends Pharmacol. Sci. 37 (2016) 290–302.
[60] L.P. Freedman, G. Venugopalan, R. Wisman, Reproducibility2020: progress and priorities, F1000Res 6 (2017) 604.
[61] J.N. Weinstein, P.L. Lorenzi, Cancer: discrepancies in drug sensitivity, Nature 504 (7480) (2013) 381–383.
[62] L. Trastulla, J. Noorbakhsh, F. Vazquez, J. McFarland, F. Iorio, Computational estimation of quality and clinical relevance of cancer cell lines, Mol. Syst. Biol. 18 (7) (2022), e11017.
[63] S.V. Frye, M.R. Arkin, C.H. Arrowsmith, P.J. Conn, M.A. Glicksman, E.A. Hull-Ryde, B.S. Slusher, Tackling reproducibility in academic preclinical drug discovery, Nat. Rev. Drug Disco 14 (2015) 733–734.
[64] D.J. Drucker, Never waste a good crisis: confronting reproducibility in translational research, Cell Metab. 24 (2016) 348–360.
[65] D. Fanelli, Opinion: Is science really facing a reproducibility crisis, and do we need it to? Proc. Natl. Acad. Sci. USA 115 (2018) 2628–2631.
[66] G. Samsa, L. Samsa, A guide to reproducibility in preclinical research, Acad. Med 94 (2019) 47–52.
[67] M.R. Munafò, C.D. Chambers, A.M. Collins, L. Fortunato, M.R. Macleod, Research culture and reproducibility, Trends Cogn. Sci. 24 (2020) 91–93.
[68] N.A. Karp, A.O. Speak, J.K. White, D.J. Adams, M. Hrabé de Angelis, Y. Hérault, R. F. Mott, Impact of temporal variation on design and analysis of mouse knockout phenotyping studies, PLoS One 9 (2014), e111239.
[69] K.F. Ding, D. Finlay, H. Yin, W.P.D. Hendricks, C. Sereduk, J. Kiefer, A. Sekulic, P. M. LoRusso, K. Vuori, J.M. Trent, N.J. Schork, Analysis of variability in high throughput screening data: applications to melanoma cell lines and drug responses, Oncotarget 8 (2017) 27786–27799.
[70] N. Kafkafi, I. Golani, I. Jaljuli, H. Morgan, T. Sarig, H. Würbel, S. Yaacoby, Y. Benjamini, Addressing reproducibility in single-laboratory phenotyping experiments, Nat. Methods 14 (2017) 462–464.
[71] K.F. Idrisova, H.U. Simon, M.O. Gomzikova, Role of patient-derived models of cancer in translational oncology, Cancers (2022) 15.
[72] U.T. Shankavaram, S. Varma, D. Kane, M. Sunshine, K.K. Chary, W.C. Reinhold, Y. Pommier, J.N. Weinstein, CellMiner: a relational database and query tool for the NCI-60 cancer cell lines, BMC Genom. 10 (2009) 277.