Received: 10 May 2021 | Accepted: 1 July 2021

DOI: 10.1002/asmb.2641

SPECIAL ISSUE PAPER

WILEY

Sample size determination to estimate mediation effects in cell transformation assays: A Bayesian causal model

Federico M. Stefanini[1] | **Alessandro Magrini**[2]

[1]Department of Environmental Science and Policy, University of Milan, Milan, Italy

[2]Department of Statistics, Computer Science, Applications, University of Florence, Florence, Italy

Correspondence
Federico M. Stefanini, Department of Environmental Science and Policy, University of Milan, Via Celoria 2, I-20133 Milan, Italy.
Email: federico.stefanini@unimi.it

Abstract

Cell transformation assays (CTAs) are in vitro methods used in the preliminary assessment of the carcinogenic potential of substances. CTAs are promising tests for cosmetic, food, and pharma companies because they are not only quick-and-cheap, but also able to reduce animal-based testing. An assay has the simple structure of a randomized one-way experiment, where the experimental factor is defined by 5 increasing concentrations. Different families of distributions have been proposed to evaluate the effect of a substance on counts of Type III foci, but all models proposed so far do not consider differences in the number of viable cells and in the total number of foci occurring among Petri dishes. In this article, a Bayesian structural causal model is proposed to distinguish total, direct, and indirect effects of a carcinogen in CTA experiments. The recommended sample size is calculated by Monte Carlo simulation given the type of effect and the magnitude to detect. An informative joint prior distribution on parameters elicited for BALB/c 3T3 CTAs is exploited to obtain the posterior distribution from each simulated dataset.

KEYWORDS

Bayesian SCM, causal inference, CTA, in vitro test, structural causal model

1 | INTRODUCTION

Cell transformation assays (CTAs) are in vitro methods used to screen chemical substances for carcinogenic potential while reducing animal-based testing.[1-3] In a typical CTA experiment (Figure 1), a preliminary step is performed with the aim of identifying five dose levels d_0, d_1, d_2, d_3, d_4 of a substance of interest characterized by an increasing degree of toxicity:[1(sec.4)] untreated negative control (d_0), highest nontoxic dose (d_1), median lethal dose (LD50, d_4) and two further intermediate dose levels, thus $d_0 < d_1 < d_2 < d_3 < d_4$. In the second step of a CTA, 10 Petri dishes for each dose level, $d_i \in \Omega_D$, are seeded with activated cells, BALB/c 3T3 in this work,[1] and after 24 h they are treated for 72 h according to the reference protocol. At 96 h from seeding (24 + 72), all Petri dishes are washed and the recovery step starts: it lasts for 5 weeks during which cells still alive-viable after treatment replicate, either with or without transformation. Transformed cells loose contact inhibition therefore they quickly replicate piling up within areas of a Petri dish with recognizable morphology, called foci. At the end of recovery, all Petri dishes are fixed and stained so that several foci eventually appear as dark-blue spots (Figure 1, right) and they are classified according to size and morphology into Type I, Type II, and Type

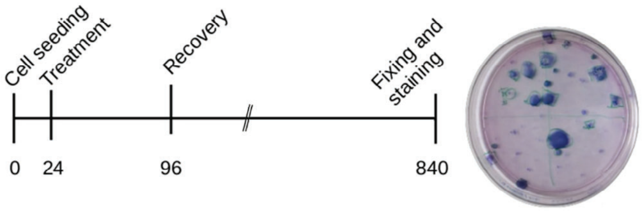

FIGURE 1 Time table of a CTA with an example of Petri dish after fixing and staining: dark-blue spots are foci, and some of them are fully transformed foci (also called Type III). Time is expressed in hours, with the end of the experiment after 5 weeks

III foci. The experimental outcome is the number of fully transformed (Type III) foci visually scored within each Petri dish at each dose level: this is the only class of foci that cause tumors when injected into rats.

The paramount importance of this class of experiments for cosmetic, food, and pharma companies is due to the possibility of reducing animal-based testing and the overall time required to perform each test: 5 weeks instead of 2 years are a substantial reduction of time in view of thousands chemicals yet to be tested.[4] CTAs are not designed to fully replace the 2-year cancer bioassay test, but they may be useful in preliminary screening for carcinogenicity. One way to support wider adoption of CTAs in the industry is the definition of a statistical model whose structure reflects the main features of a CTA, while accounting for differences of shape and variability that may occur in the marginal distributions of observable counts across different experiments: transformation into cancer is the result of several interacting mechanisms and complex multistep cell transitions involving specific signaling.[3]

In the literature on in vitro toxicology, the one-way structure underlying many assays has been recognized and contrasts of means have been considered in the general linear model framework or using nonparametric techniques.[5] An international expert group gathered at the European Centre for the Validation of Alternative Methods (ECVAM), formulated recommendations about the statistical analysis suited to the specific features of CTA (BALB/c 3T3 system).[6,7] In Hoffman et al.,[8] the most recent paper to our knowledge specifically dealing with the statistical methodology for BALB/c 3T3 CTA, the authors concluded that none of the approaches in use are suited to CTA data. They proposed two different approaches instead, one based on the negative binomial generalized linear model, and another resorting to the general linear model for Nishiyama-transformed counts.[8,9] A crucial point in the analysis of fully transformed foci pertains to the choice of a convenient family of probability distribution functions. In a recent work dealing with limitations of the Nishiyama transformation and with the presence of underdispersion at low doses, counts of Type III foci have been modeled through the discretization of continuous Beta latent variables to estimate causal effects while increasing model flexibility.[10] All models proposed so far do not consider differences in the number of viable cells and in the total number of foci occurring among Petri dishes.

We maintain that an improved class of models may be achieved from the perspective of causal analysis, in particular by developing a parametric Bayesian structural model that exploits protocol features and expert information. In CTAs, the typical sample size at each dose level is 10 Petri dishes, thus parametric models are a convenient option because they have the ability of extracting information from data very efficiently when the underlying assumptions, at least approximately, hold. Prior information may be exploited not only by eliciting the expert degree of belief, but also by recognizing specific features of the CTA protocol, like cell toxicity found at each dose level in the preliminary step: expert degree of belief about the number of cells still viable-alive after treatment can be elicited. It is worth noting that the dose-response relationship is not among the goals of the analysis, given the small number of dose levels and of dishes.

In this work, a parametric Bayesian structural causal model for BALB CTA experiments is proposed to improve the description of the data generating process by considering the total number of foci and the viability of treated cells besides the number of fully transformed foci. The proposed model enables mediation analysis, thus expressions to calculate (natural) direct and indirect effects are presented and exploited to determine the sample size required to reduce the expected uncertainty of estimated effects below a preassigned threshold, using Monte Carlo simulation.

This article is structured as follows. In Section 2, we detail the Bayesian structural causal model and the procedure for sample size determination. Results on sample size determination are reported and discussed in Section 3. Section 4 contains concluding remarks and considerations on potential improvements of CTAs based on our elaboration.

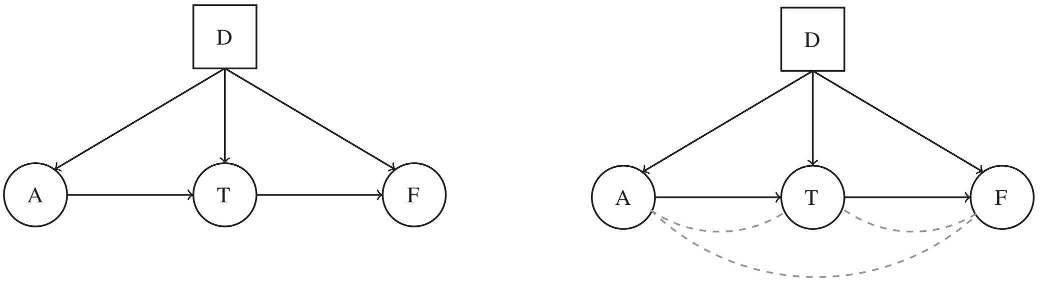

FIGURE 2 Causal DAGs for a CTA experiment. A circle around a letter denotes a random variable, while a square indicates a quantity under the control of the researcher (intervention variable). Left panel: DAG for (one batch) homogeneous experimental units. Right panel: mixed graph for (multibatch) heterogeneous experimental units

2 | A BAYESIAN STRUCTURAL CAUSAL MODEL FOR BALB CTA EXPERIMENTS

In this section, the proposed Bayesian structural causal model is detailed. We start from the definition of the causal directed acyclic graph (Subsection 2.1), then the structural causal model is specified in a Bayesian parametric formulation (Subsection 2.2), and the joint prior distribution on parameters is elicited for the class of CTA experiments based on BALB/c 3T3 cell lines[8] (Subsection 2.3). Afterward, mediation analysis is described (Subsection 2.4) and sample size determination is performed with the direct effect for a selected dose as target estimand (Subsection 2.5).

2.1 | Causal DAG and assumptions

A directed acyclic graph (DAG) is a collection of vertices (nodes) linked by directed edges (arrows) such that any sequence of equi-oriented edges (head-to-tail) does not visit the same node more than one time.[11] Nodes are labels associated to random variables evaluated on one experimental unit, here a Petri dish. We follow the common practice of avoiding the distinction between labels and variables, thus, in Figure 2, D is the variable "dose level" of a substance, A is the random variable "number of viable-alive cells" after treatment, T is the random variable "total number of foci", either fully transformed or not, and F is the random variable "number of fully transformed foci": four variables are assessed on each Petri dish included in a CTA.

DAGs are suited to represent a useful class of conditional independence relationships among random variables in a qualitative way, a feature that characterizes the class of graphical models,[12] but stronger relationships may be also represented, as it happens in nonparametric structural causal models (NP-SCMs).[11] Here, two nodes are linked by an arrow, say $D \rightarrow A$ (Figure 2), because the intervention that sets the manifest variable D equal to d determines a change in the distribution of A at least for some values in sample space Ω_D: this is an instance of direct causation at the considered granularity of representation. If the number of viable cells A after 72 h of treatment on a Petri dish is smaller than the nominal value at seeding time, then a smaller number T of transformed foci is expected at the end of the experiment ($A \rightarrow T$), and, in turn, a smaller number F of fully transformed (Type III) foci will be observed ($T \rightarrow F$). Oriented edges $D \rightarrow T$ and $D \rightarrow F$ represent the action of both toxicity and transformation processes because they are jointly determined by concentration and type of substance. In other terms, if the number of viable cells A is artificially made constant, then different concentrations of substance promote early steps in cell transitions with different probability, as well as they do with late steps toward full transformation. The square around D (Figure 2) emphasizes that this variable is under the control of the researcher: random assignment of Petri dishes to dose levels is always performed in CTAs.

The DAG in Figure 2, left panel, captures the hypothesized causal relationships among nodes D, A, T, F in a qualitative way: the four nodes suffice to represent one experimental unit of a typical CTA performed using just one batch of reagents and of cells. The causal DAG can be rewritten as a system of nonparametric structural equations, $A = h_A(d, U_A)$, $T = h_T(d, a, U_T)$, and $F = h_F(d, t, U_F)$, where each endogenous variable is a deterministic function, h, of its parent variables in the DAG and of one exogenous random variable, U, that may or may not be explicitly indicated in the DAG.[11(sec.1.4,sec.7)] The exogenous random variables U_A, U_T, and U_F, not shown in Figure 2, represent the action of all the other causes not explicitly considered in the model respectively on A, T, and F. For the intervention node D, the nonparametric equation $D = h_D(U_D)$ reduces to $D = d_i$, where d_i is the value fixed by manipulation.

The elicitation of deterministic functions and of the joint probability distribution on exogenous variables is a daunting task if the causal model does not stem from mechanistic explanations of cellular behavior at molecular level. Here we avoided this exercise by considering the conditional probability distributions induced on each endogenous variable,[11(sec.1.4.2)] as motivated hereafter. Firstly, our causal DAG (Figure 2, left panel) is Markovian, which means that all common causes of any pair of variables in the DAG, whether measured or not, are also included in the DAG. Instead, the mixed graph in Figure 2, right panel, is an example where some common causes of A, T, F are not explicitly considered, like the case of a lab protocol in which heterogeneous batches of cells are used in different Petri dishes. Secondly, modularity of functions holds, thus the intervention operated on a given variable does not change how variables, other than its children, behave. Finally, the factorization into the product of marginals follows from the Markovian condition on random variables U_A, U_T, U_F, that is, $p(u_D, u_A, u_T, u_F) = \prod_k p(u_k)$, and univariate conditional distributions are induced on endogenous variables D, A, T, F.[11(theorem1.4.1)] For example, $A(d_i) = h_A(d_i, U_A)$ is the notation for the random variable A induced by the exogenous variable U_A when the dose level of a chemical is equal to d_i.

2.2 | Parametric Bayesian specification

A carefully selected parametric model may improve the quality of inference when the sample size is small, as it happens for CTAs adopted in production. Expert degree of belief is a key ingredient to build effective statistical models,[13] and parametric Bayesian models are able to exploit such information in a quantitative way.

In Figure 2, left panel, random variables measured in one experimental unit (Petri dish) are shown. In what follows, index $i \in \{0, 1, 2, \dots\}$ refers to dose levels, with $d_i \in \Omega_D$ a specific dose, while index $j = 1, 2, \dots$ refers to Petri dishes treated at the same dose level i, thus $F_{i,j}$ is the random variable "number of fully transformed foci" in dish j at dose level i. In what follows, we consider $|\Omega_D| = 5$, because this is a very common number of dose levels selected in practice.

Parameters are indicated by Greek letters, for example, θ_A is the vector of parameters required to specify the conditional distribution of alive-viable cells. Constants, like the initial number of cells in each Petri dish, are indicated by letter c, for example, $c = 10^6$ is the typical number of cells initially seeded before treatment. All other parameters are unknown and included into the model as random variables, and $\theta = (\theta_A, \theta_T, \theta_F)$ is the vector including all the parameters.

The joint distribution of random variables given $D = d_i$ is factorized according to the DAG in Figure 2:

$$p(a_i, t_i, f_i | d_i, \theta) = \prod_j p(a_{i,j} | d_i, \theta_A) \cdot p(t_{i,j} | d_i, a_{i,j}, \theta_T) \cdot p(f_{i,j} | d_i, t_{i,j}, \theta_F), \tag{1}$$

where $a_i = (a_{i,1}, a_{i,2}, \dots)$, $t_i = (t_{i,1}, t_{i,2}, \dots)$, $f_i = (f_{i,1}, f_{i,2}, \dots)$ are vectors collecting variables at the same dose level. The joint distribution over doses is defined using (1):

$$p(a_{<v>}, t_{<v>}, f_{<v>} | d_{<v>}, \theta) = \prod_i p(a_i, t_i, f_i | d_i, \theta), \tag{2}$$

where v indicates that vectors over index i are considered, for example $a_{<v>} = (a_1, a_2, \dots)$.

The proposed Bayesian model consists of three hierarchical levels defined by the causal DAG defined in Subsection 2.1. The top level is made by the number of viable-alive cells that were modeled by choosing a smooth family of distributions because a sudden jump of probability value at subsequent counts a and $a + 1$ is unlikely for viable cells, whichever a, thus it is neither plausible that $P[a + 1] >> P[a]$, nor that $P[a + 1] << P[a]$. Another important feature to address is that counts are not larger than 1×10^6 cells, that is, the initial number of seeded cell within each Petri dish: at low or null dose levels, the distribution of A is mostly concentrated on counts values close to 1×10^6 cells. Accordingly, the following generalized logit transformation was adopted:

$$\lambda(x; c) = \ln\left(\frac{x}{c - x}\right), \tag{3}$$

where $\lambda(x; c)$ is the analogous of the *logit* transformation for values of x ranging between 0 and c, rather than for values ranging from 0 to 1, thus, for $c = 1$, (3) equates to the logit function. The inverse of (3) is:

$$\lambda^{-1}(l;c) = c \cdot \frac{\exp(l)}{1+\exp(l)} \tag{4}$$

with $l \in (-\infty, +\infty)$. By recognizing the large size of the sample space of A and the order of magnitude of the expected variability, a Normal likelihood function on the generalized logit scale was specified for the number of viable cells:

$$\lambda(A_{i,j}; c_A) \equiv \ln\left(\frac{A_{i,j}}{c_A - A_{i,j}}\right) \sim N(\mu_{A,i}, \sigma_{A,i}), \tag{5}$$

where $c_A = 1000$ thousands cells; parameters $\mu_{A,i} \in \mathbb{R}$ and $\sigma_{A,i} \in \mathbb{R}^+$ play the role, respectively, of mean and standard deviation of viable-alive cells on the generalized logit scale. It is worth noting that the choice of the Normal family of distributions for transformed counts is not new in the literature, for example, it has been used after the Nishiyama transformation,[8] but here alive-viable cells are modeled, instead of foci, and the generalized logit transformation guarantees the respect of boundaries.

The total number of foci on a Petri dish is naturally bounded by the size of a standard Petri dish, whose diameter is 10 cm. Besides eliciting the maximum value of T over dose levels and alive cells, some anchoring during elicitation is obtained from the comparison with dense packings of congruent circles in a circle,[14] although foci are only approximately circular, have different size and are well separated one from another, that is, scoring is performed only if they are not confluent into one composite focus. Another option is to check extreme quantiles of an approximating distribution, like the Poisson one, so that if the expected value of T at a given dose level is equal to 30 then the two quantiles 0.999 and 0.001 are respectively equal to 48 and 15. A rough calculation based on circular foci all of radius 0.67 cm and on the useful portion of a Petri dish from its center, which has radius equal to 4.75 cm, provides an estimate of the maximum number of foci equal to $4.75^2/0.67^2 = 50.26$. In this estimate, differences of size and the distance among foci are not taken into account. All things considered, including the evaluation of our expert, the sample space of T was defined as $\Omega_T = \{0, 1, \dots, 50\}$. At level two, the likelihood of the total number of transformed foci is defined as a function of the number of viable-alive cells through the following Normal kernel:

$$p_T(t|\mu_{T,i,j}, \sigma_{T,i}) \propto \exp\left\{-\frac{1}{2}\left(\frac{t-\mu_{T,i,j}}{\sigma_{T,i}}\right)^2\right\} \qquad t = 0, 1, \dots, 50, \tag{6}$$

where $\sigma_{T,i} \in \mathbb{R}^+$ is the scale parameter at dose level i and $\mu_{T,i,j}$ the mode. An additive decomposition of $\mu_{T,i,j}$ on the logit scale is introduced after considering its dependence on the number of alive cells:

$$\lambda(\mu_{T,i,j}, 50) = \lambda(A_{i,j}, c_A) + \lambda(\tau_i, 1) \tag{7}$$

with λ defined in (3). Parameter $\tau_i \in (0, 1)$ is the dose-dependent transformation rate, that is, the number of transformed cells over thousand viable cells. From (7), it follows that:

$$\mu_{T,i,j} = \frac{50 \cdot A_{i,j}\tau_i}{c_A(1-\tau_i) - A_{i,j}(1-2\tau_i)} \tag{8}$$

thus $\mu_{T,i,j}$ are entirely defined in terms of already defined quantities.

At the third level, the likelihood of the number of fully transformed foci is assumed to follow the binomial distribution with sample size equal to the number of fully transformed foci:

$$F_{i,j} \sim \text{Bin}(\phi_i, T_{i,j}), \tag{9}$$

where ϕ_i is a parameter representing the probability of full transformation at dose level i.

The proposed model is a flexible starting point open to refinement, especially as regards prior distributions. We do not exclude the possibility that specific classes of substances exist for which simpler models and stronger prior information work well: here a widely applicable model was developed by exploiting the common features of BALB/c 3T3 CTAs. Nevertheless, the small sample size of typical CTAs makes the investigation of more general families of distributions difficult outside specifically designed experiments. In our previous empirical investigations, the Poisson and the Negative Binomial families did not pass posterior predictive checks based on discrepancy measures. Secondly, here the quantitative concentration of the considered chemical was partitioned into discrete dose levels, in agreement with the recommendations from the literature, which recognizes the limited number of observations and of distinct doses levels.[8]

2.3 | Elicitation of an informative joint prior distribution on parameters

The proposed Bayesian causal model at dose level i depends on parameters $(\mu_{A,i}, \sigma_{A,i}, \tau_i, \sigma_{T,i}, \phi_i)$. The CTA protocol prescribes the presence of a negative control, $i = 0$, of one or few dose levels close to the no-observed-adverse-effect-level (NOAEL), and the last dose level just above the median lethal dose that kills 50% of cells (LD50). Here, we detail our procedure to elicit an informative joint prior distribution on parameters by considering a generic carcinogenic chemical under testing in the typical case of five dose levels ($d_i,\ i = 0, 1, 2, 3, 4$). We assume that each parameter is a priori independent of each other conditionally to the dose level, and the parameter vectors at different dose levels are a priori independent:

$$p(\mu_{A,i}, \sigma_{A,i}, \tau_i, \sigma_{T,i}, \phi_i) = p(\mu_{A,i})p(\sigma_{A,i})p(\tau_i)p(\sigma_{T,i})p(\phi_i) \quad i = 0, 1, 2, 3, 4. \tag{10}$$

The elicitation has been performed through interview to an expert with 15 years of experience in BALB/c 3T3 CTA tests of genotoxic and nongenotoxic substances, either sampled from the environment or from pure stock of chemicals.

2.3.1 | Prior distribution on $\mu_{A,i}$

At dose level $i = 0, \ldots, 4$, uncertainty on $\mu_{A,i}$ is represented by a Normal distribution:

$$\mu_{A,i} \sim N(m_{A,i}, s_{A,i}), \tag{11}$$

where the elicitation of hyperparameters $m_{A,i}$ and $s_{A,i}$ was performed by asking to the expert the following question: "Consider a large number of Petri dishes at the same dose level i. What are the first and the 99th percentiles of the mean number of viable-alive cells per dish?". Denote the requested percentiles as $a_{0.01,i}$ and $a_{0.99,i}$, respectively. We determined $m_{A,i}$ and $s_{A,i}$ by matching expected value and standard deviation of the Normal distribution with the first and 99th percentiles respectively equal to $\lambda(a_{0.01,i}, c_A)$ and $\lambda(a_{0.99,i}, c_A)$. For dose level d_0 (the negative control), the expert stated $a_{0.01,i} = 990$ and $a_{0.995,i} = 999.5$. In this case, $\lambda(a_{0.01,i}, c_A) = 4.595$ and $\lambda(a_{0.99,i}, c_A) = 7.6$, leading to $m_{A,i} = 6.098$ and $s_{A,i} = 0.646$. Values of hyperparameters of $m_{A,i}$ and $s_{A,i}$ at different dose levels were obtained in a similar way (Table 1).

2.3.2 | Prior distribution on $\sigma_{A,i}$

The uncertainty on parameters $\sigma_{A,i},\ i = 0, 1, 2, 3, 4$ is here described by uniform distributions:

$$\sigma_{A,i} \sim \mathrm{Unif}(u_{1,A,i}, u_{2,A,i}), \tag{12}$$

TABLE 1 Results of the elicitation of hyperparameters $m_{A,i}$ and $s_{A,i}$

Dose	$a_{0.01,i}$	$a_{0.99,i}$	$\lambda(a_{0.01,i}, c_A)$	$\lambda(a_{0.99,i}, c_A)$	$m_{A,i}$	$s_{A,i}$
d_0	990.0	999.5	4.595120	7.600402	6.097761	0.645923
d_1	980.0	999.5	3.891820	7.600402	5.746111	0.797082
d_2	800.0	840.0	1.386294	1.658228	1.522261	0.058446
d_3	640.0	680.0	0.575364	0.753772	0.664568	0.038345
d_4	475.0	525.0	−0.100084	0.100084	0.000000	0.043022

TABLE 2 Results of the elicitation of hyperparameters $u_{1,A,i}$ and $u_{2,A,i}$

Dose	$m_{A,i}$	$a_{0.5,i}$	$\delta_{0.01,i}$	$\delta_{0.99,i}$	$\lambda(a_{0.5,i} - \delta_{0.01,i}, c_A)$	$\lambda(a_{0.5,i} + \delta_{0.99,i}, c_A)$	$u_{1,A,i}$	$u_{2,A,i}$
d_0	6.097761	997.757	0.01	0.5	6.102240	6.350487	0.001925	0.108636
d_1	5.746111	996.815	0.01	1.0	5.749266	6.123950	0.001356	0.162417
d_2	1.522261	820.871	0.01	2.0	1.522329	1.535923	0.000029	0.005872
d_3	0.664568	660.286	0.01	3.0	0.664613	0.677971	0.000019	0.005762
d_4	0.000000	500.000	0.01	5.0	0.000040	0.020001	0.000017	0.008597

where $u_{1,A,i}$ and $u_{2,A,i}$ represent the minimum and the maximum plausible values of $\sigma_{A,i}$. The median number of viable-alive cells for dose level i given the elicited value $m_{A,i}$ is equal to:

$$a_{0.5,i} = \lambda^{-1}(m_{A,i}, c_A) = \frac{c_a \cdot \exp(m_{A,i})}{1 + \exp(m_{A,i})} \tag{13}$$

the elicitation of $\sigma_{A,i}$ was performed by asking to the expert the following question: "Consider the average of a large number of Petri dishes at the same dose level i. Based on your previous statements, the median number of alive-viable cells per dish is equal to $a_{0.5,i}$. At each dose level, how much should you increase (decrease) $a_{0.5,i}$ to reach the 99th (first) percentile for the average number of alive-viable cells?". The requested percentiles are $a_{0.99,i} = a_{0.5,i} + \delta_{0.99,i}$ and $a_{0.01,i} = a_{0.5,i} - \delta_{0.01,i}$, respectively. If we assume that $\lambda(A_{i,j}, c_A)$ follows the Normal distribution with mean $m_{A,i}$ and standard deviation $\sigma_{A,i}$, the minimum and the maximum plausible value of $\sigma_{A,i}$ are:

$$\begin{aligned} u_{1,A,i} &= \frac{\lambda(a_{0.5,i} - \delta_{0.01,i}, c_A) - m_{A,i}}{z_{0.01}}, \\ u_{2,A,i} &= \frac{\lambda(a_{0.5,i} + \delta_{0.99,i}, c_A) - m_{A,i}}{z_{0.99}}, \end{aligned} \tag{14}$$

where $z_{0.99}$ is the 99th percentile of the standard Normal distribution. The resulting values of hyperparameters $u_{1,A,i}$ and $u_{2,A,i}$ are shown in Table 2.

2.3.3 | Prior distribution on τ_i

The distribution on parameter τ_i, $i = 0, \dots, 4$ is assumed to follow a Normal distribution on the logit scale:

$$\lambda(\tau_i, 1) \equiv \log\left(\frac{\tau_i}{1 - \tau_i}\right) \sim N(m_{\tau,i}, s_{\tau,i}). \tag{15}$$

The elicitation of hyperparameters $m_{\tau,i}$ and $s_{\tau,i}$ at each dose level i was anchored to $a_{i,0.5}$, that is, the median number of viable-alive cells at dose level i obtained from the elicited value $m_{A,i}$ shown in (13). We asked to the expert the following question: "Consider a large number of Petri dishes at the same dose level i. Consider the case in which the number of alive-viable cells at dose level i is always equal to the median $a_{i,0.5}$, a value obtained in a previous step of the elicitation. What are the first and the 99th percentiles for the average total number of transformed foci per dish?". Denote the requested percentiles as $\mu_{0.01,T,i}$ and $\mu_{0.99,T,i}$, respectively. From (8), it follows that:

$$\tau_i = \frac{\mu_{T,i,j}(c_A - A_{i,j})}{50A_{i,j} + \mu_{T,i,j}(c_A - 2A_{i,j})} \tag{16}$$

with c_A the initial number of seeded cells. The first and 99th percentiles of the transformation rate τ_i are:

$$\tau_{0.01,i} = \frac{\mu_{0.01,T,i}(c_A - a_{0.5,i})}{50 \cdot a_{0.5,i} + \mu_{0.01,T,i}(c_A - 2 \cdot a_{0.5,i})},$$

TABLE 3 Results of the elicitation of hyperparameters $m_{\tau,i}$ and $s_{\tau,i}$

Dose	$\boldsymbol{m_{A,i}}$	$\boldsymbol{a_{0.5,i}}$	$\boldsymbol{\mu_{0.01,T,i}}$	$\boldsymbol{\mu_{0.99,T,i}}$	$\boldsymbol{\lambda(\tau_{0.01,i},1)}$	$\boldsymbol{\lambda(\tau_{0.99,i},1)}$	$\boldsymbol{m_{\tau,i}}$	$\boldsymbol{s_{\tau,i}}$
d_0	6.097761	997.757	0.1	1.2	−12.310367	−9.803170	−11.056769	0.538870
d_1	5.746111	996.815	0.5	1.6	−10.341231	−9.155608	−9.748419	0.254825
d_2	1.522261	820.871	5	13	−3.719486	−2.568230	−3.143858	0.247438
d_3	0.664568	660.286	12	23	−1.817247	−0.824911	−1.321079	0.213282
d_4	0.000000	500.000	22	38	−0.241162	1.152680	0.455759	0.299577

$$\tau_{0.99,i} = \frac{\mu_{0.99,T,i}(c_A - a_{0.5,i})}{50 \cdot a_{0.5,i} + \mu_{0.99,T,i}(c_A - 2 \cdot a_{0.5,i})} \tag{17}$$

thus $m_{\tau,i}$ and $s_{\tau,i}$ are obtained by matching expected value and standard deviation of the Normal distribution to the first and 99th percentiles, that is, $\lambda(\tau_{0.01,i},1)$ and $\lambda(\tau_{0.99,i},1)$. The resulting values of hyperparameters $m_{\tau,i}$ and $s_{\tau,i}$ are shown in Table 3.

2.3.4 | Prior distribution on $\sigma_{T,i}$

Uncertainty about $\sigma_{T,i}$ is here modeled through a uniform distribution:

$$\sigma_{T,i} \sim \text{Unif}(u_{1,T,i}, u_{2,T,i}),$$

where $u_{1,T,i}$ and $u_{2,T,i}$ represent the plausible lowest and highest values that $\sigma_{T,i}$ can take. The elicitation of hyperparameters $u_{1,T,i}$ and $u_{2,T,i}$ is anchored to the median value of τ and A at dose level i, values already elicited as $m_{A,i}$ and $m_{\tau,i}$:

$$t_{0.5,i} = \lambda^{-1}(m_{A,i} + m_{\tau,i}, 50) = \frac{50 \cdot \exp(m_{A,i} + m_{\tau,i})}{1 + \exp(m_{A,i} + m_{\tau,i})}. \tag{18}$$

The following question posed to the expert drove the elicitation: "Consider a large number of Petri dishes at the same dose level i. Given the median value of the number of transformed foci per dish $t_{i,0.5}$ elicited in a previous step, choose a non negative integer $d_{T,i}$ and consider the interval defined by counts $|t - t_{0.5,i}| \leq d_{T,i}$: this interval represents a plausible collection of counts for T (with probability above 0.75) that contains the median at dose level i. May you state the minimum $\pi_{T,i,min}$ and the maximum value $\pi_{T,i,max}$ that you may observe for the relative frequency of Petri dishes whose total number of foci is not larger (in absolute value) than $d_{T,i}$?". The elicited quantities refer to the probability:

$$P(|t - \mu_{T,i,j}| \leq d_{T,i}) = \pi_{T,i} \tag{19}$$

with $(\pi_{T,i,min}, \pi_{T,i,max})$ the elicited interval for $\pi_{T,i}$. The implied values of hyperparameters $u_{1,T,i}$ and $u_{2,T,i}$ are then obtained as the solution of:

$$\begin{aligned} u_{1,T,i} &= \arg_u \left(\sum_{t:|t-\mu_{T,i,j}| \leq d_{T,i}} p_T(t|t_{0.5,i}, u) = \pi_{T,i,max} \right), \\ u_{2,T,i} &= \arg_u \left(\sum_{t:|t-\mu_{T,i,j}| \leq d_{T,i}} p_T(t|t_{0.5,i}, u) = \pi_{T,i,min} \right) \end{aligned} \tag{20}$$

and they are shown in Table 4.

TABLE 4 Results of the elicitation of hyperparameters $u_{1,T,i}$ and $u_{2,T,i}$

Dose	$t_{0.5,i}$	$d_{T,i}$	$\pi_{T,i,min}$	$\pi_{T,i,max}$	$u_{1,T,i}$	$u_{2,T,i}$
d_0	0.35	1	0.95	0.99	0.562815	0.736778
d_1	0.90	1	0.85	0.99	0.364162	0.649950
d_2	8.25	1	0.80	0.95	0.778991	1.179497
d_3	17.08	2	0.80	0.95	1.305425	1.970565
d_4	30.60	3	0.85	0.95	1.708549	2.372781

TABLE 5 Results of the elicitation of hyperparameters $m_{F,i}$ and $s_{F,i}$

Dose	$\phi_{0.01,i}$	$\phi_{0.99,i}$	$\lambda(\phi_{0.01,i}, 1)$	$\lambda(\phi_{0.99,i}, 1)$	$m_{F,i}$	$s_{F,i}$
d_0	0.01	0.99	−4.595120	4.595120	0.000000	1.975251
d_1	0.01	0.99	−4.595120	4.595120	0.000000	1.975251
d_2	0.10	0.99	−2.197225	4.595120	1.198948	1.459873
d_3	0.15	0.99	−1.734601	4.595120	1.430259	1.360442
d_4	0.20	0.99	−1.386294	4.595120	1.604413	1.285580

2.3.5 | Prior distribution on ϕ_i

The Normal family of distributions is exploited to describe the uncertainty on $\phi_i,\ i = 0, \dots, 4$ on the logit scale:

$$\lambda(\phi_i, 1) \equiv \ln\left(\frac{\phi_i}{1-\phi_i}\right) \sim N(m_{F,i}, s_{F,i}). \tag{21}$$

Elicitation of hyperparameters $m_{F,i}$ and $s_{F,i}$ was performed by asking to the expert the following question: “Consider a large number of Petri dishes at the same dose level i all with the same value of observed total number of foci, say $T_{i,j}$ constant over i. What are the first and the 99th percentiles for the fraction of fully transformed foci per dish?”. Denote the requested percentiles as $\phi_{0.01,i}$ and $\phi_{0.99,i}$, respectively. Table 5 shows values of $m_{F,i}$ and $s_{F,i}$ computed by matching expected value and standard deviation of the Normal distribution to the first and 99th percentiles, respectively computed as $\lambda(\phi_{0.01,i}, 1)$ and $\lambda(\phi_{0.99,i}, 1)$.

2.3.6 | Revision of the elicited joint prior distribution on parameters

The final step of the elicitation consisted in the inspection of the marginal distribution of the manifest variables A, T, and F implied by the elicited joint prior distribution on parameters. At this purpose, marginal distributions were approximated by a random sample of size 10,000. The expert did not find any implausible feature neither in graphical nor in numerical summaries (Figure 3 and Table 6).

2.4 | Mediation analysis

Mediation analysis in a CTA experiment aims at evaluating the way a chemical determines changes in the outcome, that is, the number of fully transformed foci. Changes in the outcome may be determined through different paths: (1) increase/decrease of the number of alive-viable cells and of the total number of foci, which are mediating variables determining the indirect effect; (2) increase/decrease in the number of fully transformed foci given the total number of foci in a Petri dish, which represents the direct effect of a chemical on F. Here we denote the reference dose as d_0 because negative control (water) is the typical choice to declare a chemical carcinogenic, but a foregoing dose d_{i-1} as reference for d_i could also be of interest, for example, in the analysis of turning points.

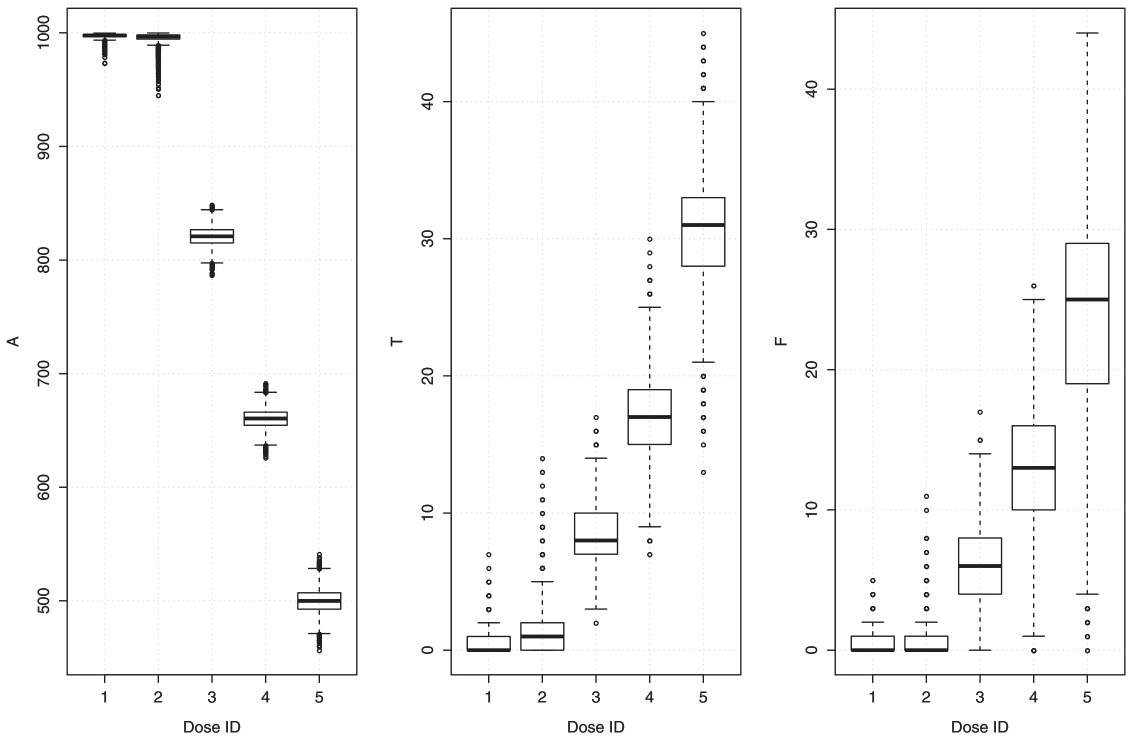

FIGURE 3 Distribution of a random sample of size 10,000 from the elicited joint prior distribution on parameters

The total effect (TE) of dose $d_i, i > 0$, of a substance with respect to d_0 is the combined result of all paths from D to F:[11,15,16(sec.4.5)]

$$\text{TE}(d_0, d_i) = \text{E}[h_F(d_i, h_T(d_i, h_A(d_i, u_A), u_T), u_F)] - \text{E}[h_F(d_0, h_T(d_0, h_A(d_0, u_A), u_T), u_F)], \tag{22}$$

where deterministic functions h_A, h_T, h_F were introduced in Section 2. The natural direct effect (NDE) and the natural indirect effect (NIE) at dose d_i are defined as:

$$\text{NDE}(d_{0,}d_i) = \text{E}[h_F(d_i, h_T(d_0, h_A(d_0, u_A), u_T), u_F)] - \text{E}[h_F(d_0, h_T(d_0, h_A(d_0, u_A), u_T), u_F)], \tag{23}$$

$$\text{NIE}(d_0, d_i) = \text{E}[h_F(d_0, h_T(d_i, h_A(d_i, u_A), u_T), u_F)] - \text{E}[h_F(d_0, h_T(d_0, h_A(d_0, u_A), u_T), u_F)], \tag{24}$$

where, for example, $h_F(d_0, h_T(d_i, h_A(d_i, u_A), u_T), u_F)$ cannot be experimentally measured because it pretends that treatment d_i is active in paths mediated by T, while d_i is not active along path $D \rightarrow F$.

The above equations are formulated in terms of nonparametric individual components, but our interest is on population averaged effects, therefore estimates may be equivalently obtained using Pearl's formulas.[16-18] The DAG in Figure 2, left panel, satisfies the conditions for identifying natural effects,[17] as the treatment is randomized and confounding of the relationship between the mediators and the outcome is absent. In a CTA experiment, mediation effects implied by our Bayesian structural causal model take the following forms (see the proof in the Appendix):

$$\text{NDE}(d_{0,}d_i) = (\phi_i - \phi_0) \cdot \text{E}[T|D = d_0], \tag{25}$$

$$\text{NIE}(d_0, d_i) = \phi_0 \cdot \{\text{E}[T|D = d_i] - \text{E}[T|D = d_0]\}, \tag{26}$$

$$\text{TE}(d_0, d_i) = \phi_i \cdot \text{E}[T|D = d_i] - \phi_0 \cdot \text{E}[T|D = d_0]. \tag{27}$$

In (25), NDE is null either if the fraction of fully transformed foci at each of the two dose levels is equal, or when the average of the total number of foci is null at the reference dose level. The average of T in CTAs is typically above zero and

TABLE 6 Quantiles of the distribution of a random sample of size 10,000 from the elicited joint prior distribution on parameters

Number of viable-alive cells (A)

Dose	0%	25%	50%	75%	100%
d_0	973.3	996.6	997.8	998.6	999.8
d_1	945.1	994.5	996.8	998.2	999.9
d_2	786.8	815.0	820.9	826.7	848.7
d_3	626.1	654.5	660.5	666.1	691.7
d_4	456.5	492.7	500.0	507.0	541.3

Total number of transformed foci (T)

Dose	0%	25%	50%	75%	100%
d_0	0	0	0	1	7
d_1	0	0	1	2	14
d_2	2	7	8	10	17
d_3	7	15	17	19	30
d_4	13	28	31	33	45

Number of fully transformed foci (F)

Dose	0%	25%	50%	75%	100%
d_0	0	0	0	1	5
d_1	0	0	0	1	11
d_2	0	4	6	8	17
d_3	0	10	13	16	26
d_4	0	19	25	29	44

below 6 when the reference dose is water. The value of NIE in (26) is null either when the fraction of fully transformed foci is null in the reference dose, or when the expected values $\text{E}[T|D = d_i]$ and $\text{E}[T|D = d_0]$ are equal, whatever the value of ϕ_0. With d_0 set to the negative control, ϕ_0 is often not null and the difference in the expected value of T is large at high dose levels. Large expected values of T in the negative control and at large dose levels entail the following benefits: (i) improved estimates of ϕs; (ii) large values of NDE; (iii) a null NIE in a revised CTA protocol where $\text{E}[T|D = d_i] = \text{E}[T|D = d_0]$, because TE is equal to NDE, thus the total effect also describes the direct effect of a chemical on the number of Type III foci.

In a CTA experiment, NDE describes the ability of a chemical to promote all transformation steps toward full carcinogenicity, while NIE is mostly about the ability of a chemical to activate only some (initial) steps of transformation that increase the total number of foci.

2.5 | Sample size determination

In CTAs, a chemical is declared to be carcinogenic when the number of Type III foci, here called number of fully transformed foci, is large at dose $d_i,\ i > 0$, and small at d_0 (the negative control), as a result of a deeply activated transformation. The interpretation is that the tested chemical was able to increase the number of fully transformed foci beyond the level naturally present in BALB/c 3T3 cells. On these grounds, toxicologists are not currently interested in the total number of foci because they do not cause tumors when injected into mice, a phenomenon possibly due to the occurrence of a limited set of modifications in cells that originate these type of foci. From the joint prior distribution on parameters elicited in Subsection 2.3, the implied distributions of TE, NDE, and NIE with respect to d_0 shown in Table 7 appear coherent with

TABLE 7 Mean and quantiles of TE and NDE with respect to dose d_0 implied by the elicited joint prior distribution on parameters (Monte Carlo approximation based on 100,000 draws)

Total effect (TE)						
Dose	**Mean**	**2.5%**	**25%**	**50%**	**75%**	**97.5%**
d_1 vs. d_0	0.323	−0.776	−0.105	0.144	0.556	2.446
d_2 vs. d_0	5.602	0.926	3.977	5.673	7.241	10.226
d_3 vs. d_0	12.473	3.391	9.963	12.970	15.361	19.434
d_4 vs. d_0	23.318	8.260	19.582	24.281	27.990	33.491
Natural direct effect (NDE)						
Dose	**Mean**	**2.5%**	**25%**	**50%**	**75%**	**97.5%**
d_1 vs. d_0	−0.03	−0.631	−0.152	−0.001	0.149	0.617
d_2 vs. d_0	0.119	−0.366	−0.033	0.090	0.242	0.781
d_3 vs. d_0	0.143	−0.304	−0.013	0.109	0.260	0.803
d_4 vs. d_0	0.162	−0.258	0.000	0.124	0.271	0.843
Natural indirect effect (NIE)						
Dose	**Mean**	**2.5%**	**25%**	**50%**	**75%**	**97.5%**
d_1 vs. d_0	0.329	−0.534	−0.002	0.120	0.461	2.295
d_2 vs. d_0	3.916	0.152	1.571	3.702	5.903	9.358
d_3 vs. d_0	8.317	0.332	3.438	8.127	12.781	18.144
d_4 vs. d_0	15.015	0.606	6.238	14.789	23.244	31.812

the above described practice: mean values of NDE always close to zero, large mean values of NIE, and nonoverlapping 95% credibility intervals for TE at all dose levels, $i > 0$.

Nevertheless, we conjecture that NDE should be considered instead. Why should TE be recommended as causal estimand if the only feature of interest is the increase in the number of Type III foci, whatever the total number of foci? A large value of TE may be due to a large difference between the average values of T at d_i and d_0 even if the difference $\phi_i - \phi_0$ is small. A small difference $\phi_i - \phi_0$ indicates that the considered chemical is not effective in promoting the complete transformation of foci, thus such a chemical could be declared carcinogenic because it is strong in enhancing the starting steps of transformation, although weak in completing all the essential steps toward full carcinogenicity. Thus, from now on, NDE will be the causal estimand and the target quantity considered in sample size determination.

The optimal sample size in a CTA experiment is here defined as the minimum number of Petri dishes that must be collected at the considered dose level so that the expected value of an objective function quantifying uncertainty is equal or below the threshold defined by the toxicologist. In order to illustrate the approach, we consider an expected value of T at dose levels d_0 equal to 4 and a difference of fractions $\phi_3 - \phi_0 = 0.125$, a small but relevant value. Then, from (25), we obtain the target value NDE $= 0.5$. Here, we determine the expected value of the width of the 95% credibility interval for NDE by Monte Carlo simulation.

Synthetic datasets may be simulated from a parametric model after assigning a numerical value to all elements of the vector of model parameter θ. The selected target quantity is NDE $= 0.5$, which is a function of θ and of the probability distribution of T, which also depends on θ. In other terms, several distinct values of θ may lead to the value of NDE specified by the expert. The collection of θs leading to NDE $= 0.5$ was found by means of a preliminary Monte Carlo simulation in which 1×10^5 parameter values were randomly drawn from the joint prior distribution elicited in Subsection 2.3, then the value of *NDE* was calculated for each sampled θ: 10 vector values $\{\theta^*_1, \dots, \theta^*_{10}\}$ produced an NDE value in the interval 0.5 ± 10^{-4}, thus they were further exploited in the simulation of synthetic datasets.

A Monte Carlo simulation was performed as detailed in Algorithm 1 for values of sample size $n = 10, 20, 30, 50$ at dose level d_3 with reference d_0. Specifically, we simulated 100 synthetic datasets for each selected parameter value θ^*_k, $k =$

1, … , 10, and we performed a MCMC simulation for each generated dataset and the prior distribution in Subsection 2.3 to obtain the corresponding posterior distribution of model parameters and the implied value of NDE.

We computed the following three statistics for NDE with respect to datasets: (i) the average posterior mean, denoted as $\overline{m}$; (ii) the percentage of 95% posterior credibility intervals not containing value 0, denoted as $\overline{z}$; (iii) the average width of 95% posterior credibility intervals, denoted as $\overline{w}$. Statistics $\overline{m}$ and $\overline{z}$ were computed to check the validity of the results obtained by running Algorithm 1: as the sample size increases, $\overline{m}$ should converge to 0.5 and $\overline{z}$ should converge to 100 to indicate, respectively, consistency of the estimation and an increasing statistical power (i.e., Lindley's test). The optimal sample size for a given dose level is identified by comparing the value of statistic $\overline{w}$ at several different sample size values with a threshold defined by the toxicologist.

Algorithm 1. Computation of statistics $\overline{m}$, $\overline{z}$, and $\overline{w}$ for the predicted natural direct effect (NDE) in a CTA experiment

```
Input:

• target value of NDE, e.g., NDE = 0.5;
• $n$: the considered sample size;
• $d_i, i>0$: the considered dose level;
• $J$: the number of synthetic datasets to be simulated at the considered sample size $n$ for eachparameter value $\theta_k$, e.g., $J = 100$.

Steps:

1. Select values $\theta_1^*, \dots, \theta_k^*, \dots, \theta_K^*$ of model parameters among the $1 \times 10^5$ values sampled from the joint prior distribution that provide the specified target value of NDE, e.g., NDE = 0.5;
2. Initialize M, Z and W as empty matrices with $K$ rows and $J$ columns.
3. For $k = 1, \dots, K$:
4.     For $j = 1, \dots, J$:
         • simulate a dataset $\mathcal{D}_{k,j}$ of size $n$ given $\theta_k^*$;
         • approximate the posterior distribution $p(\theta \mid \mathcal{D}_{k,j})$ byMarkov Chain Monte Carlo (MCMC) simulation with the prior distribution elicited in Subsection 2.3;
         • calculate the NDE for the current posterior distribution;
         • set M in position $(k,j)$ as the mean of NDE$(d_0, d_i)$ with respect to $p(\theta \mid \mathcal{D}_{k,j})$;
         • compute the 95% credibility interval for NDE$(d_0, d_i)$ with respect to $p(\theta \mid \mathcal{D}_{k,j})$, denoted as $\mathcal{I}_{k,j} = (I_{1,k,j}, I_{2,k,j})$;
         • set Z in position $(k,j)$ as 1 if $\mathcal{I}_{k,j}$ does not contain value 0, otherwise as 0;
         • set W in position $(k,j)$ as $I_{2,k,j} - I_{1,k,j}$.

Output:

• $\overline{m}$ as the average of M;
• $\overline{z}$ as the average of Z;
• $\overline{w}$ as the average of W;
```

3 | RESULTS

In the elicitation of the joint prior distribution on parameters performed in Subsection 2.3, we considered five dose levels, the minimum number recommended for CTAs, but d_2 and d_3 mostly provide information on the carcinogenic potential: d_1 is a dose selected to bear almost no effect; toxicity at d_4 kills half of the initial cells, and it may dominate and even interfere with the carcinogenic transformation. For these reasons, we illustrate our method for sample size determination

TABLE 8 Results of sample size determination for the natural direct effect (NDE) of dose d_3 with respect to dose d_0 (the negative control). The specified target value of NDE is 0.5, and the results are based on 1000 synthetic datasets (100 for each of the 10 parameter values implying NDE $= 0.5 \pm 10^{-4}$)

Statistic $\overline{m}$					
n	**Mean**	**SD**	**Median**	**2.5th perc.**	**97.5th perc.**
10	0.404	0.124	0.391	0.188	0.673
20	0.445	0.098	0.438	0.266	0.651
30	0.459	0.083	0.460	0.299	0.629
50	0.475	0.066	0.472	0.355	0.609
Statistic $\overline{z}$					
n	**Mean**	**SD**	**Median**	**2.5th perc.**	**97.5th perc.**
10	97.7	15.0	100.0	100.0	100.0
20	99.9	3.2	100.0	100.0	100.0
30	100.0	0.0	100.0	100.0	100.0
50	100.0	0.0	100.0	100.0	100.0
Statistic $\overline{w}$					
n	**Mean**	**SD**	**Median**	**2.5th perc.**	**97.5th perc.**
10	0.475	0.076	0.467	0.346	0.644
20	0.367	0.055	0.362	0.275	0.476
30	0.310	0.046	0.308	0.229	0.393
50	0.247	0.037	0.248	0.183	0.309

Note: Statistic $\overline{m}$: average posterior mean. Statistic $\overline{z}$: percentage of 95% posterior credibility intervals not containing value 0. Statistic $\overline{w}$: average width of 95% posterior credibility intervals.

by focusing on dose level d_3 only. We investigated the following sample size values: $n = 10, 20, 30, 50$, because $n = 10$ is the typical sample size of a CTA experiment and $n = 50$ is currently considered a large sample size.

A Markov chain Monte Carlo (MCMC) simulation was performed on each synthetic dataset generated as described in Algorithm 1 using the Stan software with R,[19] by means of the `rstan` package.[20,21] In particular, a single chain was run for each synthetic dataset, discarding the first 20,000 draws and keeping the following 40,000 thinned by 4. The Geweke's convergence diagnostic was always satisfactory.[22] Algorithm 1 was executed on the Google Cloud Platform by defining an instance in Compute Engine (https://cloud.google.com/compute/) with C2-standard-16 machine type made by 16 virtual CPUs, 64 GB of RAM and a permanent disk of 100 GB (maximum used space 96% at sample size 50). By running 10 Markov chains in parallel (one for each virtual CPU), the execution of Algorithm 1 took, at $n = 10, 20, 30$, and 50, respectively, $2^h\ 4^m$, $5^h\ 1^m$, $7^h\ 37^m$, and $13^h\ 20^m$ (data available on request from the authors).

Results obtained from Algorithm 1, were checked for convergence of $\overline{m}$ to the specified target value NDE $= 0.5$. Since the joint prior distribution on model parameters is highly informative and the specified target value NDE $= 0.5$ is little likely a priori, bias is expected and indeed present, but it definitely decreases as the sample size increases (Table 8). Therefore, we recommend a larger number of synthetic datasets, say 1000, to improve the precision of sample size determination when the specified target value for NDE is a priori very unlikely, say less than our value of 0.5. Furthermore, longer MCMC simulations could be needed in case of bad Markov chain mixing. Interestingly, our results show that $n = 50$, which is currently considered a very large sample size in CTA experiments, does not guarantee an unbiased estimate of NDE when its true value is 0.5 (Table 8).

For what concerns the statistic $\overline{z}$, the mean value for $n = 10$ is 97.7, which increases to 99.9 for $n = 20$, and is equal to 100 for $n = 30$ or more. These findings indicate a very high chance of getting a significant estimate when the true NDE value is 0.5, even with a sample size typically adopted in CTA experiments. The statistic $\overline{w}$, that is, the average width of 95% posterior credibility intervals, is our objective function for sample size determination: the higher is the statistic, the higher is uncertainty in the estimate of the considered mediation effect at a given sample size. As expected, the mean of $\overline{w}$ decreases monotonically as the sample size increases. After eliciting from the toxicologist a maximum uncertainty on NDE equal to 1/3, we calculated a sample size equal to 27 by linear interpolation, which is almost three times the minimum size recommended in the literature and generally adopted in practice. It is worth noting that the sample size found in this work depends not only on the selected effect size but also on the type of effect, that is, NDE, NIE or TE, thus it is specific for such target.

4 | CONCLUDING REMARKS

CTAs are relatively cheap and fast tests when compared with in vivo tests to perform the preliminary screening on carcinogenicity of compounds. In order to find the minimum sample size to adopt in a CTA experiment and to overcome the limitations of some statistical approaches proposed in the literature, we developed a causal DAG which improves the description of the data-generating process and supports mediation analysis. A parametric Bayesian causal model was built by exploiting the main features of the BALB/c CTA protocol, then an informative joint prior distribution on parameters was elicited from an expert. Afterward, we derived the expression of mediation effects for CTAs and performed Monte Carlo simulations to determine the sample size required to reduce the expected uncertainty of estimates below a preassigned threshold. Our algorithm may accommodate several different choices of target estimand, but we provided some arguments in favor of the natural direct effect (NDE). Sample size determination was illustrated for a target value of NDE equal to 0.5.

Users of the proposed model should be aware of some limitations. The current model formulation does not address the case in which the same CTA is replicated with different batches of cells and/or reagents, for example, in different laboratories. Similarly, a protocol extension in which Petri dishes are seeded using different batches of cells and/or reagents is not currently covered. The mixed graph in Figure 2, right panel, accounts for the inherent heterogeneity of cell lines and serum that may determine confounding between mediators and the outcome. Another limitation pertains to the transition of a cell toward cancer, a change of state that was implicitly assumed always to pass through non-Type III foci: this feature has not yet been confirmed by experiments focused on measurements at molecular level, thus it would be interesting to extend the proposed model to also include these events. Last, despite that we elicited an informative prior distribution on parameters that should encompass the whole set of substances that may be tested, refinements of the proposed prior distributions might be considered for chemicals belonging to specific classes of compounds, for example, genotoxic substances.

In the last few years, the improvement of the basic BALB CTA protocol lost momentum. One direction of research dealt with the automation of visual scoring to reduce subjectivity during the attribution of foci to Type I, II, III classes.[23,24] Other authors modified the basic BALB CTA by adding substances of therapeutic interest, then they combined the BALB CTA with several endpoint applications for protein analysis as a tools to elucidate cancer mechanism at higher resolution.[3] Current interest is mostly devoted toward the creation of a panel of tests covering multiple biological traits to be jointly considered as an integrated approach for the testing and assessment (IATA) of chemical nongenotoxic carcinogens.[25]

We envision several areas of future research related to Balb CTA. Firstly, the code implementing our Monte Carlo algorithm is not currently optimized neither for speed nor for memory usage, thus larger simulations may benefit from improvements in these directions. Secondly, in our elicitation with the toxicologist, we found that NDE is often very small and even null at each dose level. A revision of the CTA protocol could be considered by toxicologists to obtain a very high T and a very low F in the negative control (water), in order to improve the estimate of NDE and therefore the declaration of carcinogenicity. Thirdly, a model with higher level of granularity where single cells are experimental units deserves attention because molecular features (e.g., mRNA and metabolites) and inter/intracellular signals are the fundamental determinants of cell survival, transformation and complete transformation. This level of description still represents a challenge for the available lab techniques. Fourthly, the proposed Bayesian model could be also of interest outside the considered causal framework, especially after marginalization with respect to T and A in order to obtain a flexible class of probability distributions for F. Such marginalized model is potentially able to explain atypical samples of counts where the within-dose sample variance of F is not only substantially smaller than the mean at the same dose, but sometimes even null.[10] Indeed, specifically designed experiments involving dozens of known carcinogens are needed to evaluate the utility of this marginalized model, as well as the generality of the family of distributions employed in the proposed causal model. As fifth and last place, elicitation from a panel of experts seems a natural way to support IATA, an increasing line of research,[25] and it seems also useful to gather beliefs maturated in experiments on widely different classes of substances.

Finally, we offer a new challenge to the toxicologist: is the magnitude of NDE correlated with any useful feature pertaining carcinogenicity in humans? We conjecture that NDE is a better candidate than TE in this sense, as NDE better describes the ability of a chemical to cause deep alterations of cell metabolism, not just enhanced replication. If this should not be the case, then "able to deeply alter metabolism" is the event on which causal inference could alternatively be directed.[10]

ACKNOWLEDGMENTS

The authors thank Chiara Urani (University of Milano Bicocca, Italy) for several stimulating discussions on CTAs. The authors also thank Raffaella Corvi (Institute for Health and Consumer Protection, JRC, Ispra, Italy), the reviewers, and the guest editor for their careful reading of the manuscript and their constructive criticism. The authors acknowledge the financial support provided by the "Dipartimenti Eccellenti 2018-2022," ministerial funds, Italy.

DATA AVAILABILITY STATEMENT

Data are available upon request from the authors.

ORCID

Federico M. Stefanini https://orcid.org/0000-0003-4248-6275
Alessandro Magrini https://orcid.org/0000-0002-7278-5332

REFERENCES

1. Sasaki K, Bonhenberger S, Hayashi K, et al. Recommended protocol for the BALB/c 3T3 cell transformation assay. *Mutat Res*. 2012;744:30-35. https://doi.org/10.1016/j.mrgentox.2011.12.014
2. Corvi R, Aardema MJ, Gribaldo L, et al. ECVAM prevalidation study on *in-vitro* cell transformation assays: general outline and conclusion of the study. *Mutat Res*. 2012;744:12-19. https://doi.org/10.1016/j.mrgentox.2011.11.009
3. Poburski D, Thierbach R. Improvement of the BALB/c-3T3 cell transformation assay: a tool for investigating cancer mechanisms and therapies. *Sci Rep*. 2016;6(32966):1-8. https://doi.org/10.1038/srep32966
4. Vanparys P, Corvi R, Aardema MJ, et al. Application of *in-vitro* cell transformation assays in regulatory toxicology for pharmaceuticals, chemicals, food products and cosmetics. *Mutat Res*. 2012;744(1):111-116. https://doi.org/10.1016/j.mrgentox.2012.02.001
5. Bretz F, Hothorn LA. Statistical analysis of monotone or non-monotone dose-response data from in vitro toxicological assays. *Altern Lab Anim ATLA*. 2003;1(Suppl 32):81-96. https://doi.org/10.1177/026119290303101s06
6. EURL ECVAM Validation Management Team . BALB/c 3T3 cell transformation assay pre-validation study report; 2010.
7. Creton S, Aardema MJ, Carmichael PL, et al. Cell transformation assays for prediction of carcinogenic potential: state of the science and future research needs. *Mutagenesis*. 2012;27(1):93-101. https://doi.org/10.1093/mutage/ger053
8. Hoffmann S, Hothorn LA, Edler L, et al. Two new approaches to improve the analysis of BALB/c 3T3 cell transformation assay data. *Mutat Res Genet Toxicol Environ Mutagen*. 2012;744(1):36-41. https://doi.org/10.1016/j.mrgentox.2011.12.002
9. Nishiyama H, Omori T, Yoshimura I. A composite statistical procedure for evaluating genotoxicity using cell transformation assay data. *Environmetrics*. 2002;14:183-192. https://doi.org/10.1002/env.575
10. Stefanini FM. Causal analysis of cell transformation assays, pp 949 - 954, In: Statistics and Data Science: New Challenges, New Generations.Proceedings of the Conference of the Italian Statistical Society. Eds: Alessandra Petrucci, Rosanna Verde ; June 28–30; 2001; Florence, IT.
11. Pearl J. *Causality: Models, Reasoning, and Inference*. 2nd ed. NY, USA: Cambridge University Press; 2009.
12. Koller D, Friedman N. *Probabilistic Graphical Models. Principles and Techniques*. Cambridge, MA: The MIT Press; 2009.
13. Vining GG. Towards a foundational theory of statistical engineering. Paper presented at: Proceedings of the Statistics and Innovation for Industry 4.0 (StEering Workshop 2020); February 20–21; 2020:411-420; Florence (IT).
14. Graham RL, Lubachevsky BD, Nurmela KJ, Ostergard PRJ. Dense packings of congruent circles in a circle. *Discret Math*. 1998;181(1-3):139–154. https://doi.org/10.1016/S0012-365X(97)00050-2
15. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*. 1992;3(2):143-155. https://doi.org/10.1097/00001648-199203000-00013
16. Pearl J. Direct and indirect effects. In: Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence (UAI 2001). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. ; 2001: 411-420.
17. Pearl J. Interpretation and identification of causal mediation. *Psychol Methods*. 2014;19(4):459-481. https://doi.org/10.1037/a0036434
18. Pearl J. Causal and counterfactual inference. In: Knauff M, Spohn W, eds. *Handbook of Rationality*. Cambridge, MA: MIT Press; 2019.
19. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2017. https://www.R-project.org/.
20. Carpenter B, Gelman A, Hoffman DM, et al. Stan: a probabilistic programming language. *J Stat Softw*. 2017;76(1):1–32. https://doi.org/10.18637/jss.v076.i01
21. Stan Development Team RStan: the R interface to stan. R package version 2.21.2; 2020. http://mc-stan.org/.

22. Geweke J. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In: Bernardo JM, Berger JO, Dawid AP, Smith AFM, eds. *Bayesian Statistics*. Vol 4. Oxford, UK: Clarendon Press; 1992;169–193.
23. Callegaro G, Malkoc K, Corvi R, Urani C, Stefanini FM. A comprehensive statistical classifier of *foci* in the cell transformation assay for carcinogenicity testing. *Toxicol In Vitro*. 2017;45(3):351-358. https://doi.org/10.1016/j.tiv.2017.04.030
24. Urani C, Corvi R, Callegaro G, Stefanini FM. Objective scoring of transformed *foci* in BALB/c 3T3 cell transformation assay by statistical image descriptors. *Toxicol In Vitro*. 2013;27(6):1905-1912. https://doi.org/10.1016/j.tiv.2013.06.011
25. Jacobs MN, Colacci A, Corvi R, et al. Chemical carcinogen safety testing: OECD expert group international consensus on the development of an integrated approach for the testing and assessment of chemical non-genotoxic carcinogens. *Arch Toxicol*. 2020;94:2899-2923. https://doi.org/10.1007/s00204-020-02784-5

How to cite this article: Stefanini FM, Magrini A. Sample size determination to estimate mediation effects in cell transformation assays: A Bayesian causal model. *Appl Stochastic Models Bus Ind*. 2021;37:973-989. doi: 10.1002/asmb.2641

APPENDIX 1. COMPUTATION OF MEDIATION EFFECTS

The natural direct effect (NDE) of dose level d_i with respect to dose level d_0 is:

$$\begin{aligned}\mathrm{NDE}(d_{0,}d_i) &= \sum_t \int_a \{\mathrm{E}[F|D=d_i, A=a, T=t] - \mathrm{E}[F|D=d_0, A=a, T=t]\} \cdot p(a,t|D=d_0) \cdot \mathrm{d}a \\ &= \sum_t (\phi_i - \phi_0) \cdot t \int_a p(t|D=d_0, A=a) \cdot p(a|D=d_0) \cdot \mathrm{d}a \\ &= \phi_i \sum_t t \cdot p(t|D=d_0) - \phi_0 \sum_t t \cdot p(t|D=d_0) \\ &= \phi_i \mathrm{E}[T|D=d_0] - \phi_0 \mathrm{E}[T|D=d_0] \\ &= (\phi_i - \phi_0)\mathrm{E}[T|D=d_0]. \end{aligned} \tag{A1}$$

The natural indirect effect (NIE) of dose level d_i with respect to dose level d_0 is:

$$\begin{aligned}\mathrm{NIE}(d_0, d_i) &= \sum_t \int_a \mathrm{E}[F|D=d_0, A=a, T=t] \cdot \{p(a,t|D=d_i) - p(a,t|D=d_0)\} \cdot \mathrm{d}a \\ &= \sum_t \phi_0 \cdot t \int_a \{p(t|D=d_i, A=a) \cdot p(a|D=d_i) - p(t|D=d_0, A=a) \cdot p(a|D=d_0)\} \cdot \mathrm{d}a \\ &= \sum_t t \cdot \{p(t|D=d_i) - p(t|D=d_0)\} \\ &= \phi_0 \{\mathrm{E}[T|D=d_i] - \mathrm{E}[T|D=d_0]\}. \end{aligned} \tag{A2}$$

The total effect (TE) of dose level d_i with respect to dose level d_0 is:

$$\begin{aligned}\mathrm{TE}(d_0, d_i) &= \mathrm{E}[F|do(D=d_i)] - \mathrm{E}[F|do(D=d_0)] \\ &= \sum_t \int_a \phi_i \cdot t \cdot p(t|D=d_i, A=a) \cdot p(a|D=d_i) \cdot \mathrm{d}a \\ &- \sum_t \int_a \phi_0 \cdot t \cdot p(t|D=d_0, A=a) \cdot p(a|D=d_0) \cdot \mathrm{d}a \\ &= \phi_i \cdot \sum_t t \cdot p(t|D=d_i) - \phi_0 \cdot \sum_t t \cdot p(t|D=d_0) \\ &= \phi_i \mathrm{E}[T|D=d_i] - \phi_0 \mathrm{E}[T|D=d_0]. \end{aligned} \tag{A3}$$