Predictive *in silico* Modeling of Emetic Potency of Liquid Cleaning Products Using an Historical *in vivo* Database

S. Li[1], F. Meli[1], P. Vinson[1], H.W. Broening[1], P.L.R. Andrews[2] and J F. Nash[1]

[1]The Procter & Gamble Company, 8700 Mason-Montgomery Road, Mason, Ohio, 45040 USA.

[2]St George's University of London, Cranmer Terrace, London, SW17 0RE, UK.

Keywords: emesis, *in silico*, multivariate analysis, nausea, 3Rs, recursive partitioning analysis, vomiting

Corresponding author:

Email: nash.jf@pg.com

Abstract

The induction of vomiting by activation of mechanisms protecting the body against ingested toxins is not confined to natural products but can occur in response to manmade medicinal and non-medicinal products such as liquid cleaning products where it is a commonly reported adverse effect of accidental ingestion. The present study examined the utility of an historic database (>30 years old) reporting emetic effects of 98 orally administered liquid cleaning formulations studied *in vivo* (canine model) to objectively identify the main pro-emetic constituents and to derive a predictive model. Data were analysed by categorizing the formulation constituents into 10 main groups followed by using multivariate correlation, partial least squares and recursive partitioning analysis. Using the ED_{50} we objectively identified high ionic strength, non-ionic surfactants (alcohol ethoxylate) and alkaline pH as the main pro-emetic factors. Additionally, a mathematical model was developed which allows prediction of the ED_{50} based on formulation. The limitations of the use of historic data and the model are discussed. The results have practical applications in new product formulation and safety but additionally the principles underpinning this *in silico* study have wider applicability in demonstrating the potential utility of such archival data in current research contributing to animal replacement.

Highlights

- Constituents causing the emetic effect of ingested liquid cleaning formulations are unknown.
- Recursive partitioning was used to model historic *in vivo* data on 98 liquid cleaning formulations.
- Emesis was positively associated with ionic strength, non-ionic surfactant, and high alkaline pH.
- The mathematical model predicted the ED_{50} from formulation composition.
- Application to product development, safety and wider assessment of emetic liability is discussed.

1. Introduction

Emesis or vomiting, i.e., the forceful oral expulsion of gastric contents, is one of the body's initial responses to toxins following ingestion as a food constituent or a contaminant (Davis et al., 1986). A wide range of substances of plant and animal origin with diverse structures can evoke nausea and vomiting in humans, e.g. muscarine, (Diaz, 2015); vomitoxin, (Wu et al., 2014); domoic acid, (Sobel and Painter, 2005); tetrodotoxin (Hayama and Ogura, 1963). Additionally, bacteria and their toxins, e.g. *Staphylococcus aureus* enterotoxins, (Angeles et al., 2010) and viruses, e.g. norovirus, (Baker et al., 2011); rotavirus, (Crawford et al., 2017) and COVID-19, (Andrews et al., 2020a) have nausea and vomiting as symptoms. The mechanisms and pathways by which these naturally occurring emetics induce nausea and vomiting can also be triggered by synthetic therapeutic drugs where nausea and vomiting then become side-effects, e.g. cancer chemotherapeutic agents such as cyclophosphamide, (Andrews and Rudd, 2015).

The activation of emetic pathways which evolved to protect the body against ingested natural toxins is not only confined to synthetic therapeutic agents but can also occur in response to non-medicinal synthetic products as exemplified by liquid cleaning products. Vomiting is the most common effect reported in cases of accidental ingestion of cleaning products in both children and adults (Day et al., 2019a, Day et al., 2019b, Smith et al., 2014). While there are reports decades earlier showing that cleaning products produced emesis in a canine model (Snyder et al., 1964, Weaver and Griffith, 1969), such accounts are sporadic and limited.

Considering, the large number of concentrated cleaning products currently marketed, e.g., laundry packets and tablets, further investigation into the components and physiochemical properties responsible for such events is warranted.

Previously we systematically reviewed and critiqued an historical database comprised of original study reports of liquid cleaning products tested in a canine model of emesis (Andrews et al., 2020b). The purpose was to determine if historical studies might inform and be reapplied to current formulations without additional animal testing thus contributing to the Replacement element of the "3Rs" (Replacement, Reduction, Refinement; (Russell and Burch, 1959)). The initial study determined that historical data could be used, with some limitations, to characterize the latency and magnitude of the emetic response and to demonstrate dose-response relationships for the incidence of emesis. Furthermore, detailed analysis of a sub-group of 15 formulations for which a complete data set (latency, intensity and ED_{100}) was available enabled calculation of a "vomiting index" (VI) showing an association between a high VI, a high percentage of non-ionic surfactants, high ionic strength, and a pH of ~10 which was proposed to be causally linked with the possible mechanism(s) discussed. Additionally, we found that the ED_{50} (the calculated dose evoking emesis in 50% of the group tested) provided a metric derived in a relatively consistent manner in all such studies, which serves as a dependent variable when assessing emetic potency for this group of cleaning products.

The present study extends the findings from Andrews et al. (2020b) by developing an *in silico* model to predict emetic potency of liquid cleaning products. The constituents of liquid cleaning

products, e.g., surfactants, polymers, hydrotropes, solvents, and physiochemical measures, e.g., pH, ionic strength, were placed into categories and, using multivariate and recursive partitioning statistical models, used to predict their contribution to emetic potency based on ED_{50}. It was hypothesised that surfactants, i.e., anionic, cationic and nonionic, would emerge as the categories responsible for emetic potency. However, we found that emetic potency of complex liquid cleaning product mixtures were most influenced by the ionic strength and concentration of non-ionic surfactant, i.e., alcohol ethoxylate. This study illustrates the value of using historical data to model, in this case emesis, without conducting additional animal testing. Moreover, this predictive modeling helps understand emetic potency of liquid cleaning products in cases of accidental ingestion.

2. Materials and methods

2.1. Data set

The studies used in this analysis were performed between 1973 and 1987 as part of toxicological testing commonly performed during this time period. None of the authors participated in the conduct of the studies which were performed in accordance with the ethical and regulatory requirements in place at the time. For an overview of the methodology and experimental details the reader is referred to Andrews et al. 2020b. This study of 74 liquid cleaning products (Andrews et al. 2020b) focused on the overall emetic characteristics of the formulations and explored the relationship between ingredient composition and the vomiting index (VI). Calculation of the VI required detailed reporting of emetic data (latency and magnitude in 4 animals at the ED_{100}) and so was only possible for a relatively small (20%) subset

of formulations. Whilst the VI approach provided insights into the relationship between emesis and formulation characteristics, the data requirements limited its wider use. However, as experimental records usually reported the ED_{50}, an additional 24 studies for a total of 98 data records were identified with both well-defined liquid cleaning product formula details and ED_{50} values (see below) defining the emetic potency. Statistical relationships were examined between formulae variations and this measure of emetic potency, i.e., ED_{50}.

2.2. Measure of formulation emetic potency

The ED_{50}, i.e., the calculated dose of the undiluted liquid formulation in mL/Kg which induced vomiting in 50% of the treatment group within 120 min, was determined using the "up and down" procedure according to Brownlee et al. (Brownlee et al., 1953) but over the 14 years of the study period other recognized comparable methods for determining the ED_{50} such as Dixon (Dixon and Mood, 1948), Weil (Weil, 1952) and Probit (Finney, 1947) were also used. The smaller the ED_{50} value, the more potent the emetic effect of the formulation. It should be noted that based on a limited data set of 20 formulations where both an ED_{100} value was achieved and an ED_{50} value was calculated for the same formulation that the two were significantly linearly correlated (Andrews et al., 2020b)

Although the historical reports reliably reported the ED_{50} as noted in Andrews et al. (2020b) information about vomiting onset time, repeated episodes and duration of effect were not reported consistently so are of limited utility in developing a predictive model requiring a large

data set and therefore these parameters are not in the scope of this manuscript (see Discussion).

2.3. Grouping of formulation Ingredients - dimension reduction

As part of data curation, different formulation ingredients and key physicochemical properties were grouped into the following 10 categories: (1) alkyl sulfonate – anionic surfactant, (2) alkyl sulfate – anionic surfactant, (3) ethoxylated alkyl sulfate – anionic surfactant (AES), (4) alcohol ethoxylate – nonionic surfactant (NI), (5) amine oxide/amine/amide/ - cationic surfactant (Zwitterionic/Cationic), (6) fatty acid, (7) solvent, (8) hydrotrope, (9) pH, and (10) ionic strength (*IS*). The *IS* is a function of the concentration of all ions present in each formulated product (IUPAC, 1997) according to equation 1:

$$IS = \frac{1}{2} \sum_{i=1}^{n} C_i Z_i^2 \qquad \text{(eq. 1)}$$

where c_i is the molar concentration of ion *i* (mol/L), z_i is the charge number of that ion, and the sum is taken over all ions in the solution.

2.4. Statistical methods

JMP software (version 12.2, SAS Institute Inc., Cary, NC) was employed as the statistical evaluation tool. Considering that the dataset did not originate from a statistical design of experiments, exploratory analysis including possibility distribution of variables and multivariate correlations were initially performed on the data, followed by three different regression

163 methods to investigate the formulation drivers for emesis. These different models were
164 developed to check for concordance.
165
166 *2.4.1 Probability Distribution of the Variables*
167 A probability distribution is a mapping of all the possible values of a random variable to their
168 corresponding probabilities. A histogram graph and outlier box plot were reported for each
169 variable. The key aspects of the outlier box plot [SAS JMP Software (version 12.2) user manual,
170 SAS Institute Inc., Cary, NC] include: (i) The ends of the box represent the first and third
171 quartiles; (ii) The horizontal line within the box represents the median sample value; (iii) The
172 confidence diamond contains the mean (the middle of the diamond) and the upper and lower
173 95% of the mean (top and bottom points); (iv) The red bracket outside of the box identifies the
174 shortest half, which is the most dense 50% of the observations (Rousseeuw and Leroy, 1987);
175 and, (v) The whiskers extend from the ends of the box to the outermost data point that falls
176 within the distances computed as follows:
177 *first quartile - 1.5*(difference between the first and third quartiles)*
178 *third quartile + 1.5*(difference between the first and third quartiles)*
179 If the data points do not reach the computed ranges, then the whiskers are determined by the
180 upper and lower data point values (not including outliers).
181
182 *2.4.2 Multivariate correlation*
183 The multivariate correlation analysis calculates the pairwise correlation between multiple
184 variables. For variables *x* and *y*, the Pearson correlation coefficient, *r*, is computed as follows:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \quad \text{(eq. 2)}$$

where the value *r* = 1 indicates an exact positive linear correlation, the value *r* = -1 indicates an exact negative linear correlation, and the value *r* = 0 means there is no linear relationship between *x* and *y*. Besides the correlation matrix, a scatterplot matrix is also reported to demonstrate how the variables relate to each other.

2.4.3 Recursive Partitioning Analysis (RPA)

Recursive partition (Kass and Hawkins, 1982, Kass, 1980) creates a decision tree that strives to correctly classify the response variable *Y* by splitting it into subsets where the distribution of *Y* is successively more homogeneous, based on a vector of independent variables *X*. The process is termed recursive because each subset may in turn be split an indefinite number of times until a particular stopping criterion is reached.

2.4.4 Multivariable Linear Regression (MLR)

Multiple linear regression is the most common form of linear regression analysis. It models the relationship between one continuous dependent variable y and two or more independent variables *X*. Multiple linear regression makes three key assumptions that need to be checked along the model development process (Kutner et al., 2005): (i) Multivariate normality — residuals of the regression are normally distributed; (ii) No multicollinearity — the independent

variables *X* are not highly correlated with each other; (iii) Homoscedasticity — the variance of residuals need to be similar across the values of the independent variables *X*.

2.4.5 Partial Least Squares (PLS)

Partial least squares regression is an extension of the multiple linear regression model and bears some relation to principal components regression. It fits linear models based on factors, namely, linear combinations of the independent variables. These factors are obtained in a way that attempts to maximize the covariance between the independent variables and the responses. PLS exploits the correlations between the independent variables and the responses to reveal underlying latent structures. PLS regression is especially useful when the independent variables are highly collinear, or when there are more independent variables than observations (Cox and Gaudard, 2013).

3. Results

3.1 Exploratory Analysis

Probability distributions of both independent, i.e., formulation variables, and dependent, i.e., ED_{50}, variables are presented in Figure 1. The ED_{50} value ranges from 0.0125 mL/kg to 32 mL/kg. To better fit the "multivariate normality" assumption underlying regression, a natural Log transformation was applied to the ED_{50} values to achieve a near normal distribution. As shown in the plot labeled "Ln (ED_{50}), the resulting Ln (ED_{50}, mL/kg) ranges from -4.4 to 3.5 for the 98 samples, with the mean of -0.9 and median of -1.0. The first and third quartiles of Ln (ED_{50}, mL/kg) are at -2.3 and 0.7, respectively. All the points are within the whiskers.

The probability distribution of formulation ingredients and physicochemical properties are also plotted in Figure 1. The formulation ingredients, including alkyl sulfonate, alkyl sulfate, AES, NI, zwitterionic/cationic, fatty acid, solvent, and hydrotrope, carry units of weight percentage (%) in the finished products (y-axis). The physicochemical properties include pH and *IS* with units of mol/L (y-axis). Since this historical dataset was not originated from statistically designed experiments, the distribution of many independent variables is peaked near the lower boundary of 0% with the tail on the higher concentration side, except for solvent and pH. The concentration of alkyl sulfonate ranges from 0% to 40.8%, with the mean of 6.1%, first quartile at 0%, medium at 2.4%, and third quartile at 7.1%. The concentration of alkyl sulfate ranges from 0% to 32.8%, with the mean of 1.4% and all the points in the fourth quartile. The concentration of AES ranges from 0% to 36.3%, with the mean of 5.9%, medium of 0% and third quartile at 9.2%. NI ranges 0% to 64% with the mean of 6.0%, medium of 0% and third quartile at 4.0%. Zwitterionic/Cationic ranges from 0% to 40%, with the mean of 1.4%, medium of 0%, and third quartile at 2.7%. FA ranges from 0% to 28%, with the mean of 1.9%, medium of 0% and third quartile at 0.3%. The solvent concentration ranges from 0% to 18.4%, with the mean of 4.9%, medium of 5.0%, first quartile at 0% and third quartile at 7.0%. The hydrotrope concentration ranges from 0% to 9%, with the mean of 2.0%, medium and third quartile at 0% and 3%, respectively. pH ranges from 5.1 to 11.9 with the mean of 8.9, first quartile at 7.1, median at 8.5, and third quartile at 10.5. The *IS* ranges from 0 to 28.4 mol/L, with the mean of 5.4 mol/L, first quartile, medium, and third quartile at 1.2 mol/L, 2.3 mol/L, and 4.5 mol/L, respectively.

Multivariate correlation of Ln ED_{50} and formulation variables are summarized in Table 1. A scatterplot matrix is also reported in the supplementary file (Supplementary Figure S1). Six samples were excluded from the multivariate analysis due to missing values of pH. As shown in Table 1, there is relatively strong correlation between Ln ED_{50} and AES (0.45), NI (-0.32), solvent (0.35), pH (-0.36) and *IS* (-0.55). Among the independent variables, pH has negative correlation with all the surfactants (*r* ranges from -0.46 to -0.20). IS also shows negative correlation with solvent (-0.53) and positive correlation with pH (0.39). Overall, the formulation variables are not highly correlated with each other, which satisfies the "no multicollinearity" assumption of the MLR analysis in 3.3.

3.2 Recursive Partitioning Analysis

The partitioning tree of RPA are illustrated in Figure 2. The actual Ln ED_{50} values against predicted Ln ED_{50} values by this RPA model are included in the Supplementary file (Supplementary Figure S2), with a fitting R^2 of 0.76. As shown by Figure 2, there are 5 partitions that split the 98 samples: (i) The first split is based on the criteria if *IS* ≥ 5.09 mol/L. The samples meeting the criteria are put on the left side predicted with lower Ln ED_{50} values and the samples that do not meet the criteria are placed on the right predicted with higher Ln ED_{50} values. (ii) The second split is based on all the samples with *IS* < 5.09 mol/L, among which the samples with NI ≥ 2.75% are predicted to have lower Ln ED_{50} values and the samples with less than 2.75% NI have higher Ln ED_{50} values. (iii) The third split is based on samples with *IS* < 5.09 mol/L and NI ≥ 2.75%. The samples with NI ≥ 18.00% are predicted to have lower Ln ED_{50} values

than the samples with NI < 18.00%. (iv) The fourth split is based on samples with *IS* < 5.09 mol/L and NI < 2.75%. The samples with *IS* ≥ 2.80 mol/L is predicted to have lower Ln ED_{50} values than the samples with *IS* < 2.80 mol/L. (v) The fifth split is based on the samples with NI < 2.75% and *IS* < 2.80 mol/L. Samples with pH ≥ 10.3 is predicted to have smaller Ln ED_{50} values than the samples with pH < 10.3.

A leaf report[1] that displays the mean Ln ED_{50}, mL/kg, and count of each leaf node is presented in Table 2. The 23 samples with *IS* $\underline{\geq}$ 5.09 mol/L have the smallest mean Ln (ED_{50}, mL/kg) of -2.75. The 9 samples with *IS* < 5.09 mol/L and NI $\underline{\geq}$ 18% have the second smallest mean Ln (ED_{50}, mL/kg) at -2.67. The Ln ED_{50} value increases with the reduction of *IS* and NI in the formulation. For the samples with *IS* < 2.80 mol/L and NI < 2.75%, the 11 samples with pH >= 10.3 has mean of Ln (ED_{50}, mL/kg) at -1.56 and the other 11 samples with pH < 10.3 has the highest mean of Ln (ED_{50}, mL/kg) at 1.19 mL/kg among all the 98 samples. Overall, results of RPA suggest the lower the *IS*, and concentration of NI, and the lower the pH (< 10.3), the higher the ED_{50} value, which is a reduction in the emetic potency of these liquid cleaning products.

3.3 Multivariable Linear Regression

The two graphs in Figure 3 illustrate the fitting quality of the experimentally determined, i.e., "measured", Ln ED_{50} vs. model predicted value (Left panel) and the residue of model prediction (right panel), with R^2 of 0.6. As shown, the MLR model satisfies the "Homoscedasticity"

[1] Each "leaf node" corresponds to one of the rectangles in Figure 2, which depicts the RPA "tree".

assumption. Due to the existence of missing values, pH was excluded from the formulation variable selection to enable all 98 formulations to be used to train the model. Response surface effects (including linear terms, quadratic terms, and binary interaction terms) of the formulation variables were taken into consideration during regression. Stepwise regression was used with minimum Bayesian Information Criterion as the stopping rule (Burnham and Anderson, 2004). The developed MLR model can be described as:

$$\mathrm{Ln}\ (\mathrm{ED50}) = 0.730 - 0.072 * \mathrm{NI} - 0.379 * IS + 0.014 * (IS - 5.368)^2 \quad \text{(eq. 3)}$$

where ED_{50} has unit of mL/kg, NI has unit of % concentration, and *IS* has unit of mol/L. The model indicates that increasing NI and *IS* leads to smaller Ln ED_{50}, and therefore greater emetic potency. The quadratic term describes the plateau effect of extremely high *IS* on Ln ED_{50} value.

3.4 Partial Least Squares Regression

Similar to the MLR, pH was also excluded from the formulation variable selection to enable all 98 formulations to be used to train the PLS model. As shown in Figure 4, two factors (X1, X2) cover a total of 62.8% of the variation in Ln ED_{50}, with factor X1 capturing 54.2% of Ln ED_{50} variance and factor X2 capturing 8.6% of Ln ED_{50} variance. Figure 5 summarizes the model coefficients based on centered and scaled data. As illustrated in Figure 5, *IS* and NI are identified as the two major negative drivers for Ln ED_{50} with model coefficients of -0.43 and -0.39, respectively. Alkyl sulfonate and AES are identified as major positive drivers for Ln ED_{50} with coefficients of 0.33 and 0.32, respectively, which is counter intuitive and probably due to

their intrinsic negative correlation with pH (as shown in Table 1). It can be the lower pH that actually drives higher Ln ED_{50}. The other formulation variables are less significant based on smaller absolute values of model coefficient.

4. Discussion

In a previous study of historical data (from >30 years ago) from canine studies of the emetic response to liquid cleaning formulations we established its utility to characterize the emetic response and to identify pro-emetic physicochemical factors (Andrews et al., 2020). The present statistical analysis of a larger data set reporting the ED_{50} of 98 liquid cleaning formulations extends the previous study by developing a mathematical model for predicting the ED_{50}. In the sections below we discuss the factors influencing emetogenicity, the limitations of the model, and the specific and more general applicability of the findings from this study.

4.1. *Factors influencing emetogenicity identified by the model.*

The original experimental data used to develop the model included the ED_{50} and the formulation composition categorized by 10 key constituents. Statistical analysis using three regression models revealed that the formulae components driving emetic potency were ionic strength, non-ionic surfactant (alcohol ethoxylate) and, to a lesser extent highly alkaline pH. Thus, a lower ED_{50} indicative of a higher emetic liability at a lower dose is associated with higher ionic strength and concentration of alcohol ethoxylate, i.e., non-ionic surfactant. This conclusion is consistent with the preliminary analysis in Andrews et al. (2020) using the

336 "vomiting index" as the outcome measure but which was limited to 11% of formulations
337 because calculation of the "vomiting index" requires a more detailed data-set which was not
338 available for the large number of formulations studied here. A discussion of the mechanisms by
339 which the above key constituents activate the pathways inducing the nausea and vomiting is
340 outside the scope of the present paper. Potential emetic mechanisms were discussed in
341 Andrews et al. (2020) with a focus on the effect of the key constituents with mucosally located
342 enteroendocrine cells in the stomach and small intestine releasing mediators locally to act on
343 terminals of the abdominal vagal afferents. Additionally, Andrews et al., (2020) compared the
344 profile of the emetic response to liquid cleaning formulations with that reported in the
345 literature for a wide range of emetics also given by gavage in the canine model.
346
347 4.2. *Limitations of the model*
348 The model developed and the resulting predictions depend upon the quality of the data derived
349 from the original historical studies and the mathematical/statistical methodology used in its
350 genesis. The challenges and limitations in using historic data including variability of data
351 collection, protocol variations with time, nature of the data collected, and controls were
352 discussed in Andrews et al. (2020b) and will not be reiterated here. However, here we focus on
353 the data used in the current model and its limitations.
354
355 The ED_{50} is a well-established metric of the potency of a biologically active substance and in the
356 historic studies although differing methods (see section 2.2) were used for its derivation, each is
357 valid, likely yielding comparable values. Our previous study included twenty of the same

formulations analyzed here where the dose (ED_{100}) producing an emetic response in all animals in a group was established allowing us to show a significant linear correlation (R^2= 0.84) between the ED_{50} and the ED_{100} (Andrews et al., 2020b, Supplementary material). In view of the latter finding, we are confident that the use of the ED_{50} is a valid metric reflecting the emetic potency of a given formulation. However, it must be noted that the ED_{50} is a reflection of the incidence of emesis (i.e. the probability of occurrence) which although directly relevant to safety assessment of a consumer product, does not reflect either the magnitude (number of vomits) or latency (time for onset of vomiting) of the response. Future development of the mathematical model should ideally incorporate a measure of the magnitude of the emetic response although it should be noted that the previous study of a subset of formulations showed no obvious relationship between the ED_{100} and the magnitude or latency of the response for 18 formulations for which a full data-set was available (Andrews et al., 2020b). Nevertheless, the analysis based on ED_{50} values has identified constituents increasing the probability of emesis and enabled derivation of a mathematical predictive model.

The three regression models used in the present analysis are straightforward with respect to their application and utility. The selection of multiple methods to analyse the current dataset was an attempt to determine concordance amongst the key findings using more than one statistical model. However, the most significant limitations were the limited number of studies available and the diversity of formulae used in the analysis. Ideally, for such an exercise there would be hundreds to thousands of studies which could be used in such a retrospective regression evaluation. Realistically, however, historical *in vivo* databases may be limited in

380 useable data and/or number of completed studies (see Andrews et al., 2020b for further
381 discussion). Even so, we have shown there is some value even with a limited number of studies.
382
383 The liquid cleaning formulae were not created to test a specific hypothesis related to emesis.
384 These were products made for marketing and consumer use and, as such, lack broad formula
385 diversity in the concentration and selection of ingredients. This is both an advantage and
386 disadvantage. The advantage is in grouping of formula components; it is achievable in that the
387 assortment of ingredients in these formulations is limited. Of course, this is also a disadvantage
388 to the extent that liquid cleaning products are not monolithic in design or make up requiring
389 the qualification of current findings. Moreover, grouping of formula components is relatively
390 broad. For example, amongst non-ionic surfactants of the alcohol ethoxylates, there are many
391 different chemistries based on alkyl chain length and number of ethoxylates which may
392 influence the biological potency (Broening et al., 2019). Even with such limitations, there was
393 agreement in the model findings with the systematic evaluation of individual studies as
394 reported in Andrews et al. 2020b.
395
396 4.3 *Practical applications.*
397 The findings presented here have two main practical implications. Firstly, the immediate impact
398 is to inform the development of new liquid cleaning product formulation and hence further
399 improve product safety. Additionally, this modeling of liquid cleaning products identified non-
400 ionic surfactants (alcohol ethoxylates) and ionic strength as key factors predicting emetic
401 potency of such products, which significantly improves our ability to anticipate the

consequences of accidental ingestion. Finally, the development of this predictive model further contributes to the commitment of Procter& Gamble to eliminating the use of animals in product testing (https://us.pg.com/policies-and-practices/animal-welfare-policy/) and the creation of new state-of-the-art approaches to evaluate this important endpoint. Secondly, this work is, to some extent, a proof of concept with applications beyond liquid cleaning products. Institutions, i.e., commercial, government, often have vast caches of experimental data that were never published. Such data are often based on animal models that have been abandoned for one reason or another yet have some value with respect to endpoint evaluation. The canine emesis model is one such example. During the 1970-1980s, such studies were performed routinely in the commercial sector usually as part of product safety assessment requirements. Such data often remains stored or even "lost" in company archives because of a short corporate memory and hence is unused, with little effort to examine its applicability to current product development or research questions. Such *In vivo* studies are no longer performed routinely, so the historic data represents a potentially valuable resource which with the techniques now available to analyse large data sets, together with greater mechanistic understanding, can contribute to answering current practical and research problems. Although this *in silico* study has focused on liquid cleaning products the methodology here has wider applicability in assessing emetic liability in toxicology. For example, assessing the emetic potential of novel drugs intended for therapeutic use where nausea and vomiting as side-effects can both curtail drug development and reduce patient compliance; the potential use of historical data in this area has been discussed previously (Holmes et al., 2009, Percie du Sert et al., 2012).

5. Conclusion

This study, together with a previous related one (Andrews et al., 2020b) has demonstrated the utility of data from historic *in vivo* animal studies to identify pro-emetic constituents of liquid cleaning formulations and perhaps of greater significance to develop a predictive model. Despite the acknowledged limitations, the results have practical applications in new product formulation and safety but additionally the principles underpinning this *in silico* study have wider applicability in demonstrating the potential utility of such archival data in current research contributing to animal replacement. The approach taken here has wide applicability as similar unique data sets from animal studies are likely to be in the archives of many organizations and could contribute to replacement of the use of animals. The approach taken here to develop a predictive model based on analysis of historic data exemplifies the potential of predictive toxicology for “next generation” risk assessment which could inform regulatory decisions, e.g., (Fitzpatrick et al., 2020).

Author's contribution

S. Li undertook the analysis generating the model and together with JFN and PLRA drafted the paper. PLRA and JFN revised the draft manuscript with concurrence from the other authors.

Declaration of competing interest

J F. Nash, S. Li, F. Meli, P. Vinson and H Broening are employees of P&G. PLRA is in receipt of an unrestricted P&G educational grant and has acted as a consultant to P&G.

Acknowledgements.

We wish to thank the researchers who collected the original data >30 years ago and made the current analysis possible, as well as Drs. Sabaliunas, Roggeband, Takechui, Ms. Johnson and Mr. White for their review and thoughtful comments.

Supplementary data.

Supplementary data to this article can be found online.

Table 1. Pairwise multivariate correlation of Ln ED_{50} and formulation variables

	Ln (ED_{50})	Alkyl Sulfonate	Alkyl Sulfate	AES	NI	Zwitterioinc /Cationic	Fatty Acid	Solvent	Hydrotrope	pH	IS
Ln (ED_{50})	1.00										
Alkyl Sulfonate	0.25	1.00									
Alkyl Sulfate	0.04	0.22	1.00								
AES	0.45	-0.18	0.09	1.00							
NI	-0.32	-0.13	-0.08	-0.21	1.00						
Zwitterioinc/Cationic	0.07	-0.13	0.04	0.23	0.23	1.00					
Fatty Acid	-0.04	0.05	0.03	-0.08	-0.02	-0.06	1.00				
Solvent	0.35	-0.01	-0.01	0.17	0.12	-0.07	0.14	1.00			
Hydrotrope	0.08	-0.09	-0.11	0.14	-0.27	-0.04	-0.20	0.10	1.00		
pH	-0.36	-0.27	-0.20	-0.46	-0.33	-0.34	-0.12	-0.18	0.13	1.00	
IS	-0.55	-0.02	0.01	-0.25	-0.21	-0.14	-0.04	-0.53	-0.23	0.39	1.00

** 6 samples were excluded from the multivariate analysis due to missing values of pH*

456 **Table 2**. The leaf report of the decision tree for Recursive Partitioning Analysis

Leaf of the decision tree	**Mean of Ln (ED_{50}, mL/kg)**	**Sample count**
IS > 5.09	-2.75	23
IS < 5.09 & NI > 18.00	-2.67	9
IS < 5.09 & 2.75 ≤ NI < 18.00	-0.49	18
2.80 ≤ *IS* < 5.09 & NI < 2.75	-1.56	11
IS < 2.80 & NI < 2.75 & pH 10.3	-0.16	11
IS < 2.80 & NI < 2.75 & pH < 10.3	1.19	26

457

458 **Figure legends**

459 **Figure 1**. Probability Distribution of independent variables (formulation variables) and
460 dependent variable (ED_{50}). Ten (10) independent variables Alkyl sulfonate – anionic surfactant,
461 Alkyl sulfate – anionic surfactant, Alkyl ethoxylate sulfate (AES) – anionic surfactant, Nonionic
462 surfactant (NI) – alcohol ethoxylate, Zwitterionic/Cationic surfactant, Fatty acid, Solvent, and
463 Hydrotrope are presented as percent concentration (y-axis) and probability distribution (x-axis).
464 For pH and ionic strength (IS) – mol/L, the y-axis are these measures versus probability
465 distribution (x-axis). The Ln (ED%) for all 98 formulations analysed are presented. The right-
466 hand panel of each constituent independent variable shows a box and whisker plot indicating
467 the median, confidence diamond and first and third quartile for each distribution and these are
468 also labelled in the Ln (ED_{50}) panel at the extreme lower right of the figure.

469

470 **Figure 2**. Partitioning tree of the Recursive Partitioning Analysis. IS=Ionic strength; NI=non-ionic
471 surfactant; pH= $-\log_{10}[H^+]$

472

473 **Figure 3**. Model quality of the Multivariable Linear Regression. The left panel plots the
474 relationship between the measured Ln ED_{50} vs. model predicted value. The diagonal red line
475 shows the linear regression (R^2=0.6). The right panel plots the residue of model prediction vs.
476 model predicted Ln ED_{50} and shows that they are not correlated.

477

478 **Figure 4**. Scatterplots of the X and Y scores for each extracted factor in the Partial Least Squares
479 Regression together with the linear correlation (diagonal red line). From the X-Y scores plots,

480 the two extracted factors (X1, X2) cover a total of 62.8% of the variation in Ln ED_{50}, with factor
481 X1 capturing 54.2% of Ln ED_{50} variance and factor X2 capturing 8.6% of Ln ED_{50} variance.
482
483 **Figure 5.** Key formulation drivers for Ln ED_{50} from Partial Least Squares Regression; negative
484 drivers are plotted to the left and positive drivers to the right. See text for details but note that
485 AES and alkyl sulfonate have an intrinsic negative correlation with pH (Table 1).

References

ANDREWS, P. L. R., CAI, W., RUDD, J. A. & SANGER, G. J. 2020a. Nausea, vomiting and COVID-19. *Gastroenterol Hepatol,* in press.

ANDREWS, P. L. R., LI, S., MELI, F., VINSON, P., BROENING, H. W. & NASH, J. F. 2020b. Evaluation of historic in vivo data to characterise the emetic properties of liquid cleaning products and provide a framework for the development of an in silico predictive algorithm. *Food Chem Toxicol,* 143**,** 111553.

ANDREWS, P. L. R. & RUDD, J. A. 2015. The physiology and pharmacology of nausea and vomiting induced by anti-cancer chemotherapy in humans. *In:* NAVARI, R. M. (ed.) *Management of Chemotherapy-induced Nausea and Vomiting: New Agents and New Uses of Current Agents.*

ANGELES, M. A., MENDOZA, M. C. & RODICIO, M. R. 2010. Food poisoning and *Staphylococcus aureus* enterotoxins. *Toxins,* 2**,** 1751-1773.

BAKER, K., MORRIS, J., MCCARTHY, N., SALDANA, L., LOWTHER, J., COLLINSON, A. & YOUNG, M. 2011. An outbreak of norovirus infection linked to oyster consumption at a UK restaurant, February 2010. *J Public Health (Oxf),* 33**,** 205-11.

BROENING, H. W., LA DU, J., CARR, G. J., NASH, J. F., TRUONG, L. & TANGUAY, R. L. 2019. Determination of narcotic potency using a neurobehavioral assay with larval zebrafish. *Neurotoxicology,* 74**,** 67-73.

BROWNLEE, K. A., HODGES, J. L., JR. & ROSENBLATT, M. 1953. The up-and-down method with small samples. *Journal of the American Statistical Association,* 48**,** 262-277.

BURNHAM, K. P. & ANDERSON, D. R. 2004. Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods Research,* 33**,** 261-304.

COX, I. & GAUDARD, M. 2013. *Discovering partial least squares with JMP,* Cary, NC, SAS Institute Inc.

509 CRAWFORD, S. E., RAMANI, S., TATE, J. E., PARASHAR, U. D., SVENSSON, L., HAGBOM, M., FRANCO, M.
510 A., GREENBERG, H. B., O'RYAN, M., KANG, G., DESSELBERGER, U. & ESTES, M. K. 2017. Rotavirus
511 infection. *Nat Rev Dis Primers,* 3**,** 17083.
512 DAVIS, C. J., HARDING, R. K., LESLIE, R. A. & ANDREWS, P. L. R. 1986. The organisation of vomiting as a
513 protective reflex: a commentary on the first day's discussions. *In:* DAVIS, C. J., LAKE-BAKAAR, G.
514 V. & GRAHAME-SMITH, D. G. (eds.) *Nausea and Vomiting: Mechanisms and Treatment.* Berlin:
515 Springer-Verlag.
516 DAY, R., BRADBERRY, S. M., JACKSON, G., LUPTON, D. J., SANDILANDS, E. A., S, H. L. T., THOMPSON, J. P.
517 & VALE, J. A. 2019a. A review of 4652 exposures to liquid laundry detergent capsules reported to
518 the United Kingdom National Poisons Information Service 2008-2018. *Clin Toxicol (Phila)***,** 1-8.
519 DAY, R., BRADBERRY, S. M., THOMAS, S. H. L. & VALE, J. A. 2019b. Liquid laundry detergent capsules
520 (PODS): a review of their composition and mechanisms of toxicity, and of the circumstances,
521 routes, features, and management of exposure. *Clin Toxicol (Phila)***,** 1-11.
522 DIAZ, J. H. 2015. *Atlas of Human Poisoning and Envenoming,* Boca Raton, USA, CRC Press.
523 DIXON, W. J. & MOOD, A. M. 1948. A Method for Obtaining and Analyzing Sensitivity Data. *Journal of the*
524 *American Statistical Association,* 43**,** 109-126.
525 FINNEY, D. J. 1947. *Probit analysis; a statistical treatment of the sigmoid response curve,* Oxford,
526 England, Macmillan.
527 FITZPATRICK, S. C., DABT, E. R. T. U. S. F. & DRUG, A. 2020. Predictive toxicology for regulatory decisions:
528 Implementing new approaches at US Food and Drug Administration. *Toxicol In Vitro,* 63**,** 104659.
529 HAYAMA, T. & OGURA, Y. 1963. Site of emetic action of tetrodotoxin in dog. *J Pharmacol Exp Ther,* 139**,**
530 94-6.
531 HOLMES, A. M., RUDD, J. A., TATTERSALL, F. D., AZIZ, Q. & ANDREWS, P. L. 2009. Opportunities for the
532 replacement of animals in the study of nausea and vomiting. *Br J Pharmacol,* 157**,** 865-80.

IUPAC 1997. *Compendium of Chemical Terminology (the "Gold Book"),* Oxford, Blackwell Scientific Publications.

KASS, G. V. 1980. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics,* 29**,** 119-127.

KASS, G. V. & HAWKINS, D. M. 1982. Automatic interaction detection. *In:* HAWKINS, D. M. (ed.) *Topics in applied multivariate analysis.* Cambridge, UK: Cambridge University Press.

KUTNER, M. H., NACHTSHEIM, C. J., NETER, J. & LI, W. 2005. *Applied Linear Statistical Models,* New York, McGraw-Hill Irwin.

PERCIE DU SERT, N., HOLMES, A. M., WALLIS, R. & ANDREWS, P. L. 2012. Predicting the emetic liability of novel chemical entities: a comparative study. *Br J Pharmacol,* 165**,** 1848-1867.

ROUSSEEUW, P. J. & LEROY, A. M. 1987. *Robust regression and outlier detection,* New York, John Wiley & Sons.

RUSSELL, W. M. S. & BURCH, R. L. 1959. *The Principles of Humane Experimental Technique,* London, Methuen & Co LTD.

SMITH, E., LIEBELT, E. & NOGUEIRA, J. 2014. Laundry detergent pod ingestions: is there a need for endoscopy? *J Med Toxicol,* 10**,** 286-91.

SNYDER, F. H., OPDYKE, D. L., GRIFFITH, J. F., RUBENKOENIG, H. L., TUSING, T. W. & PAYNTER, O. E. 1964. Toxicologic Studies on Household Synthetic Detergents. I. Systemic Effects. *Ther Ggw,* 103**,** 133-40.

SOBEL, J. & PAINTER, J. 2005. Illnesses caused by marine toxins. *Clin Infect Dis,* 41**,** 1290-6.

WEAVER, J. E. & GRIFFITH, J. F. 1969. Induction of emesis by detergent ingredients and formulations. *Toxicol Appl Pharmacol,* 14**,** 214-20.

WEIL, C. S. 1952. Tables for convenient calculation of median-effective dose (LD_{50} or ED_{50}) and instructions in their use. *Biometrics,* 8**,** 51-54.

557 WU, W., ZHOU, H. R., BURSIAN, S. J., PAN, X., LINK, J. E., BERTHILLER, F., ADAM, G., KRANTIS, A., DURST,
558 T. & PESTKA, J. J. 2014. Comparison of anorectic and emetic potencies of deoxynivalenol
559 (vomitoxin) to the plant metabolite deoxynivalenol-3-glucoside and synthetic deoxynivalenol
560 derivatives EN139528 and EN139544. *Toxicol Sci,* 142**,** 167-81.

561