Journal of Neuroscience Methods 397 (2023) 109931

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

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

Methodical advances in reproducibility research: A proof of concept qualitative comparative analysis of reproducing animal data in humans

Cathalijn H.C. Leenaars[a,b,*], Steven Teerenstra[c], Franck L.B. Meijboom[b], André Bleich[a]

[a] Institute for Laboratory Animal Science, Hannover Medical School, Hannover, Germany
[b] Department of Animals in Science and Society - Faculty of Veterinary Sciences, Utrecht University, Utrecht, the Netherlands
[c] Department for Health Evidence (section biostatistics), Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, the Netherlands

ARTICLE INFO

Keywords:
Reproducibility
Prediction
Extrapolation
Validity
Modelling
Scoping review
Qualitative comparative analysis
Regression analysis

ABSTRACT

Background: While the term reproducibility crisis mainly reflects reproducibility of experiments between laboratories, reproducibility between species also remains problematic. We previously summarised the published reproducibility between animal and human studies; i.e. the translational success rates, which varied from 0% to 100%. Based on analyses of individual factors, we could not predict reproducibility.

Several potential analyses can assess effect of combinations of predictors on an outcome. Regression analysis (RGA) is common, but not ideal to analyse multiple interactions and specific configurations (≈ combinations) of variables, which could be highly relevant to reproducibility.

Qualitative comparative analysis (QCA) is based on set theory and Boolean algebra, and was successfully used in other fields. We reanalysed the data from our preceding review with QCA.

Results: This QCA resulted in the following preliminary formula for successful translation:

~Old*~Intervention*~Large*MultSpec*Quantitative

Which means that within the analysed dataset, the combination of relative recency (~ means not; >1999), analyses at event or study level (not at intervention level), $n < 75$, inclusion of more than one species and quantitative (instead of binary) analyses always resulted in successful translation (>85%). Other combinations of factors showed less consistent or negative results. An RGA on the same data did not identify any of the included variables as significant contributors.

Conclusions: While these data were not collected with the QCA in mind, they illustrate that the approach is viable and relevant for this research field. The QCA seems a highly promising approach to furthering our knowledge on between-species reproducibility.

1. Background

While the debate on the relevance and acceptability of animal experimentation remains polarized (Baker et al., 2019; Genzel et al., 2020; Herrmann and Kimberly, 2019; Hobson-West, 2010), animal experiments are still hard to avoid in the process of new drugs reaching the market. However, the predictive value of animal experiments has limits, and poor reproducibility of results from animal experiments in humans may contribute to the high attrition rates in drug development (Kola and Landis, 2004). Explaining attrition can contribute to more efficient drug development, which is one of the reasons why we analyse translational success. Another one is animal welfare; we cannot defend using animals for translational experiments that do not provide relevant information.

While the most common approach to evaluating translation is mechanistic and qualitative, we started focussing on quantitative studies in a scoping review of reviews (Leenaars et al., 2019). In that review, we observed translational success rates from 0% to 100% (median: 64%; interquartile range (IQR): 44–79%). To identify factors contributing to translational success, we visualised these data by several potentially predictive factors, which are explained further below. Relevant for this paper are: definition type (binary vs. continuous definitions of translation), unit of analysis (UoA; our UoA was the study included in our

Abbreviations: csQCA, crisp-set Qualitative Comparative Analysis; IQR, InterQuartile Range; mvQCA, multivalue Qualitative Comparative Analysis; QCA, Qualitative Comparative Analysis; RGA, ReGression Analysis; SE, Standard Error; UoA, Unit of Analysis.

* Corresponding author at: Institute for Laboratory Animal Science, Hannover Medical School, Hannover, Germany.
E-mail address: Leenaars.Cathalijn@mh-hannover.de (C.H.C. Leenaars).

https://doi.org/10.1016/j.jneumeth.2023.109931
Received 14 February 2023; Received in revised form 16 May 2023; Accepted 28 July 2023
Available online 29 July 2023

analyses, but within the included studies, we recognised three categories of analysis units: events, interventions or studies, all further explained below), species, the number of included observations (events, interventions or studies), and the year of publication. There was no apparent relationship between any of these individual factors and the percentage of translational success. However, the effects of combinations of these potential factors on translational success could still be relevant. We thus performed additional analyses on our previously-collected data, which are described in this paper.

The following five factors were further analysed based on their theoretical relevance: publication age, UoA, analysis size, inclusion of multiple species and type of definition for translation (binary vs. continuous). Publication age was included as the state of science and the quality of animal models are thought to improve over time, and because animal-to-human translation is getting more attention in the last decade, which may result in improvements.

The UoA was previously extracted as a categorical variable with 3 possible values: event, intervention or study. Events were mainly specific adverse events observed in animals, humans or both, where the observed translation (e.g. the percentage of adverse events observed in both animals and humans) depends on study size and chance; larger studies have a larger chance of picking up rare events (Jacobson et al., 2001). Interventions were mostly specific drugs, where certain groups of interventions may translate better than others (Mahmood, 2006), and the observed translational success rate depends on the sampling, e.g. which group of drugs was analysed. Studies were the UoA, in meta-analyses of similar outcome data from multiple animal- and human references (Yen et al., 2014; Faggion et al., 2010), or in analyses that followed up on specific sets of preclinical publications (Hackam and Redelmeier, 2006). Analyses at the level of individual studies heavily depend on multiple factors, comprising included data set and experimental design of the compared studies (Leenaars et al., 2020).

The number of included observations was counted at the level of the UoA, and could thus reflect a number of events, interventions or studies. It was included in the current analyses as a proxy for power, as underpowered studies can result in erroneous conclusions which may impact translation (Ioannidis, 2005). While some authors argue that species differences introduce uncertainties that seriously limit their validity (Pound and Ritskes-Hoitinga, 2018), investigating multiple species can at least theoretically improve translation, as successful transfer of a first species barrier may be predictive of crossing a second.

The definition type for animal to human translation could be binary, i.e., there was successful translation or there was not, or continuous, which could refer to a percentage success, a correlation coefficient between animal and human data, a percentage overlap in confidence intervals, etc. Binary definitions can of course be expressed as percentages success, but the type of definition may impact the observed translational success.

As multiple roads lead to Rome, multiple combinations of these factors may lead to translational success. In scientific terms, there possibly is causal complexity; comprising equifinality (i.e., there are multiple routes to success) (Kast and Rosenzweig, 1972) and conjunctural causation (i.e., combinations of factors may be involved instead of individual factors (Aus, 2009)). Besides, causation may be asymmetrical (Befani, 2013); while the presence of a factor may contribute to success, its absence does not necessarily result in failure (and vice versa). Thus, we have a configural research question: "Which factors, individually or in combination, are necessary or sufficient for successful animal-to-human translation?".

To analyse effects of multiple potential predictors on an outcome, regression analysis (RGA) is common. However, RGA is not specifically suitable for research questions comprising multiple interactions and specific configurations (≈ combinations) of predictors. Qualitative comparative analysis (QCA) is an approach developed for configural research questions (Ragin, 1999). It is based on set theory and Boolean algebra. QCA is increasingly used to identify specific configurations of factors predicting an outcome in other fields (Hanckel et al., 2021; Roig-Tierno et al., 2017). We reanalysed the data from our preceding review with a crisp-set QCA (csQCA) (Vink and Van Vliet, 2009). To test the added value of this QCA-approach, we compared it with a classical regression analysis (RGA).

2. Methods

2.1. Data collection and selection

We reanalysed the data published in our systematic scoping review (Leenaars et al., 2019) for this study. This preceding scoping review was an umbrella review of reviews that addressed animal-to-human translation quantitatively, comparing the results of studies including at least 2 species with one being human. Data were extracted from the included publications to Microsoft Excel. When an included paper described multiple studies or analyses on different data, all those compliant with the inclusion criteria were included as a separate "case" into our analyses. When the original authors did not express translation as a percentage, but provided the data needed to do so, we calculated the percentage and added it to the respective case.

From these already published data, we selected the following factors as theoretically relevant for further combined analyses (as explained in the introduction): definition type (binary vs. continuous), UoA (event, intervention or study), species, the number of included observations and the year of publication. Cases with missing data for any of the analysed factors were excluded from the analyses (numbers are mentioned in the results).

All analyses were performed in R (Anon, 2022), version 4.0.3 ("Bunny-Wunnies Freak Out"), via RStudio (Version 1.3.1093). Data were imported from excel using the Readxl package (readxl, 2023). Where needed, data were selected with functions from the Dplyr package (dplyr, 2023).

2.2. Qualitative Comparative Analysis

As described in the introduction, QCA is an approach developed to deal with causal complexity (Ragin, 1999). It is based on set theory (sets being collections of cases) and Boolean algebra of these sets (combined configurations with the operators "AND", "OR", and "NOT"). QCA comprises several steps; the main steps that can be distinguished are 1.) case selection and data collection, 2.) data calibration, 3.) creation of a truth table, 4.) logical minimisation and 5.) interpretation (Ragin, 1999; Hanckel et al., 2021; Roig-Tierno et al., 2017; Melendez-Torres et al., 2018; Naims and Eppinger, 2022; Skaaning, 2011). We selected csQCA, one of the main types of QCA (Roig-Tierno et al., 2017), as it allows for more straightforward definitions and a conservative approach to calibration (Vink and Van Vliet, 2009). (In the field of data analysis, "conservative" means cautious to prevent false positive conclusions.) In csQCA, both the outcome and the explanatory conditions need to be dichotomised (Skaaning, 2011). "Calibration" refers to selecting the threshold value for the dichotomy. The truth table is a reorganisation of the data, into lines with the same "configuration", the same combination of analysed set memberships. Logical minimisation is the process of summarising the truth table into a logical formula. Interpretation depends on the presence of configurations with inconsistent outcomes and logical remainders within the truth table (further explained below).

We performed two separate QCAs; the first for translational success, and the second for translational failure. All cases with calibrated translational success or failure (see below) were included in both QCAs.

2.3. Data Calibration & data matrix

We calibrated all data to create so-called crisp sets as described in Table 1. Data were calibrated on theoretical grounds and based on expert opinions from within our network. For example, cut-offs for old

Table 1
Data calibration for QCA. UoA: Unit of Analysis.

Set name	Definition IN set	Definition OUT set
Old	Publication date < 2000	Publication date ≥ 2000
Intervention	"Interventions" were the UoA	"events" or "studies" were the UoA
Large	k > 75	k ≤ 75
MultSpec	At least two animal species were analysed	At most one animal species was analysed
Quantitative	Translation was calculated in a continuous manner	Translation was defined in a binary manner
Translation (outcome)	Success: > 80%	Failure: < 45%

publication age at around the start of the century were based on the use of the internet becoming increasingly common in research. Also, > 75 observations is considered a large study in the qualitative field. Calibration was not performed in a blinded manner, data were known at the time of calibration. Knowledge of the data informed our set definitions, i.e., the thresholds for dichotomisation, to the extent that (near) empty sets were consciously prevented. During calibration, data were only assessed as distributions/ number of counts per category per variable, they were not analysed at the case- or configuration level.

UoAs in the included reviews could be interventions, publications & studies, and particular (e.g., adverse) events. Refer to the introduction above or to our original publication (Leenaars et al., 2019) for a further explanation of the UoAs. We distinguished observations at the intervention level from those at the event and study level, as the latter two are both chance processes, while at the intervention level we can imagine a clear distinction between compounds that translate well (i.e., have similar outcomes in animals and humans) and those that do not (i.e., have different outcomes in animals and humans, due to differences in pharmacokinetics, absence versus presence of specific receptors, etc.). Translational success is difficult to define quantitatively (Leenaars et al., 2019). For our QCAs, we selected the less disputable percentages only: success was defined as > 80% correspondence, failure as < 45%. We excluded the reviews with percentages from 45% to 80% (which were included in the RGA described below).

Set memberships scores were added to the data file in separate columns.

2.4. Truth table creation & logical minimisation

A truth table was created with the truthTable function from the QCA package (Dusa, 2019). We analysed the truth table for sufficiency and necessity of individual factors before addressing combinations of factors. Sufficiency is the presence of the outcome in all cases with the occurrence of a predicting factor, and the factor is never present without the outcome (F → O, If F then O); necessity is the presence of a factor in all cases with the occurrence of an outcome, but the factor can also be present in cases without the outcome (F ← O, If O then F). Next, logical minimisation of configurations was performed with the minimize function from the QCA package.

We anticipated both logical inconsistencies and logical remainders in the truth table. Logical inconsistencies are rows with inconsistent outcomes (i.e., configurations that had both translational success and translational failure). This reanalysis of available data does probably not include all predictive factors relevant to translational success, which would result in perfectly defined sets without logical inconsistencies. Logical remainders are theoretically possible configurations that are not present in the data. Configurations with inconsistent translation and logical remainders were accepted, but they were not used to inform logical minimisation (in the QCA package's truthTable function: incl.cut = 1, n.cut = 1, pri.cut = 0). Also, because of our awareness of missing relevant factors in this proof of concept study, we did not analyse coverage of the solutions; i.e., which part of the cases could be explained with the final formula.

2.5. Regression analysis

All cases with complete data were included for the RGA. The following variables were included in the RGA: definition type (binary vs. continuous), UoA (event, intervention or study), multiple species, the number of included observations and the year of publication. Compared to the QCA, we included more data into the RGA. Because we did not have to dichotomise data into crisp sets, we included all cases with full data, also those with translational success from 45% to 80%, with the original percentage as the outcome. The variables for the number of included observations and the year of publication were also included as numbers instead of dichotomising them. The variables definition type, UoA and multiple species were included as binary variables, exactly like in the QCA.

Regression analysis was performed with the lm function from R's basic stats package (Anon, 2022). We tested a single model, including all variables included in the QCA individually. To provide a comparison with the QCA outcome we further added an interaction term for the variables "MultSpec" and "Quantitative".

3. Results

3.1. QCA

In our original review, we included 232 cases from 121 references. From these original cases, the 104 without missing data but with clearly successful or clearly unsuccessful translation were included in the QCA. Of these, 50 showed successful translation and 54 did not. The number of cases included in each set as defined in Table 1 is shown in Table 2.

The different observed set configurations with the outcomes are summarised in a truth table (Table 3). The truth table shows that 9 configurations had inconsistent (both successful and unsuccessful translation) results. For 16 configurations, no cases were observed; these are the so-called logical remainders. None of the configurations was deemed implausible.

The configurations with inconsistent results indicate that none of the analysed factors was individually sufficient; successful (or unsuccessful) translation was not consistently present with occurrence of any of the analysed predictive factors. We further checked configurations for individual necessary conditions. None of the included factors was always present (or absent) when successful translation occurred, indicating that none of the included factors was individually necessary for translational success. Similarly, none of the included factors was always present (or absent) when translational failure occurred, so none of the individual factors was necessary for translational failure either.

Following our conservative approach excluding inconsistent configurations, the only consistent configuration corresponds with the solution from the logical minimisation process, i.e., the following formula for translational success:

~Old*~Intervention*~Large*MultSpec*Quantitative -> Success

This means that the combination of relative recency (~ means not), analyses at event or study level, n < 75, inclusion of more than one species and quantitative analyses resulted in successful translation

Table 2
Number of cases per set.

Set name	N IN set (%)
Old	30 (29%)
Intervention	93 (89%)
Large	15 (14%)
MultSpec	26 (25%)
Quantitative	73 (70%)

Table 3
Truth table. Bold: configuration consistent with translational success. *Italics*: Configuration with inconsistent results. E.g., in configuration 2, the second line, the first one in *italics,* there are 5 cases with this configuration of potential predictors, of which 4 show translational success, and 1 translational failure. Underlined: configuration consistent with translational failure. Plain text: logical remainders.

Configuration	Old	Intervention	Large	MultSpec	Quantitative	N cases (n successful translation)	Cases
1	0	0	0	0	0	0 (0)	-
2	*0*	*0*	*0*	*0*	*1*	*5 (4)*	*Sultan_2017 (5x)*
3	0	0	0	1	0	0 (0)	-
4	0	0	0	1	1	2 (2)	Yen_2014 (2x)
5	0	0	1	0	0	0 (0)	-
6	0	0	1	0	1	0 (0)	-
7	0	0	1	1	0	2 (0)	Olson_2000 (2x)
8	0	0	1	1	1	0 (0)	-
9	*0*	*1*	*0*	*0*	*0*	*10 (3)*	*Dong_2011, DeBuck_2007, Evans_2006, Sanoh_2014, Tang_2005, Wang_2010, Ward_2005, Ward_2008 (3x)*
10	*0*	*1*	*0*	*0*	*1*	*33 (17)*	*Akabane_2010, Cao_2006, Cheng_2008 (2x), Chiou_2000a, Chiou_2000b (2x), Chiou_2002 (2x), Jones_2012 (2x), Grime_2013, Jones_2016 (4x), Kalvass_2007 (2x), Lennernas_2007, Ling_2009, Musther_2014, Paine_2011, Rocchetti_2007, Sanoh_2012, Walton_2004, Ward_2009 (2x), Whiteside_2008, Whiteside_2010, Wong_2004 (4x)*
11	*0*	*1*	*0*	*1*	*0*	*4 (2)*	*Corpet_2005, Fagerholm_2007a, Fourches_2010, Goteti_2010*
12	*0*	*1*	*0*	*1*	*1*	*5 (4)*	*Mahmood_2001, Mahmood_2004, Mahmood_2013 (2x) Wajima_2002*
13	*0*	*1*	*1*	*0*	*0*	*7 (2)*	*Monticello_2017 (3x), Nagilla_2004, Weaver_2003 (3x)*
14	0	1	1	0	1	2 (0)	Musther_2014 (2x)
15	*0*	*1*	*1*	*1*	*0*	*4 (2)*	*Ennever_2003, Fourches_2010, Monticello_2017 (2x)*
16	0	1	1	1	1	0 (0)	-
17	1	0	0	0	0	1 (0)	Litchfield_1961
18	1	0	0	0	1	0 (0)	-
19	1	0	0	1	0	1 (0)	Steinberg_1987
20	1	0	0	1	1	0 (0)	-
21	1	0	1	0	0	0 (0)	-
22	1	0	1	0	1	0 (0)	-
23	1	0	1	1	0	0 (0)	-
24	1	0	1	1	1	0 (0)	-
25	1	1	0	0	0	1 (0)	Schein_1973b
26	*1*	*1*	*0*	*0*	*1*	*19 (10)*	*Boxenbaum_1982 (4x), Chiou_1998 (2x), Crouch_1979 (2x), Fagerholm_1996, Freireich_1966 (4x), He_1998 (2x), Schneider_1999 (2x), Sietsema_1989 (2x)*
27	1	1	0	1	0	1 (0)	Schein_1973a
28	*1*	*1*	*0*	*1*	*1*	*7 (4)*	*Bachmann_1989, Mahmood_1996b, Mahmood_1996c, Mahmood_1998a, Mahmood_1998b (2x), Sietsema_1989*
29	1	1	1	0	0	0 (0)	-
30	1	1	1	0	1	0 (0)	-
31	1	1	1	1	0	0 (0)	-
32	1	1	1	1	1	0 (0)	-

(>85%).

Further evaluation of the two cases consistent with this formula shows that they were both derived from the same publication; they are two meta-analyses (for different outcomes) including both animal and human data (further described in the discussion). The results from both meta-analyses showed a high degree of overlap between the animal and the human data.

A separate QCA on the reverse outcome results in the following formula for translational failure:

Old*~Large*~Quantitative + ~Old*~-Intervention*Large*MultSpec*~Quantitative + ~Old*-Intervention*Large*~MultSpec*Quantitative -> Fail

This formula arose from logical minimisation of the 6 configurations with a consistent negative outcome in Table 3. It shows that there are 3 combinations of factors that consistently combine with translational failure; first, old, small studies using binary definitions of translation, second, newer large studies at the event or study level analysing multiple species using binary definitions of translation, and third, newer large studies at the intervention level analysing single species using quantitative definitions of translation.

3.2. RGA

From the 232 cases from our original review, the 197 cases without missing data for the analysed variables were included in the RGA. The RGA included five explanatory variables, corresponding to the sets included in the QCA, and an interaction term. Two of these variables were numerical and three categorical-binary. The observed values of the variables are summarised in Table 4.

A summary of the RGA is provided in Table 5. The effect estimates of the individual variables are relatively modest, but note that the effect sizes for publication age and study size are per year/ observation. The interaction term had a relatively large effect estimate, consistent with our QCA, indicating that analysing multiple species in combination with analysing translational success quantitatively may be optimal. However, none of the individual variables, nor this interaction, statistically

Table 4
Summary of data included in the RGA.

Variable	Corresponding QCA set	Type	N (0)	N (1)	Median (range)
Year of Publication	Old	Continuous			2005 (1961–2018)
Intervention	Intervention	Binary	14	183	
Study size	Large	Continuous			21 (4–951)
MultSpec	MultSpec	Binary	150	47	
Quantitative	Quantitative	Binary	87	110	
Translational success	OUT	Continuous			64 (0–100%)

Table 5
Summary of the RGA. SE: Standard Error.

Variable	Estimate	SE	p
(Intercept)	-77.0	301.6	0.80
Year of Publication	0.07	0.15	0.65
Intervention	-1.3	7.3	0.86
Study size	-0.02	0.02	0.65
MultSpec	1.1	6.6	0.87
Quantitative	2.5	4.3	0.56
MultSpec* Quantitative	10.3	8.9	0.25

affected translational success in the RGA.

4. Discussion

We initiated these analyses as a first exploration and proof of principle of the QCA-method in meta-research of animal-to-human translation. QCAs have successfully been performed on data from systematic literature reviews in other fields (Hanckel et al., 2021; Roig-Tierno et al., 2017; Melendez-Torres et al., 2018; Thomas et al., 2014). However, to the best of our knowledge, we are the first to perform a QCA with animal metadata, and to use it to analyse animal-to-human translation.

Our QCA resulted in a preliminary success formula for translational success at the meta-level; recent small reviews with analyses at the event or study level including more than one species and using a quantitative definition of translation were consistent with successful translation. While the effect sizes and directions of the RGA were consistent with these results, hence supportive of the QCA, the RGA did not identify any of the variables, nor the interaction term, as statistically significant. This shows the strength of the QCA approach.

4.1. Cases consistent with the QCA-derived formulae

The formula for translational success was based on 2 meta-analyses, which both came from the same paper (Yen et al., 2014). The authors performed an in-depth systematic review on guided tissue regeneration for periodontal infrabony lesions. They included 13 human and 9 animal papers, with varying study quality scores. The approach in their paper can be considered exemplary in synthesising animal and human data; combining them into a sub-grouped meta-analysis of percentages of bone filling, allowing for cross-species comparisons.

The formula for translational failure was based on 4 cases from 2 papers including newer studies (Musther et al., 2014; Olson et al., 2000), combined with 4 papers combining into a single term for older studies (Litchfield, 1961; Schein and Anderson, 1973; Schein et al., 1973; Steinberg and Schlesselman, 1987). To start with the newer studies; Olson et al (Olson et al., 2000). described large analyses of animal studies in dogs, primates, rats, mice and guinea pigs. The authors were fairly optimistic in describing that 71% of human adverse events was somehow predicted in an animal model, but they also detailed low concordance rates in toxicity. Musther et al (Musther et al., 2014). described correlational analyses of oral bioavailability, and concluded that bioavailability in animals is not predictive of that in humans. They provided separate data for mice (30 compounds), rats (122 compounds) and dogs (125 compounds), which were separately included in our analyses. Their monkey data (41 compounds) were included in the RGA, but excluded from the QCA because of an intermediate translational success rate.

To continue with the older studies; Litchfield (Litchfield, 1961) concluded that many serious side effects that can occur when a drug is given to humans were not predictable from observations on dogs or rats. The rat data were included in the QCA as clear translational failure, the dog data were only included in the RGA because of intermediate translational success. Steinberg & Schlesselman (Steinberg and Schlesselman, 1987) compared the effects of pancreatitis therapeutics between 13 human studies and 25 animal studies in dogs, pigs, rats and guinea pigs, with low correspondence between the results.

Schein co-authored two publications in 1973 that both described multiple analyses included in our RGA, most with translational success rates between 45% and 80%. In one publication, Schein and Anderson carefully concluded that combining data from multiple species could reduce false negatives for prediction of human adverse events, but one of their data sets reflected translation below 45% and was included in our QCA (Schein and Anderson, 1973). In the other publication, Schein et al. concluded that animal models can predict a substantial part of the adverse events occurring in clinical use (Schein et al., 1973), but again, translation was low in one of their data sets which we included in the QCA.

The term in the formula for translational failure that combines the 4 configurations listing these older studies (Old*~Large*~Quantitative) effectively illustrates the concept of logical minimisation, and thereby the potential of the QCA-method.

The formula we here present for translational success is restricted to smaller studies, and two of the three terms in our formula for translational failure cover large studies. Because underpowered studies can result in erroneous conclusions (Ioannidis, 2005), translational success being related to smaller studies may seem counterintuitive. However, we analysed study size, and not actual power (which was rarely known). For smaller studies, particularly when they are adequately powered, increased familiarity with the data at the individual case level might benefit the quality of the work, e.g., by decreasing the error rate and by improving the interpretation of the findings. This, in specific configurations, could positively affect translational success.

4.2. Suggestions for future QCAs

The here-presented data were not collected with QCA in mind, and therefore not optimised for this approach. However, they still resulted in preliminary formulae consistent with translational success and failure, based on relatively few consistent configurations. We expected contradictory lines in the truth table; configurations that were not consistent with the outcome. We hypothesise that this is mainly due to not all relevant factors being included in this QCA.

Future studies should gather data for more factors, but also on a larger number of cases to fill the logical remainders. A 2-step approach is considered; a first large QCA could comprise multiple factors relating to the meta-level. A second QCA could be restricted to the successful configurations from the first, and address factors at the primary study level. With more cases included, and less concern about logical remainders, multivalue QCAs (mvQCAs) (Thiem, 2013) may well be preferable. With mvQCA, it is possible to have multiple values per variable instead of strict dichotomisation. For the here-described dataset, it would be advantageous to distinguish studies at the event and the study level instead of pooling them together, outside (next to) the set of studies at the intervention level. While it may seem like an attractive idea to include a factor for individual species, the resulting truth table would become incredibly large and have many logical remainders for the less frequently used species. However, a category distinguishing e.g., rodents, non-human primates and other mammals could be viable for future work.

QCAs can also be applied to other types of data than literature (Befani, 2013; Ragin, 1999; Vink and Van Vliet, 2009), which may make other types of data and variables accessible for analyses. E.g., communication and consideration of all available data in experimental design can be added as factors, or individual compounds or targets can be defined as cases in a QCA. Medical research fields could be another type of case within a QCA. A recent study from our group showed variation in translational success rates between medical fields (Van de Wall et al., 2023), and QCA could be valuable to analyse the effects of differences in practice between fields. Of note, contrary to common thoughts, translation in the field of neuroscience is not worse than in other fields (Van de Wall et al., 2023). Alternatively, cases could be individual clinical trials, and a QCA could focus on back-translation from clinical trial results to the animal data. There are indications that animal data are insufficiently considered in the design of human trials (Sievers et al., 2021; Wieschowski et al., 2018), which might partially explain translational failure. Data could be gathered from multiple sources comprising also investigators' brochures, ethics applications and patent registrations. QCA analyses including human or animal studies as cases seem most promising, as they would allow for analyses of combinations of quality measures such as blinding, randomization and sample size calculations, in specific settings. Eventually, this type of QCA might even result in actual guidance for future studies, which should focus on those quality measures that are most important to successful translation.

4.3. Implications and conclusion

While our results are not conclusive and need confirmation, analysing multiple species in combination with analysing translational success quantitatively may be the optimal approach for future animal studies aiming for reproduction in humans. Analysing animal-to-human translation quantitatively as a percentage of correspondence instead of making simplified binary yes/no distinctions fairly reflects the available data. While we do not encourage increasing the number of animal studies overall, if a study aiming at translation is considered to be necessary, we may need to get used to the idea of testing more than one species.

In this paper, we present the first QCAs addressing translational success and failure rates. While the data were not collected with this method in mind, we show that the approach is viable, relevant and promising. Further knowledge on animal-to-human translation may help to improve reproducibility in research and drug development, and to focus animal studies to where they are predictive for humans.

Ethics approval and consent to participate

Not applicable.

Funding

ZonMW (40-42600-98-417), NWO (313-99-310) and the Federal State of Lower Saxony (R2N) provided funding for this project.

CRediT authorship contribution statement

Cathalijn H.C. leenaars: Conceptualisation, Formal analysis, Investigation, Visualisation, Interpreted the data, Writing - original draft. **Steven Teerenstra:** Conceptualization, Formal analysis, Validation, Methodology, Writing - original draft. **Franck Meijboom:** Interpreted the data, Writing - review and editing, Provided the funding. **André Bleich:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing. CL & ST designed the analyses. CL performed the analyses. All authors interpreted the data. CL wrote the first draft of the manuscript, with major contributions from ST. FM and AB provided the funding. All authors read and approved the final manuscript.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Data availability

All data synthesised in this manuscript are already in the public scientific domain.

Acknowledgements

Not applicable.

Consent for publication

Not applicable.

Preprints

This work has not been submitted for publication elsewhere but has been posted on MedRXiv (Pharmacology & therapeutics, MEDRXIV/2022/270227, doi: 10.1101/2022.01.31.22270227).

References

Anon, R: A language and environment for statistical computing. R Foundation for Statistical Computing, [⟨https://www.R-project.org/⟩. Accessed [24-01-2022]].
Aus, J.P., 2009. Conjunctural causation in comparative case-oriented research. Qual. Quant. 43, 173–183.
Baker, E.J., Beck, N.A., Berg, E.L., Clayton-Jeter, H.D., Chandrasekera, P.C., Curley, J.L., Donzanti, B.A., Ewart, L.C., Gunther, J.M., Kenna, J.G., et al., 2019. Advancing nonclinical innovation and safety in pharmaceutical testing. Drug Discov. Today 24, 624–628.
Befani, B., 2013. Between complexity and generalization: addressing evaluation challenges with QCA. Evaluation 19, 269–283.
dplyr: A Grammar of Data Manipulation. R package version 1.0.3. 2023.
Dusa, A., 2019. QCA with R. A Comprehensive Resource. Springer International Publishing,.
Faggion Jr., C.M., Chambrone, L., Gondim, V., Schmitter, M., Tu, Y.K., 2010. Comparison of the effects of treatment of peri-implant infection in animal and human studies: systematic review and meta-analysis. Clin. Oral. Implants Res. 21, 137–147.
Genzel, L., Adan, R., Berns, A., van den Beucken, J., Blokland, A., Boddeke, E., Bogers, W. M., Bontrop, R., Bulthuis, R., Bousema, T., et al., 2020. How the COVID-19 pandemic highlights the necessity of animal research. Curr. Biol. 30, R1014–R1018.
Hackam, D.G., Redelmeier, D.A., 2006. Translation of research evidence from animals to humans. JAMA 296, 1731–1732.
Hanckel, B., Petticrew, M., Thomas, J., Green, J., 2021. The use of Qualitative Comparative Analysis (QCA) to address causality in complex systems: a systematic review of research on public health interventions. BMC Public Health 21, 877.
Herrmann, K., Kimberly, J., 2019. Animal Experimentation: Working Towards a Paradigm Change. Brill,.
Hobson-West, P., 2010. The role of 'public opinion' in the UK animal research debate. J. Med. Ethics 36, 46–49.
Ioannidis, J.P., 2005. Why most published research findings are false. PLoS Med. 2, e124.
Jacobson, R.M., Adegbenro, A., Pankratz, V.S., Poland, G.A., 2001. Adverse events and vaccination-the lack of power and predictability of infrequent events in pre-licensure study. Vaccine 19, 2428–2433.
Kast, F.E., Rosenzweig, J.E., 1972. General systems theory:Applications for organization and management. Acad. Management J. 15, 447–465.
Kola, I., Landis, J., 2004. Can the pharmaceutical industry reduce attrition rates. Nat. Rev. Drug Discov. 3, 711–715.
Leenaars, C., Stafleu, F., de Jong, D., van Berlo, M., Geurts, T., Coenen-de Roo, T., Prins, J.B., Kempkes, R., Elzinga, J., Bleich, A., et al., 2020. A systematic review comparing experimental design of animal and human methotrexate efficacy studies for rheumatoid arthritis: lessons for the translational value of animal studies. Animals 10.
Leenaars, C.H.C., Kouwenaar, C., Stafleu, F.R., Bleich, A., Ritskes-Hoitinga, M., De Vries, R.B.M., Meijboom, F.L.B., 2019. Animal to human translation: a systematic scoping review of reported concordance rates. J. Transl. Med. 17, 223.
Litchfield Jr, J.T., 1961. Forecasting drug effects in man from studies in laboratory animals. JAMA 177, 34–38.
Mahmood, I., 2006. Prediction of human drug clearance from animal data: application of the rule of exponents and 'fu Corrected Intercept Method' (FCIM). J. Pharm. Sci. 95, 1810–1821.
Melendez-Torres, G.J., Sutcliffe, K., Burchett, H.E.D., Rees, R., Richardson, M., Thomas, J., 2018. Weight management programmes: Re-analysis of a systematic review to identify pathways to effectiveness. Health Expect. 21, 574–584.
Musther, H., Olivares-Morales, A., Hatley, O.J., Liu, B., Rostami Hodjegan, A., 2014. Animal versus human oral drug bioavailability: do they correlate. Eur. J. Pharm. Sci. 57, 280–291.
Naims, H., Eppinger, E., 2022. Indicator-driven data calibration of expert interviews in a configurational study. MethodsX 9, 101699.
Olson, H., Betton, G., Robinson, D., Thomas, K., Monro, A., Kolaja, G., Lilly, P., Sanders, J., Sipes, G., Bracken, W., et al., 2000. Concordance of the toxicity of pharmaceuticals in humans and in animals. Regul. Toxicol. Pharmacol. 32, 56–67.
P. Schein R.D. Davis D.A. Cooney Qualitative aspects of drug toxicity in prediction from laboratory animals to man *5th Int. Congr. Pharmacol.* 1973 304 335.
Pound, P., Ritskes-Hoitinga, M., 2018. Is it possible to overcome issues of external validity in preclinical animal research? why most animal models are bound to fail. J. Transl. Med 16, 304.
Ragin, C.C., 1999. Using qualitative comparative analysis to study causal complexity. Health Serv. Res. 34, 1225–1239.
readxl: Read Excel Files. R package version 1.3.1. 2023.
Roig-Tierno, N., Gonzalez-Cruz, T.F., Llopis-Martinez, J., 2017. An overview of qualitative comparative analysis: a bibliometric analysis. J. Innov. Knowl. 15–23.

Schein, P., Anderson, T., 1973. The efficacy of animal studies in predicting clinical toxicity of cancer chemotherapeutic drugs. Int J. Clin. Pharmacol. 8, 228–238.

Sievers, S., Wieschowski, S., Strech, D., 2021. Investigator brochures for phase I/II trials lack information on the robustness of preclinical safety studies. Br. J. Clin. Pharmacol. 87, 2723–2731.

Skaaning, S., 2011. Assessing the robustness of crisp-set and fuzzy-set QCA results. Sociol. Methods Res. 40, 391–408.

Steinberg, W.M., Schlesselman, S.E., 1987. Treatment of acute pancreatitis. comparison of animal and human studies. Gastroenterology 93, 1420–1427.

Thiem, A., 2013. Clearly Crisp, and Not Fuzzy: a reassessment of the (Putative) Pitfalls of Multi-value QCA. Field Methods 25, 197–207.

Thomas, J., O'Mara-Eves, A., Brunton, G., 2014. Using qualitative comparative analysis (QCA) in systematic reviews of complex interventions: a worked example. Syst. Rev. 3, 67.

Van de Wall, G., Van Hattem, A., Timmermans, J., Ritskes-Hoitinga, M., Bleich, A., Leenaars, C., 2023. Comparing translational success rates across medical research fields - a combined analysis of literature and clinical trial data. ALTEX.

Vink, M.P., Van Vliet, O., 2009. Not Quite Crisp, Not Yet Fuzzy? assessing the potentials and pitfalls of multi-value QCA. Field Methods 21, 265–289.

Wieschowski, S., Chin, W.W.L., Federico, C., Sievers, S., Kimmelman, J., Strech, D., 2018. Preclinical efficacy studies in investigator brochures: do they enable risk-benefit assessment. PLoS Biol. 16, e2004879.

Yen, C.C., Tu, Y.K., Chen, T.H., Lu, H.K., 2014. Comparison of treatment effects of guided tissue regeneration on infrabony lesions between animal and human studies: a systematic review and meta-analysis. J. Periodontal Res. 49, 415–424.