bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.887919; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

What influences treatment response in animal models of non-alcoholic fatty liver disease? A meta-analysis with meta-regression.

Harriet Hunter*[1], Dana de Gracia Hahn*[1], Amedine Duret*[1], Yu Ri Im*[1], Qinrong Cheah[1], Jiawen Dong[1], Madison Fairey[1], Clarissa Hjalmarsson[1], Alice Li[1], Hong Kai Lim[1], Lorcán McKeown[1], Claudia-Gabriela Mitrofan[1], Raunak Rao[1], Mrudula Utukuri[1], Ian A. Rowe[2], Jake P. Mann[3]

Affiliations

1. School of Clinical Medicine, University of Cambridge, Cambridge, UK
2. Leeds Institute for Medical Research & Leeds Institute for Data Analytics, University of Leeds, Leeds, UK.
3. Institute of Metabolic Science, University of Cambridge, Cambridge, UK.

*These authors contributed equally.

Correspondence:
Dr. Jake P Mann
Metabolic Research Laboratories
Level 4, Institute of Metabolic Science
Addenbrooke's Hospital
Cambridge, CB2 0QQ
jm2032@cam.ac.uk
Tel: +44 1223 336792
Fax: +44 1223 330598

Funding
JPM is supported by a Wellcome Trust fellowship (216329/Z/19/Z).

Competing interests
The authors have no conflicts of interest to declare.

Keywords
NAFLD, meta-research, rodent models, metabolic syndrome, 3Rs

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.887919; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Abstract

The classical drug development pipeline necessitates studies using animal models of human disease to gauge future efficacy in humans, however, there is a comparatively low conversion rate from success in animals to in humans. Non-alcoholic fatty liver disease (NAFLD) is a complex chronic disease without any licensed therapies and hence a major field of animal research. We performed a meta-analysis of 414 interventional rodent studies (6,575 animals) in NAFLD to assess the mean difference in hepatic triglyceride content. 20 of 21 studied drug classes had similar efficacy with a mean difference of -30% hepatic triglyceride. However, when publication bias was accounted for, this reduced to -16% difference. Study characteristics were only able to account for a minority of variability on meta-regression, and we replicated previous findings of high risk of bias across 82% of cohorts. These findings build on previous work in preclinical neuroscience and help to explain the challenge of reproducibility and translation within the field of metabolism.

Introduction

Interventional studies in animals are an integral component of the drug development pipeline. If a disease can be suitably modelled in an animal, then the therapeutic response to a treatment observed in animals should inform about its potential efficacy in humans[1]. However, there is a well-documented translational gap between preclinical studies and subsequent outcomes in humans[2–4]. Multiple factors contribute to this, including bias within study design[5], insufficiently powered preclinical studies[6], and biological differences between species[7,8].

Systematic analyses of preclinical studies, predominantly in the field of neuroscience, have found that publication bias may account for at least a third of the estimate of efficacy in trials[9,10]. In addition, other variables of animal model design can influence the magnitude of the treatment response[11]. These findings are highly relevant in the context of the 'reproducibility crisis'[12,13] as well as having ethical implications of the use of animals in research that is not of optimum quality[14].

Non-alcoholic fatty liver disease (NAFLD) is a highly active field of animal research[15,16]. NAFLD is a common condition characterised by increased liver fat (hepatic steatosis) that may progress to inflammation (non-alcoholic steatohepatitis (NASH)) and fibrosis[17]. Cirrhosis, end-stage liver disease, and hepatocellular carcinoma develop in a small proportion of patients. However, due to the high prevalence of obesity, NAFLD is the second most common indication for liver transplant in the United States[18], predicted to overtake hepatitis C. NAFLD is intricately related with insulin resistance and therefore usually coexists with other features of the metabolic syndrome, such as type 2 diabetes and its recognised complications cerebrovascular disease, coronary artery disease and chronic kidney disease[19].

There are currently no approved pharmacological therapies for NAFLD[20]. Several Phase 3 trials are ongoing[21], but many interventions that appeared to have substantial efficacy in

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.887919; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

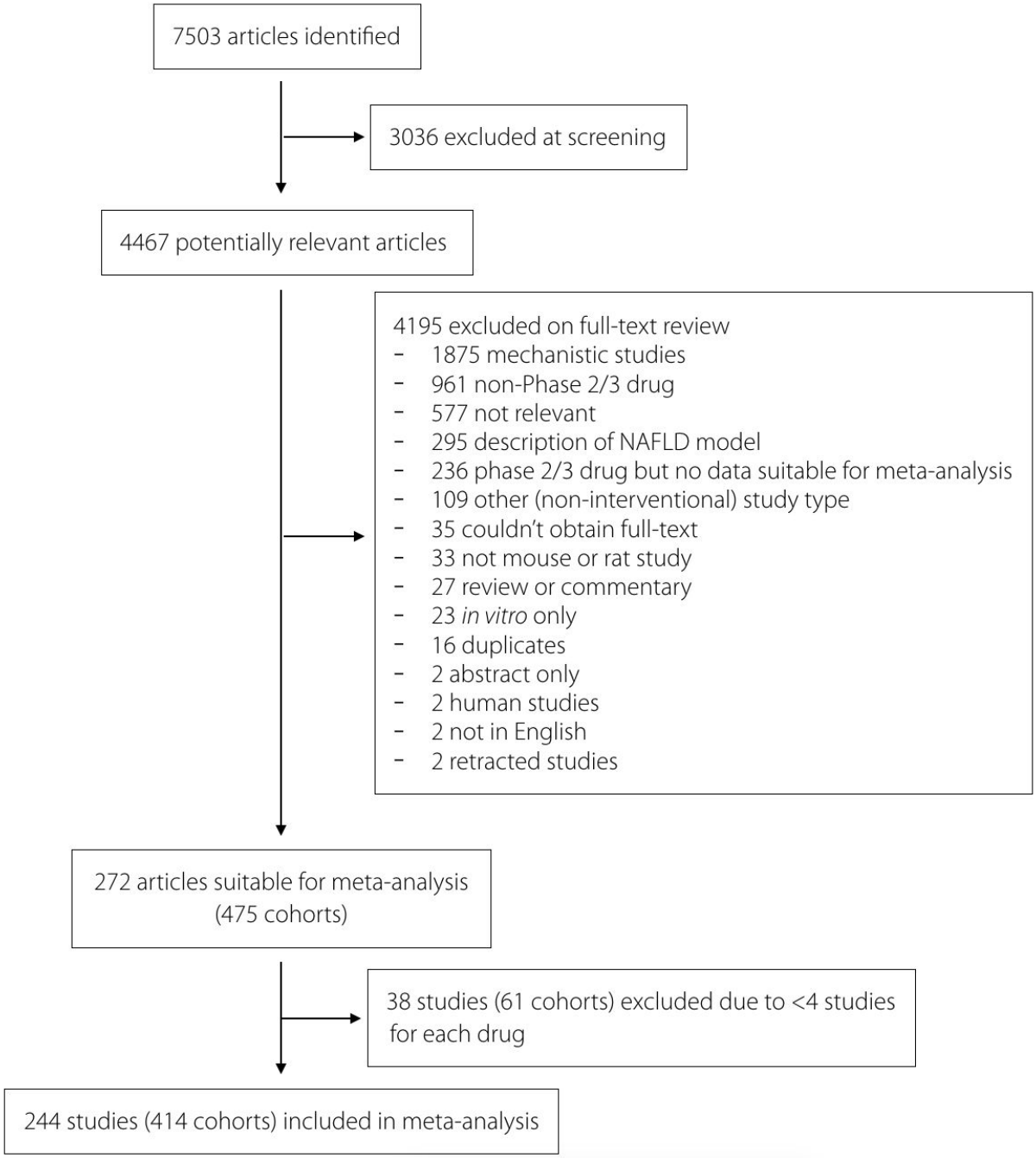
preclinical models have failed to be replicated in humans[22–24]. These studies have used a wide range of preclinical NAFLD models, including genetically modified animals (e.g. leptin deficient *ob/ob* mice), hypercaloric diets (e.g. high fat diet), and toxic insults (e.g. streptozocin injections), all of which may be used in varying combinations[25]. It is not known if, or which of, these variables influence treatment response to therapeutic agents in preclinical models of NAFLD.

Therefore, we performed a meta-analysis of interventional rodent studies of NAFLD to describe which drug classes were associated with the greatest reduction in liver fat and whether any study characteristics (or biases) were linked to the magnitude of effect.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.887919; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Results

We performed a systematic search to identify interventional studies in rodent models of NAFLD. Our searches yielded 7503 articles, which after screening gave 4467 articles for full-text review (Figure 1). Studies were included in the meta-analysis if they used a pharmacological agent that had been used in Phase 2 or 3 trials in humans for NAFLD and reported hepatic triglyceride content for control and interventional groups. 244 studies were included in the meta-analysis, comprising 414 cohorts of rodents (6,575 animals).



bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.887919; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Fig. 1: *Study inclusion and exclusion flow chart.*

21 drug classes were represented in the meta-analysis, including 34 studies on GLP-1 agonists, 15 on probiotics, and 55 on fibrates.

Meta-analysis of hepatic triglyceride content

We used random-effects meta-analysis to estimate the mean difference in hepatic triglyceride (TG) content between intervention and control groups. The overall mean change in hepatic TG content was -30.4% (95% CI -33.0%, -27.7%) with considerable between-study heterogeneity (I^2 = 91% [95% CI 90.3-91.6%]).

We hypothesised that much of this heterogeneity would be due to the different drug class interventions. However, on meta-analysis using drug class as a sub-group, we found marked similarity between the effect size of the different interventions (Figure 2). There was weak evidence of difference between drug classes (Q=36.4, p=.014). The confidence intervals of 20 out of 21 drug classes overlapped and there remained substantial or considerable heterogeneity within drug class subgroups (Table 1).

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.887919; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

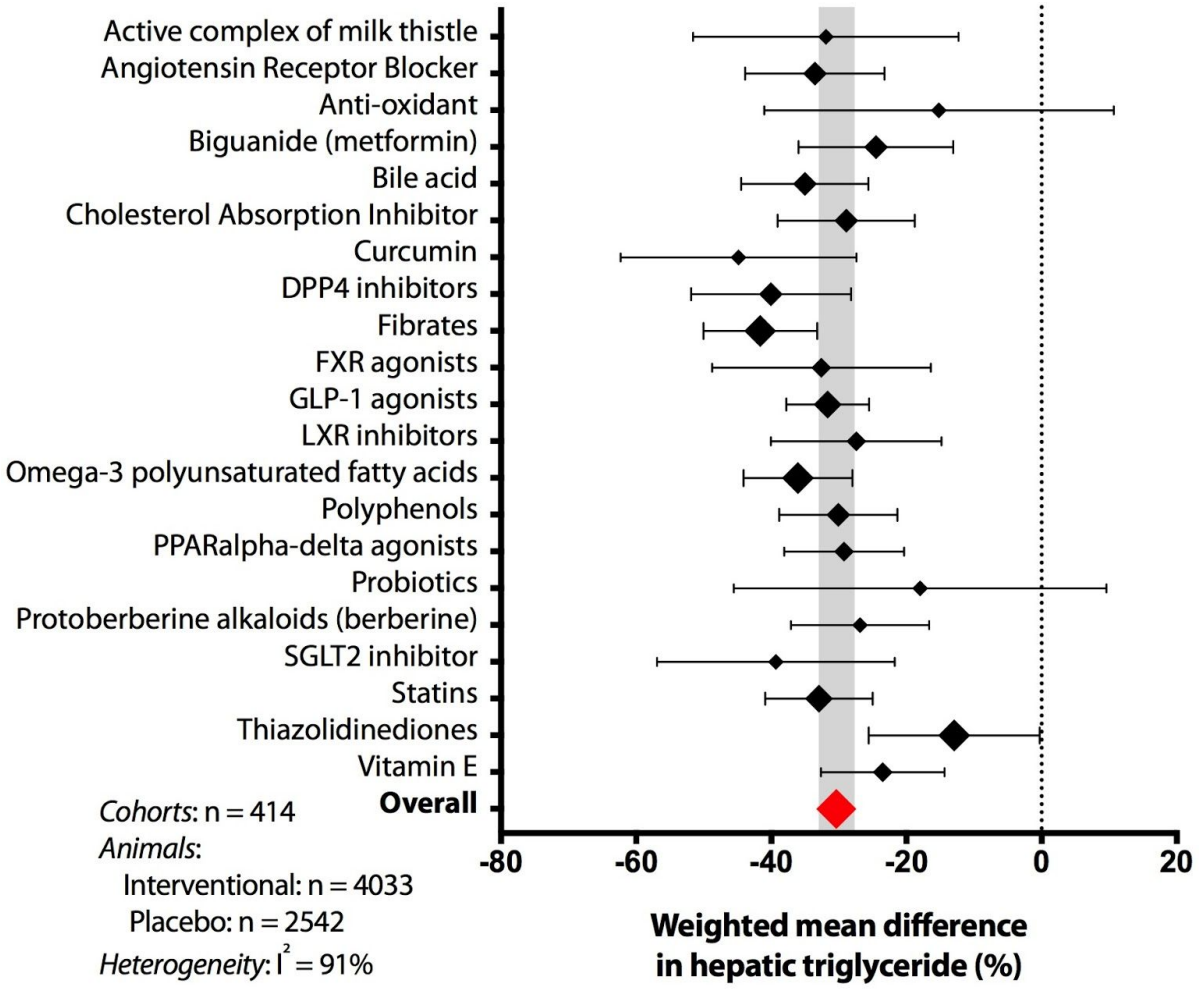


Fig 2. *Summary forest plot showing mean difference in hepatic triglyceride content between interventional and control animals in rodent studies of NAFLD. Individual studies have been hidden and only subgroup summary figures are illustrated.*

Drug class	**k (cohorts)**	**Mean difference in hepatic TG (95% CI)**	**I^2**	**τ^2**	**Cochrane's Q**
Active complex of milk thistle	7	-31.9 (-51.6, -12.3)	57.6%	277	14.2
Angiotensin Receptor Blocker	18	-33.6 (-43.9, -23.2)	77.6%	313	75.9
Anti-oxidant	8	-15.2 (-41.0, 10.7)	68.6%	723	22.3
Biguanide	24	-24.5 (-35.9, -13.1)	89.9%	562	227.0

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.887919; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Bile acid	17	-35. (-44.4, -25.6)	74.6%	230	63.0
Cholesterol Absorption Inhibitor	20	-28.9 (-39.1, -18.7)	82.5%	355	108.9
Curcumin	5	-44.9 (-62.3, -27.4)	57.4%	129	9.4
DPP4 inhibitor	10	-40.0 (-51.8, -28.2)	31.1%	168	13.1
Fibrates	44	-41.6 (-50.0, -33.2)	85.3%	613	292.1
FXR agonist	10	-32.6 (-48.8, -16.4)	56.9%	338	20.9
GLP-1 agonist	34	-31.7 (-37.8, -25.5)	90.8%	196	359.7
LXR inhibitor	10	-27.4 (-40.0, -14.8)	79.1%	142	43.0
Omega-3 polyunsaturated fatty acids	46	-36. (-44.1, -28.0)	89.3%	589	418.9
Polyphenol	23	-30.1 (-38.8, -21.3)	92.1%	315	278.4
PPARalpha-delta agonist	11	-29.2 (-38.1, -20.3)	29.0%	93	12.7
Probiotics	15	-18.0 (-45.6, 9.6)	78.9%	1670	66.4
Protoberberine alkaloid	10	-26.9 (-37.1, -16.6)	45.7%	134	16.6
SGLT2 inhibitor	6	-39.3 (-56.9, -21.7)	31.6%	159	7.3
Statin	36	-32.9 (-40.9, -25.0)	94.5%	406	638.2
Thiazolidinediones	50	-12.9 (-25.6, -0.3)	94.4%	1816	869.3
Vitamin E	11	-23.5 (-32.6, -14.4)	21.5%	106	12.7

Table 1: *Summary of the studies by drug class with estimates of within sub-group heterogeneity. DPP4, dipeptidyl peptidase-4; FXR, farnesoid X receptor; GLP-1, glucagon-like peptide-1; LXR, liver X receptor; PPAR, peroxisome proliferator-activated receptor; SGLT2, sodium-glucose transport protein 2.*

In order to test whether the heterogeneity was driven by individual outlying studies, we used a Baujat plot to identify outliers with disproportionate contribution to heterogeneity (Supplementary Figure 1). After removal of seven studies, the overall mean difference was unchanged (-31% [95% CI -33%, -28%]) and considerable heterogeneity remained (I^2 = 86.5% [95% CI 85.3%, 87.5%]).

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.887919; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Meta-regression for difference in hepatic TG content

To try and understand the variability in mean difference of hepatic TG, we performed meta-regression using both categorical and continuous variables associated. On mixed-effects meta-regression, drug class accounted for only 5.6% of heterogeneity in hepatic TG (Table 2).

Variable(s)	**k (Number of cohorts)**	***β* (SE)**	**Test of moderators (p-value)**	**r^2**
Drug class	414	*Figure 2*	0.002	5.60%
Genetic background	337	*SupTable 1*	0.08	1.65%
Age at end of study (months)	392	1.45 (.60)	0.02	1.13%
Duration of intervention (weeks)	413	.42 (.22)	0.06	0.63%
Age at start of drug (months)	392	.94 (.73)	0.2	0.12%
Model	277	*SupTable 2*	0.37	0.11%
Quality score [0-4]	414	1.3 (1.8)	0.46	0
Male sex	382	-1.9 (26.6)	0.68	0
Female sex	382	3.3 (27.2)	0.68	0
% fat in diet	219	-.06 (.15)	0.66	0
Drug dose (mg/kg)	283	.005 (.007)	0.5	0

Table 2. *Results from univariable meta-regressions. β represents the change in mean difference in hepatic TG per unit change in variable. For categorical variables, the effect sizes are shown in Figure 2 (for drug class), Supplementary Table 1 (genetic background), and Supplementary Table 2 (model type).*

When other study characteristics and variables were assessed, genetic background (Supplementary Table 1) and age at the end of intervention (Figure 3) had a very modest impact upon mean difference in hepatic TG. The type of NAFLD model used accounted for almost no heterogeneity in the data (Supplementary Table 2).

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.887919; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

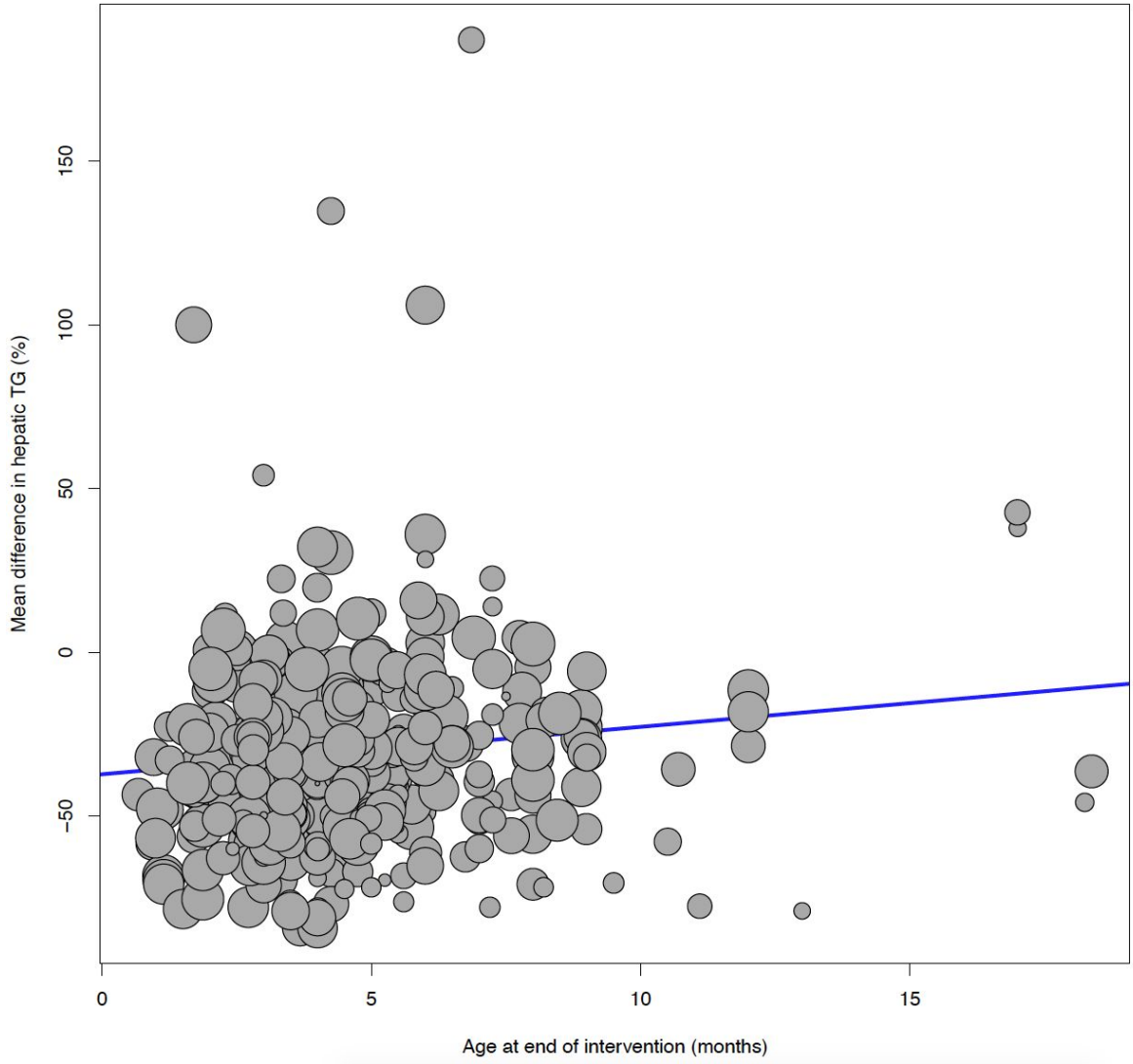


Figure 3. *Bubble plot illustrating the results of meta-regression between the age at end of intervention (in months) and mean difference in hepatic TG. Studies in older mice were weakly associated with smaller treatment responses: 1.4% less reduction in hepatic TG per month of life for the studied rodents.*

In order to estimate the total variance explained by measurable variables, we used multiple variable meta-regression in a smaller dataset (k = 71) with minimum 4 studies for each drug, genetic background, and NAFLD model. When all variables from Table 2 were included, this model was not associated with treatment response ($F_{25,45} = .07$, $p = .83$) though accounted for 47.4% of heterogeneity.

Publication (study distribution) bias

The meta-analysis result was assessed for study bias using a funnel plot (Figure 4). This showed an uneven distribution with a bias towards positive results (i.e. a reduction in hepatic

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.887919; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

TG with intervention), which was supported by Egger's test (β = -1.0 [95% CI -1.4, -.60], p = $6x10^{-5}$). Using the Trim and Fill method to account for this bias, we estimated that the true overall mean difference in hepatic TG would be half as great: -16.6% (95% CI -19.6, -13.7).

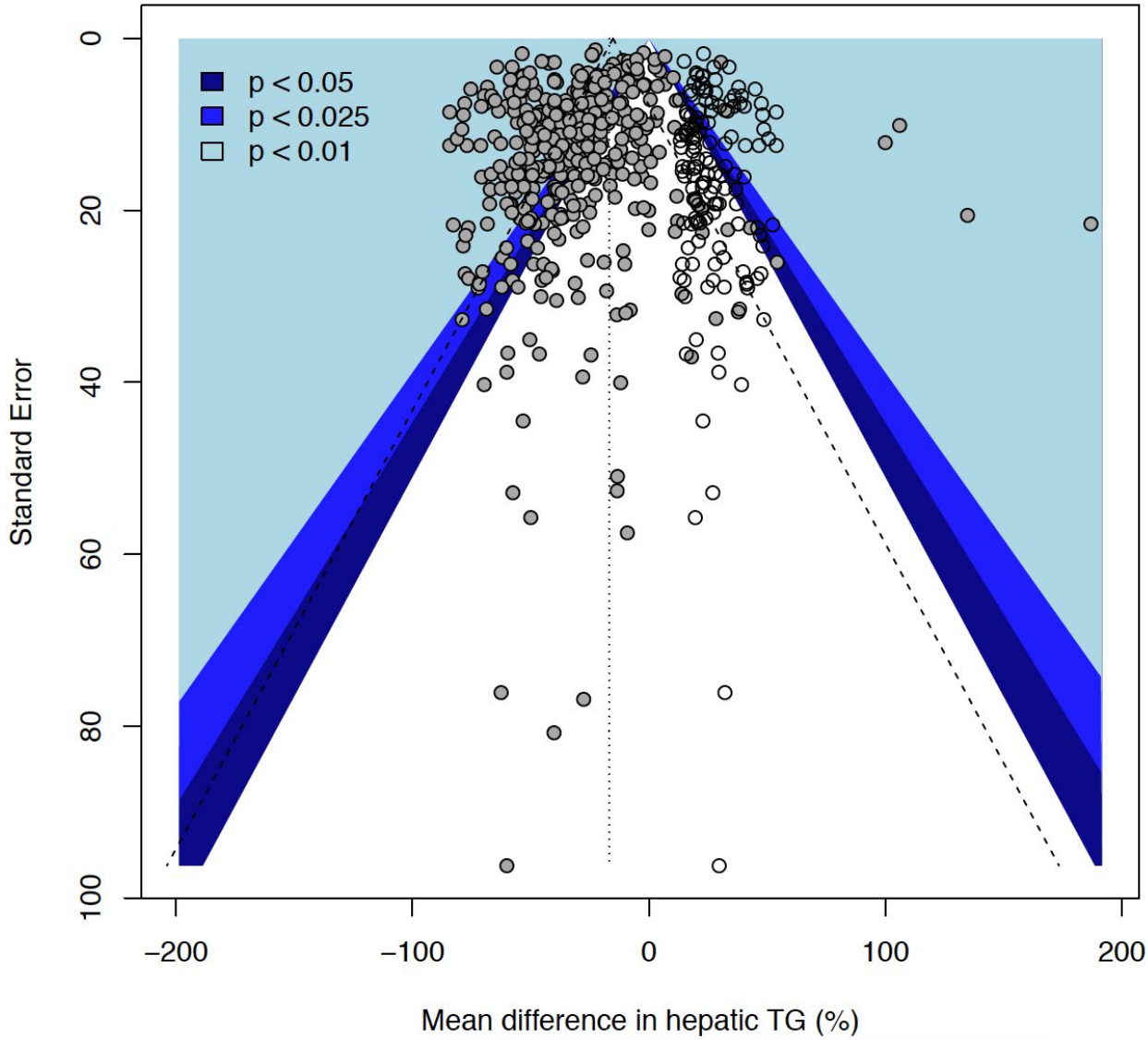


Fig 3. *Funnel plot illustrating study distribution (publication) bias in 414 original studies (solid grey circles) with 139 added studies (from trim and fill). The statistical significance associated with each study is illustrated with the coloured background.*

Study quality

We used a four-item scale to estimate study quality (Supplementary Figure 2). We found that 361/441 (82%) cohorts were at high risk of bias due to either absence of randomisation or absence of blinding. No individual component of the risk of bias score (or combined as a study quality score) was associated with magnitude of treatment response.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.887919; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

There was a bimodal distribution of study powers across the cohorts included in the meta-analysis (Supplementary Figure 3). 91% (375/414) cohorts had a power of 80% or greater on post-hoc calculation. However, using the results from this meta-analysis (mean difference -30.4% with standard deviation = 29.6), to achieve a power of 80% with significance set as p=.05, group size would need to be n=16. 3.4% (14/414) cohorts included 16 or more animals and would have met sufficient power to detect associations, based on these data.

Discussion

Through meta-analysis and meta-regression we have quantified the remarkable similarity between interventional drug classes and which study characteristics influence treatment response in rodent models of NAFLD. These findings suggest that measurable factors (e.g. animal age) can only account for a small degree of variability in animal response to drugs.

NAFLD is generally considered to be a major public health concern and drug development is a highly active field[26]. Though there are no licensed therapies, there have been over 30 drugs used in Phase 2 or 3 trials, with some demonstrating potential efficacy in well-conducted randomized controlled trials, such as for GLP-1 agonists[27] and pioglitazone[28]. Therefore, we were surprised to find that there was minimal difference between any of 21 drug classes studied in animal models (Figure 2). In fact, thiazolidinediones were the only outlier, due to studies that found an increase in hepatic TG content in response to treatment[29,30]. To our knowledge this is such similarity between multiple drugs in preclinical models has not been described before to this extent. It is generally consistent with findings reported in preclinical models of spinal cord injury where the effect size of several different types of treatment overlapped[11].

Despite overlapping confidence intervals for drug classes, there was considerable heterogeneity between studies. This is potentially expected as a consequence of including a relatively large number of cohorts in the meta-analysis (n=414). We were further surprised to find that drug class accounted for a modest amount of variability in treatment response. Similarly, it is well established that different rodent NAFLD models result in higher or lower hepatic TG concentrations[25], but our results do not suggest that they influence treatment response. Selection of animal models for studying NAFLD is a heavily debated topic[31] and a detailed discussion is beyond the scope of this meta-analysis.

Design of an interventional animal study is considerably more complex than just the ‘core’ model (e.g. leptin deficient (*ob/ob*) mice). There have been previous reports to illustrate the effect of genetic background on NAFLD[32,33] as well as in other fields, including immunology[34] and behavioural neuroscience[35,36]. We were surprised to find that genetic background accounts for only 1.6% heterogeneity in results. We anticipate that the true value is greater than this because we could only include backgrounds that had been used in multiple studies and we excluded mixed genetic backgrounds from analysis.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.887919; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Older animals had a smaller response to treatment, which is consistent with evidence that they develop more substantial NASH[37] and their physiology is less plastic[38]. However, it could be argued that younger animals do not accurately reflect chronic multisystem diseases such as NAFLD.

Using the trim and fill method, we estimated that study distribution bias (most likely publication bias in this case) may have almost doubled the reported magnitude of effect (16% reduction in hepatic TG compared to 30%). The presence of publication bias did not come as a surprise[39] and this dataset provides useful replication of the strong evidence base for this in preclinical neurological studies. The results from power calculations are also likely to reflect publication bias: based on the overall effect summary, only 3.4% of cohorts were of sufficient size to be predicted to achieve the power of 80%. However, over 90% of cohorts had results that would be consistent with power >80%. Similarly, we have replicated previously described low rates of randomization and blinding in animal studies[40].

There are several implications of these results. Firstly, it is not surprising that there are multiple reports of difficulty in reproducing preclinical studies of metabolism[13] if the choice of drug intervention plays a relatively minor role in determining the magnitude of treatment response. Variations in study design affects response and therefore could silence subtle differences, especially when combined with this risk of bias which can generate false positives. Secondly, these results also help to explain the difficulty in bridging the preclinical to human translational gap[41]. If almost all studied interventions are reported to have a similar effect in animals, then results to date are not helpful in guiding which will show efficacy in humans. This is clearly a challenge across the entire biomedical field and careful replication of animal study design is likely to be important. It could be suggested that 30% reduction in hepatic TG could be a ‘benchmark’ for future therapeutics aimed at reduced steatosis.

This apparent threshold for 30% reduction in hepatic TG does not appear to be due to the sensitivity of in vitro assays used as TG content was reported to range from 0.05μM/mg to 7.9μM/mg across the included studies.

Much of the variability in results could not be accounted for. We speculate that unmeasurable variables, such as technique of animal handling, may also influence the treatment effect. The bacterial status of mice is known to affect liver phenotypes[42], potentially via intestinal dysbiosis[43,44]; and similarly, whilst we included proportion of fat in diet as a variable, the complete diet composition is much more complex and may be of importance.

The main strength of this work is the number of included studies, interventions, and variables. This has facilitated a detailed analysis of a single disease area. Though the primary limitation of findings is that we are unable to determine the extent that these results are generalisable to other fields we have highlighted several conceptual similarities to previous meta-research in neuroscience. In addition, many studies did not report variables, for example genetic

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.887919; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

background of animals was reported in 7% (31/441), which reduced the number of studies included in meta-regression analyses.

Conclusion

Quantifiable and measurable variables in animal studies, including the drug used, have only a modest effect on the size of treatment response. There is a highly consistent 30% reduction in hepatic TG in all drugs tested to date, though the true magnitude of this value might be half as great when accounting for publication bias. Standardisation of study design and rigorous quality are needed in preclinical studies in metabolism to improve the translation and replicability of findings.

Methods

Review protocol and search strategy

The systematic review protocol was prospectively registered with SyRF (Systematic Review Facility) and is available from:
https://drive.google.com/file/d/0B7Z0eAxKc8ApQ0p4OG5SblRlRTA/view.

PubMed via MEDLINE and EMBASE were searched for published articles of experimental rodent models of fatty liver, NAFLD, or non-alcoholic steatohepatitis (NASH). The following search term was used: ("Non-alcoholic fatty liver disease" OR "Nonalcoholic fatty liver disease" OR "NAFLD" OR "non-alcoholic steatohepatitis" OR "nonalcoholic steatohepatitis" OR "NASH" OR "fatty liver" OR "hepatic steatosis") AND ("mouse" OR "animal" OR "rat" OR "murine" OR "animal model" OR "murine model" OR "rodent model" OR "experimental model") NOT ("Review"). Both databases were searched using the “Animal” filters[45,46], the results combined, and duplicates eliminated. The search was completed in November 2017.

Study selection and eligibility criteria

Our inclusion criteria were as follows: primary research articles using mice or rats to model NAFLD (to include hepatic steatosis, NASH, and NASH-fibrosis), use of pharmacological intervention with a control (or placebo) group, and that the pharmacological intervention class (e.g. statins) had been used in Phase 2 or 3 trials in humans for treatment of NAFLD/NASH. Studies were excluded if: not modelling NAFLD/NASH; studies in humans or any animal other than mice and rats; reviews, comments, letters, editorials, meta-analyses, ideas; articles not in English (unless there was an available translation); studies not reporting hepatic triglyceride content relative to hepatic protein (e.g. mg/mg or μM/mg); studies using a pharmacological agent class that had not been used in Phase 2/3 studies in humans for NAFLD; and fewer than four independent studies using any single pharmacological agent drug class.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.887919; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Abstracts and titles were screened to identify relevant studies using Rayyan[47]. Potentially relevant studies had their full text extracted and were assessed against inclusion/exclusion criteria independently by two reviewers, with discrepancies settled by discussion with JPM.

Data collection

The variables extracted were as follows: phenotypic characteristics of animal model used (sex, diet [including percentage of fat in diet], rodent age, genetic alterations, background animal strain); drug treatment (dose, drug class, duration, age at intervention), and hepatic triglyceride content. Studies frequently included multiple cohorts or interventional arms, which were defined as use of a different animal model of NAFLD, a different drug, or a different drug dose. Data were extracted for each cohort or interventional arm separately.

Quality assessment

Each paper was assessed in the following 4 areas: use of a protocol, reporting use of randomisation, reporting use of blinding, and a power calculation. These were each given a score of 1, and each paper was assigned an overall "quality score". A post-hoc power calculation was performed for each study using the means of each group and a common SD [48] using the *pwr[49]* package in r. In addition, a 'pre-test' sample size calculation was performed using the overall effect summary from meta-analysis, power = 80% and p-value = .05.

Statistical analysis

Primary outcome was the percentage difference in hepatic triglyceride (TG) content in the interventional group compared to control/placebo.

Random-effects meta-analysis using the Hartung-Knapp-Sidik-Jonkman method was used to calculate mean difference in hepatic TG for each drug class. Heterogeneity within drug classes and across the whole dataset was reported using Cochran's Q, Higgin's & Thompson's I^2, and τ^2. Potential outliers were identified using a Baujat plot[50], where all studies with excess contribution to heterogeneity were excluded as a sensitivity analysis.

Study distribution ('publication') bias was assessed using funnel plot with Egger's test. Given evidence of study distribution bias, Duval & Tweedie's trim-and-fill procedure[51] was performed to estimate the impact of bias on the overall measure.

Mixed-effects meta-regression was performed to assess which baseline variables were associated with heterogeneity in TG content across the whole dataset. Meta-regression was performed for categorical variables (drug class, sex, animal background, NAFLD model design) and continuous variables (percentage of fat in diet, age at intervention, drug dose). For each regression analysis, studies were only included where four or more studies reported each variable. For example, for analysis by animal background, C57BL/6J (used in 110 studies) was included but FVB/N (used in 2 studies) was excluded. Due to high variability and minimal replication, studies using 'Mixed' animal background were excluded from meta-regression analyses. The number of cohorts included in each regression analysis is reported with their results. Multiple variable meta-regression was performed to assess what proportion of between-study heterogeneity could be accounted for by baseline characteristics (using R^2).

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.887919; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Statistical analysis was performed using R 3.6.2 for Mac[52,53] with packages *dmetar[53]*, *meta[54]*, and *metafor[55]*. Graphs were also generated using GraphPad Prism (v8.0 for Mac, GraphPad Software, La Jolla California, USA).

Data availability

The raw dataset used for analysis, including references to individual studies, are available in the Dryad repository: https://doi.org/10.5061/dryad.pzgmsbcgc

R code used for analysis are available in Supplementary Data.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.887919; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

References

1. Howells DW, Sena ES, Macleod MR. Bringing rigour to translational medicine. Nat Rev Neurol. 2014;10: 37–43. doi:10.1038/nrneurol.2013.232

2. Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, et al. A call for transparent reporting to optimize the predictive value of preclinical research. Nature. 2012;490: 187–191. doi:10.1038/nature11556

3. Hackam DG, Redelmeier DA. Translation of research evidence from animals to humans. JAMA. 2006;296: 1731–1732. doi:10.1001/jama.296.14.1731

4. Perel P, Roberts I, Sena E, Wheble P, Briscoe C, Sandercock P, et al. Comparison of treatment effects between animal experiments and clinical trials: systematic review. BMJ. 2007;334: 197. doi:10.1136/bmj.39048.407928.BE

5. Macleod MR, McLean AL, Kyriakopoulou A, Serghiou S, de Wilde A, Sherratt N, et al. Risk of Bias in Reports of In Vivo Research: A Focus for Improvement. PLoS Biol. 2015;13: e1002273. doi:10.1371/journal.pbio.1002273

6. Macleod MR, O'Collins T, Horky LL, Howells DW, Donnan GA. Systematic review and metaanalysis of the efficacy of FK506 in experimental stroke. J Cereb Blood Flow Metab. 2005;25: 713–721. doi:10.1038/sj.jcbfm.9600064

7. Rangarajan A, Weinberg RA. Opinion: Comparative biology of mouse versus human cells: modelling human cancer in mice. Nat Rev Cancer. 2003;3: 952–959. doi:10.1038/nrc1235

8. Mestas J, Hughes CCW. Of mice and not men: differences between mouse and human immunology. J Immunol. 2004;172: 2731–2738. doi:10.4049/jimmunol.172.5.2731

9. van der Worp HB, Howells DW, Sena ES, Porritt MJ, Rewell S, O'Collins V, et al. Can animal models of disease reliably inform human studies? PLoS Med. 2010;7: e1000245. doi:10.1371/journal.pmed.1000245

10. Sena ES, Van Der Worp HB, Bath PMW, Howells DW, Macleod MR. Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. PLoS Biol. 2010;8: e1000344. Available: https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1000344

11. Watzlawick R, Antonic A, Sena ES, Kopp MA, Rind J, Dirnagl U, et al. Outcome heterogeneity and bias in acute experimental spinal cord injury: A meta-analysis. Neurology. 2019;93: e40–e51. doi:10.1212/WNL.0000000000007718

12. Baker M. 1,500 scientists lift the lid on reproducibility. Nature. 2016;533: 452–454.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.887919; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

doi:10.1038/533452a

13. von Herrath M, Pagni PP, Grove K, Christoffersson G, Tang-Christensen M, Karlsen AE, et al. Case Reports of Pre-clinical Replication Studies in Metabolism and Diabetes. Cell Metab. 2019;29: 795–802. doi:10.1016/j.cmet.2019.02.004

14. Prescott MJ, Lidster K. Improving quality of science through better animal welfare: the NC3Rs strategy. Lab Anim . 2017;46: 152–156. doi:10.1038/laban.1217

15. Brenner DA. Of Mice and Men and Nonalcoholic Steatohepatitis. Hepatology . 2018. pp. 2059–2061. doi:10.1002/hep.30186

16. Farrell G, Schattenberg JM, Leclercq I, Yeh MM, Goldin R, Teoh N, et al. Mouse Models of Nonalcoholic Steatohepatitis: Toward Optimization of Their Relevance to Human Nonalcoholic Steatohepatitis. Hepatology. 2019;69: 2241–2257. doi:10.1002/hep.30333

17. Sanyal AJ. Past, present and future perspectives in nonalcoholic fatty liver disease. Nat Rev Gastroenterol Hepatol. 2019;16: 377–386. doi:10.1038/s41575-019-0144-8

18. Younossi Z, Anstee QM, Marietti M, Hardy T, Henry L, Eslam M, et al. Global burden of NAFLD and NASH: trends, predictions, risk factors and prevention. Nat Rev Gastroenterol Hepatol. 2018;15: 11–20. doi:10.1038/nrgastro.2017.109

19. Byrne CD, Targher G. NAFLD: A multisystem disease. J Hepatol. 2015;62: S47–S64. doi:10.1016/j.jhep.2014.12.012

20. Chalasani N, Younossi Z, Lavine JE, Charlton M, Cusi K, Rinella M, et al. The diagnosis and management of nonalcoholic fatty liver disease: Practice guidance from the American Association for the Study of Liver Diseases. Hepatology. 2018;67: 328–357. doi:10.1002/hep.29367

21. Ratziu V, Sanyal AJ, Loomba R, Rinella M, Harrison S, Anstee QM, et al. REGENERATE: Design of a pivotal, randomised, phase 3 study evaluating the safety and efficacy of obeticholic acid in patients with fibrosis due to nonalcoholic steatohepatitis. Contemp Clin Trials. 2019;84: 105803. doi:10.1016/j.cct.2019.06.017

22. Sanyal AJ, Abdelmalek MF, Suzuki A, Cummings OW, Chojkier M. No significant effects of ethyl-eicosapentanoic acid on histologic features of nonalcoholic steatohepatitis in a phase 2 triaL. Gastroenterology. 2014;147: 377–384.e1. doi:10.1053/j.gastro.2014.04.046

23. Gilead Sciences, Inc. Gilead Announces Topline Data From Phase 3 STELLAR-3 Study of Selonsertib in Bridging Fibrosis (F3) Due to Nonalcoholic Steatohepatitis (NASH). In: Gilead.com [Internet]. 25 Apr 2019 [cited 22 Dec 2019]. Available:

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.887919; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

https://www.gilead.com/news-and-press/press-room/press-releases/2019/4/gilead-announces-topline-data-from-phase-3-stellar3-study-of-selonsertib-in-bridging-fibrosis-f3-due-to-nonalcoholic-steatohepatitis-nash

24. Budas G, Karnik S, Jonnson T, Shafizadeh T, Watkins S, Breckenridge D, et al. Reduction of Liver Steatosis and Fibrosis with an Ask1 Inhibitor in a Murine Model of Nash is Accompanied by Improvements in Cholesterol, Bile Acid and Lipid Metabolism. J Hepatol. 2016;64: S170. doi:10.1016/S0168-8278(16)01686-X

25. Anstee QM, Goldin RD. Mouse models in non-alcoholic fatty liver disease and steatohepatitis research. Int J Exp Pathol. 2006;87: 1–16. doi:10.1111/j.0959-9673.2006.00465.x

26. Friedman SL, Neuschwander-Tetri BA, Rinella M, Sanyal AJ. Mechanisms of NAFLD development and therapeutic strategies. Nat Med. 2018;24: 908–922. doi:10.1038/s41591-018-0104-9

27. Armstrong MJ, Gaunt P, Aithal GP, Barton D, Hull D, Parker R, et al. Liraglutide safety and efficacy in patients with non-alcoholic steatohepatitis (LEAN): A multicentre, double-blind, randomised, placebo-controlled phase 2 study. The Lancet. 2015. doi:10.1016/S0140-6736(15)00803-X

28. Sanyal AJ, Chalasani N, Kowdley KV, McCullough A, Diehl AM, Bass NM, et al. Pioglitazone, vitamin E, or placebo for nonalcoholic steatohepatitis. N Engl J Med. 2010;362: 1675–1685. doi:10.1056/NEJMoa0907929

29. Kajikawa S, Harada T, Kawashima A, Imada K, Mizuguchi K. Highly purified eicosapentaenoic acid prevents the progression of hepatic steatosis by repressing monounsaturated fatty acid synthesis in high-fat/high-sucrose diet-fed mice. Prostaglandins, Leukotrienes and Essential Fatty Acids. 2009. pp. 229–238. doi:10.1016/j.plefa.2009.02.004

30. Jia C, Huan Y, Liu S, Hou S, Sun S, Li C, et al. Effect of Chronic Pioglitazone Treatment on Hepatic Gene Expression Profile in Obese C57BL/6J Mice. Int J Mol Sci. 2015;16: 12213–12229. doi:10.3390/ijms160612213

31. Castro RE, Diehl AM. Towards a definite mouse model of NAFLD. Journal of hepatology. 2018. pp. 272–274. doi:10.1016/j.jhep.2018.05.002

32. Rogers AB. Stress of Strains: Inbred Mice in Liver Research. Gene Expr. 2018;19: 61–67. doi:10.3727/105221618X15337408678723

33. Hui ST, Parks BW, Org E, Norheim F, Che N, Pan C, et al. The genetic architecture of NAFLD among inbred strains of mice. Elife. 2015;4: e05607. doi:10.7554/eLife.05607

34. Martin MD, Danahy DB, Hartwig SM, Harty JT, Badovinac VP. Revealing the

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.887919; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Complexity in CD8 T Cell Responses to Infection in Inbred C57B/6 versus Outbred Swiss Mice. Front Immunol. 2017;8: 1527. doi:10.3389/fimmu.2017.01527

35. Homanics GE, Quinlan JJ, Firestone LL. Pharmacologic and behavioral responses of inbred C57BL/6J and strain 129/SvJ mouse lines. Pharmacol Biochem Behav. 1999;63: 21–26. doi:10.1016/s0091-3057(98)00232-9

36. Liu X, Gershenfeld HK. Genetic differences in the tail-suspension test and its relationship to imipramine response among 11 inbred strains of mice. Biol Psychiatry. 2001;49: 575–581. doi:10.1016/s0006-3223(00)01028-3

37. Liu C-Y, Chang C-W, Lee H-C, Chen Y-J, Tsai T-H, Chiau J-SC, et al. Metabolic Damage Presents Differently in Young and Early-Aged C57BL/6 Mice Fed a High-Fat Diet. International Journal of Gerontology. 2016;10: 105–111. doi:10.1016/j.ijge.2015.10.004

38. Kirkland JL, Stout MB, Sierra F. Resilience in Aging Mice. J Gerontol A Biol Sci Med Sci. 2016;71: 1407–1414. doi:10.1093/gerona/glw086

39. Tsilidis KK, Panagiotou OA, Sena ES, Aretouli E, Evangelou E, Howells DW, et al. Evaluation of excess significance bias in animal studies of neurological diseases. PLoS Biol. 2013;11: e1001609. doi:10.1371/journal.pbio.1001609

40. Bahor Z, Liao J, Macleod MR, Bannach-Brown A, McCann SK, Wever KE, et al. Risk of bias reporting in the recent animal focal cerebral ischaemia literature. Clin Sci . 2017;131: 2525–2532. doi:10.1042/CS20160722

41. Denayer T, Stöhr T, Van Roy M. Animal models in translational medicine: Validation and prediction. New Horizons in Translational Medicine. 2014;2: 5–11. doi:10.1016/j.nhtm.2014.08.001

42. Kaden-Volynets V, Basic M, Neumann U, Pretz D, Rings A, Bleich A, et al. Lack of liver steatosis in germ-free mice following hypercaloric diets. Eur J Nutr. 2019;58: 1933–1945. doi:10.1007/s00394-018-1748-4

43. Mazagova M, Wang L, Anfora AT, Wissmueller M, Lesley SA, Miyamoto Y, et al. Commensal microbiota is hepatoprotective and prevents liver fibrosis in mice. FASEB J. 2015;29: 1043–1055. doi:10.1096/fj.14-259515

44. Balmer ML, Slack E, de Gottardi A, Lawson MAE, Hapfelmeier S, Miele L, et al. The liver may act as a firewall mediating mutualism between the host and its gut commensal microbiota. Sci Transl Med. 2014;6: 237ra66. doi:10.1126/scitranslmed.3008618

45. Hooijmans CR, Tillema A, Leenaars M, Ritskes-Hoitinga M. Enhancing search efficiency by means of a search filter for finding all studies on animal experimentation in PubMed. Lab Anim. 2010;44: 170–175.

bioRxiv preprint doi: https://doi.org/10.1101/2019.12.31.887919; this version posted February 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

doi:10.1258/la.2010.009117

46. de Vries RBM, Hooijmans CR, Tillema A, Leenaars M, Ritskes-Hoitinga M. Updated version of the Embase search filter for animal studies. Laboratory animals. 2014. p. 88. doi:10.1177/0023677213494374

47. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev. 2016;5: 210. doi:10.1186/s13643-016-0384-4

48. Cohen J. Statistical power analysis for the behavioral sciences. Routledge; 1988. doi:10.4324/9780203771587

49. Champely S. pwr: Basic functions for power analysis. R package version 1.2-2. Available: https://CRAN.R-project.org/package=pwr

50. Baujat B, Mahé C, Pignon J-P, Hill C. A graphical method for exploring heterogeneity in meta-analyses: application to a meta-analysis of 65 trials. Stat Med. 2002;21: 2641–2652. doi:10.1002/sim.1221

51. Duval S, Tweedie R. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics. 2000;56: 455–463. doi:10.1111/j.0006-341x.2000.00455.x

52. Team RC. A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. http://www. R-project. org; 2014.

53. Harrer M, Cuijpers P, Furukawa TA, Ebert DD. Doing Meta-Analysis in R: A Hands-on Guide. PROTECT Lab Erlangen. 2019.

54. Schwarzer G, Others. meta: An R package for meta-analysis. R news. 2007;7: 40–45. Available: https://cran.rstudio.org/doc/Rnews/Rnews_2007-3.pdf#page=40

55. Viechtbauer W. Conducting meta-analyses in R with the metafor package. J Stat Softw. 2010;36: 1–48. Available: https://cran.r-project.org/web/packages/metafor/vignettes/metafor.pdf