scientific reports

OPEN

Performance of preclinical models in predicting drug-induced liver injury in humans: a systematic review

Hubert Dirven [1], Gunn E. Vist[2], Sricharan Bandhakavi[3], Jyotsna Mehta[4], Seneca E. Fitch[5], Pandora Pound[6], Rebecca Ram[6], Breanne Kincaid [7], Cathalijn H. C. Leenaars [8], Minjun Chen[9], Robert A. Wright [10] & Katya Tsaioun [11]✉

Drug-induced liver injury (DILI) causes one in three market withdrawals due to adverse drug reactions, causing preventable human suffering and massive financial loss. We applied evidence-based methods to investigate the role of preclinical studies in predicting human DILI using two anti-diabetic drugs from the same class, but with different toxicological profiles: troglitazone (withdrawn from US market due to DILI) and rosiglitazone (remains on US market). Evidence Stream 1: A systematic literature review of in vivo studies on rosiglitazone or troglitazone was conducted (PROSPERO registration CRD42018112353). Evidence Stream 2: in vitro data on troglitazone and rosiglitazone were retrieved from the US EPA ToxCast database. Evidence Stream 3: troglitazone- and rosiglitazone-related DILI cases were retrieved from WHO Vigibase. All three evidence stream analyses were conducted according to evidence-based methodologies and performed according to pre-registered protocols. Evidence Stream 1: 9288 references were identified, with 42 studies included in analysis. No reported biomarker for either drug indicated a strong hazard signal in either preclinical animal or human studies. All included studies had substantial limitations, resulting in "low" or "very low" certainty in findings. Evidence Stream 2: Troglitazone was active in twice as many in vitro assays (129) as rosiglitazone (60), indicating a strong signal for more off-target effects. Evidence Stream 3: We observed a fivefold difference in both all adverse events and liver-related adverse events reported, and an eightfold difference in fatalities for troglitazone, compared to rosiglitazone. In summary, published animal and human trials failed to predict troglitazone's potential to cause severe liver injury in a wider patient population, while in vitro data showed marked differences in the two drugs' off-target activities, offering a new paradigm for reducing drug attrition in late development and in the market. This investigation concludes that death and disability due to adverse drug reactions may be prevented if mechanistic information is deployed at early stages of drug development by pharmaceutical companies and is considered by regulators as a part of regulatory submissions.

Medicines save millions of lives and are considered a cost-effective intervention, effectively combating infections and making conditions previously considered incurable now manageable. However, drugs can also cause dangerous and fatal reactions in humans, both in clinical trials and after market approval. In the US, it is estimated that 2 million serious adverse drug reactions (ADRs) occur every year in hospitalized patients, with 100,000 people dying annually[1]. In the UK, it is estimated that ADRs kill more than 10,000 annually[2] and account for 6.5% of hospital admissions[3]. ADRs also result in significant costs to pharmaceutical companies when drugs have to be

[1]Department of Environmental Health, Norwegian Institute of Public Health, Oslo, Norway. [2]Division for Health Services, Norwegian Institute of Public Health, Oslo, Norway. [3]Geneia LLC, Cambridge, MA, USA. [4]Keva Health, Lexington, MA, USA. [5]ToxStrategies, Inc, Katy, TX, USA. [6]Safer Medicines Trust, Kingsbridge, UK. [7]Department of Environmental Health and Engineering, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [8]Institute for Laboratory Animal Sciences, Hannover Medical School, Hannover, Germany. [9]Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, US Food and Drug Administration, Little Rock, AK, USA. [10]Basic Science Informationist, Welch Medical Library, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [11]Evidence-Based Toxicology Collaboration, Johns Hopkins Bloomberg School of Public Health, 615 N Wolfe St, Baltimore, MD 21205, USA. ✉email: ktsaiou1@jhu.edu

1

withdrawn[4], create human suffering, and place huge burdens on health systems[5] and the economy. While some ADRs may occur as a result of inappropriate use or prescribing errors[5], a major question remains why drugs that have met the preclinical and clinical testing required to secure regulatory approval go on to cause adverse reactions in humans. Regulatory agencies require a standard battery of tests based on International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) guidelines to ensure the safety and efficacy of new drugs before they are tested in humans. This regulatory battery relies on preclinical animal testing in rodents (typically rats) and non-rodents (typically dogs or non-human primates (NHPs)). This process is lengthy, costly[6,7], and risky, considering that most new drugs in development fail to gain approval[8,9]. While several factors contribute to the high attrition rates in new drug development, including the difficulty of detecting rare events in small clinical trial populations, it is widely agreed that a predominant reason is the failure of preclinical animal models (as well as some long-established in vitro assays using mono-layer transformed cell cultures[10,11]) to accurately predict clinical efficacy[12–15] and safety[16,17]. There is increasing evidence that the current system of drug development needs to be modernized[18] and that we need to use tests that are more predictive of human outcomes[12].

In the last few decades, due to the advent of molecular biology techniques and high-throughput screening, a number of tests based on human biology have been developed and commercialised. These tests employ a variety of approaches, including stem cells, -omics-based technologies, organoids, organs-on-chips, and computational (in silico) approaches. These new tests, often referred to as new approach methodologies (NAMs), can be used to study the mechanisms of toxicity of chemicals and identify endpoints of concern, thus, allowing for more targeted follow-up of promising chemical or drug candidates, without subjecting every candidate to the recommended ICH and/or OECD guideline tests. There is now enormous optimism about NAMs[12,18,19] and accumulating evidence to support their use in regulatory contexts across various economic sectors[20]. The US Food and Drug Administration's (FDA) Center for Drug Evaluation and Research (CDER) encourages communication with stakeholders regarding NAMs and is committed to exploring the potential for NAMs to improve regulatory efficiency and expedite drug development[21]. A key government initiative has been the Toxicity Forecaster Programme (ToxCast), launched by the US Environmental Protection Agency (EPA) in 2007 to investigate the safety of industrial chemicals, cosmetics, pesticides and approved drugs using in vitro mechanistic information[22]. ToxCast uses high-throughput screening technologies based on human biology. The cells or proteins in these assays are exposed to chemicals and assessed for changes in biological activity that may suggest undesirable effects in humans. Nearly ten thousand chemicals have been screened to date against over a thousand molecular targets, which makes ToxCast the largest public in vitro database in the world.

Drug-induced liver injury (DILI) is the most frequent cause of acute liver failure in the Western world, accounting for more than half of all cases. DILI is also responsible for 3–5% of hospital admissions for jaundice[23]. Its incidence is estimated to be 14–19 cases per 100,000 persons, with jaundice occurring in 30% of cases[23]. Drug safety has become the bottleneck of drug development, with hepatotoxicity accounting for one in every 4.5 drug failures in clinical trials and one in every three market withdrawals caused by ADRs[24]. While there are standard clinical diagnostic markers of DILI, animal studies have only a limited ability to predict hepatic drug safety using these markers[25]. The aim of this study is to take an evidence-based approach[26] to investigating how well ToxCast in vitro tests compare with preclinical animal tests in predicting liver-related ADRs in humans, with human pharmacovigilance data used as the true indicator of DILI incidence in the population. The current investigation is conducted according to a pre-registered protocol[27] which outlines our intent to query ten drugs selected according to the presence or absence of documented DILI in human subjects. This is the first publication based on this protocol. Here we report data on two of the ten drugs, troglitazone and rosiglitazone maleate (henceforth referred to as rosiglitazone). This pair of anti-diabetic drugs come from the same class of thiazolidinediones but have differing effects on the human liver. Troglitazone was approved in the US in 1997 but withdrawn from the US market in 2000 after reports of deaths and severe liver failure requiring transplantation. Rosiglitazone was approved in the US in 1999 and remains on the US market[28,29]. We selected this pair of drugs because of their distinct liver safety profiles: their regulatory status is "withdrawn" for troglitazone and "on the market" for rosiglitazone, while their DILI risk classification (based on the US FDA Liver Toxicity Knowledge Base) is "most DILI concern" for troglitazone and "less DILI concern" for rosiglitazone[30].
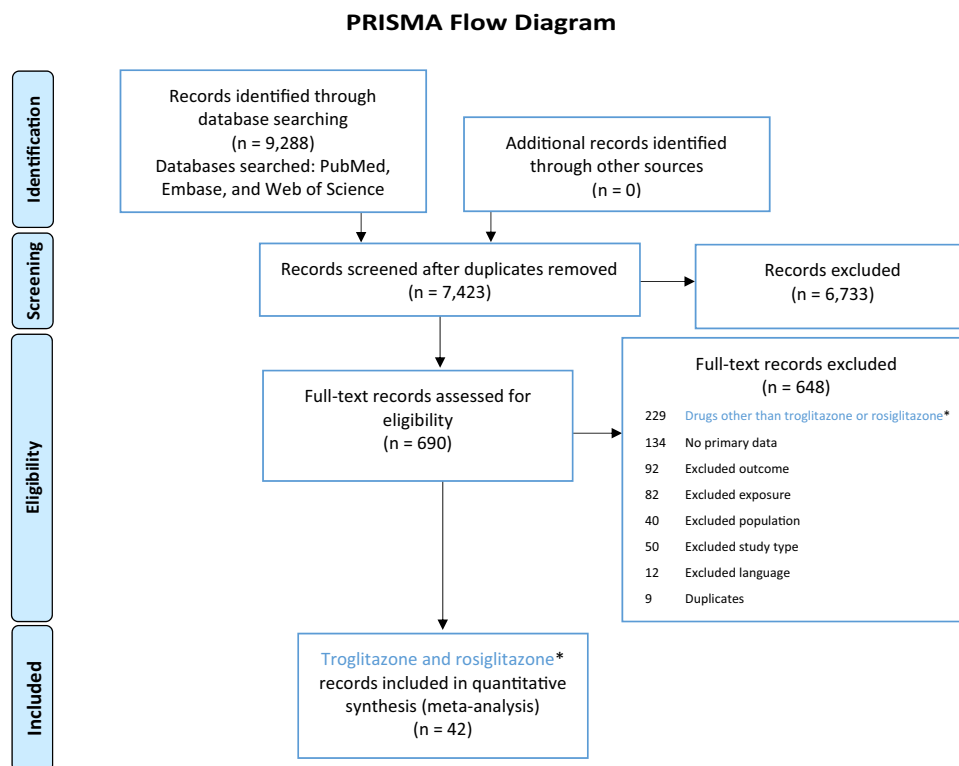
Results

Evidence stream 1: systematic review of in vivo *studies*.
The literature searches identified 9288 references. After screening the titles/abstracts for relevance, we reviewed the remaining 690 references in full text. Two hundred and seventy-one publications were retained for data extraction, 42 of which were studies of troglitazone or rosiglitazone (22 on troglitazone and 22 on rosiglitazone, with 2 studies evaluating both compounds). The other 229 publications were studies of eight other drugs that will be analysed separately (see systematic review protocol) (Fig. 1).

The included studies are presented in Table 1a (troglitazone), b (rosiglitazone) and S2. Most of the studies of troglitazone were published after drug withdrawal in 2000, probably to study the mechanisms of toxicity involved.

Risk of bias for the included studies. A summary of our risk of bias (RoB) assessments for the included studies is presented in Fig. 2a (animal studies) and b (human studies).

Animal studies. Eight of the 11 RoB questions in the OHAT tool were applicable to the animal studies (Fig. 2a). Overall, many studies failed to report the information needed for reviewers to assess potential bias. In terms of selection, exclusion, and selective reporting bias, most studies had low or definitely low RoB, with a few excep-

PRISMA Flow Diagram

* Drugs other than troglitazone and rosiglitazone will be analyzed in forthcoming studies.

Figure 1. PRISMA flow diagram depicting study inclusion and exclusion justification. *Results on drugs other than troglitazone and rosiglitazone will be published in forthcoming manuscripts.

tions. However, it is important to note that a large number of studies had at least two bias domains where there was a high RoB, including performance bias and detection bias.

Human studies. The human studies included 8 randomised controlled trials (RCTs), 4 cohort studies (Co), 3 case series (CaS), and 5 cohort/case series studies (Co/CaS). All the human studies had some methodological challenges (Fig. 2b) that impacted confidence in the effect estimates and conclusions, with lack of adequate reporting on randomisation and blinding (RCTs), selection bias, confounding and outcome assessment (Co), and selective outcome reporting (all study designs). Given the low number of studies, it is hard to draw any conclusions about RoB for the CaS and Co/CaS studies. The studies on rosiglitazone appeared to be either better reported or to have a lower RoB than the troglitazone studies.

Meta-analysis of the effect of troglitazone and rosiglitazone on liver function in animals and humans. The included studies had a variety of study designs, dosing regimens, and liver endpoints reported (Table 1). Among the included endpoints, ALT elevation was the most frequently reported outcome, closely followed by AST. ALP, total bilirubin, liver weight, and histopathology were infrequently reported. To summarize troglitazone's and rosiglitazone's effects on the liver, we conducted a meta-analysis of the reported outcomes and present the results from all species on each of the five main liver outcomes in collated forest plots (S3), with the representative forest plot for ALT (Fig. 3) and summary of all five liver outcomes (Table 2) presented below. In most studies, wide confidence intervals due to the small number of participants (animal studies) or events (human trials) restrict our ability to draw definitive conclusions about the predictive ability of any specific endpoint or animal model. Given that for most liver injury markers there were a limited number of studies on each species, we caution against over-interpretation of these results.

In the studies where mice were given troglitazone there appeared to be an increase in ALT (4 studies, total 118 mice), AST (2 studies, total 78 mice), ALP (1 study, 19 mice), liver weight (1 study, 19 mice) and inconclusive changes in total bilirubin (2 studies, total 38 mice). We found fewer rosiglitazone studies in mice. In these studies, inconclusive results are reported for ALT (2 studies, total 28 mice), AST (2 studies, total 28 mice) and liver weight (3 studies, total 51 mice). Rosiglitazone appeared to reduce ALP (1 study, 12 mice) and total bilirubin (1 study, 12 mice).

In the studies where rats were given troglitazone, results were inconclusive for ALT (4 studies, total 305 rats), AST (4 studies, total 305 rats), total bilirubin (2 studies, total 46 rats) and liver weight (2 studies, total 48 rats). However, for ALP there appears to be a dose–response increase after troglitazone compared to control (2 studies, total 38 rats). We found fewer studies on rosiglitazone in rats. In these studies, inconclusive results are

References	# animals/participants	Age	Strain/species	Sex	Drug dose	Route	Exposure time	ALT	AST	ALP	Total bilirubin	Liver weight, relative	Liver weight, absolute	Histo-pathology
(a) Troglitazone studies														
Studies in mice (n = 5)														
Bedoucha 2001	3 mice/group	13 weeks	C57BL/6J	M	0, 100 mg/kg/day	p.o. (gavage)	28 days							X
Fujimoto 2009	10 mice/group 9 mice/group	8 weeks 35 weeks	Wild type Sod2+/+	M	0, 300 mg/kg/day	p.o. (gavage)	28 days	X	X		X	X	X	X
Jia et al. 2019	4–6 mice/group	6 weeks	BALB/c	F	0, 30, 100, 300 mg/kg	ip	1 day/single dose	X	X	X				X
Mak et al. 2018	3 mice/group 5 mice histopath control	8–10 weeks	C57BL/6	F	0.2% w/w chow ad libitum	Diet	5 weeks	X	X					
Ong et al. 2007	5 mice/group	16–21 weeks	Wild type Sod2+/+	M/F	0, 30 mg/kg/day	ip	28 days	X						
Studies in rats (n = 9)														
Boitier 2011	5 rats/group	8–10 weeks	Wistar	M	0, 200, 1500 mg/kg/day	p.o. (gavage)	14 days							X
Cepa et al. 2018	10 rats/sex/group	88 to 94 days old	Sprague Dawley	M/F	0, 25, 50 mg/kg	i.v	1 day/single dose	X	X					X
Cheng et al. 2017	4–5 rats/group	Not reported	Sprague Dawley	M	0, 200 mg/kg/day	Oral gavage	7 days/daily dose	X	X	X				
Hermann 2002	60 rats per sex per group (480 rats)	7 weeks	Wistar	F / M	0, 25, 50, 200 mg/kg/day / 0, 100, 400, 800 mg/kg/day	p.o. (gavage)	104 weeks							X
Jia 2000	6–10 rats/group	12 weeks 28 weeks	Long-Evans Tokushima Otsuka (LETO)	M	0, 200 mg/100 g food / 0, 200 mg/100 g food	p.o.(diet)	60 weeks (until 72 wks old) / 44 weeks (until 72 wks old)	X	X			X	X	X
Kakiuchi-Kiyota 2011	10 rats/group	7 weeks	Wistar	F	0, 50, 200 mg/kg/day	p.o. (gavage)	4 weeks					X	X	
Kostrubsky 2001	4 rats/group	Not reported	Wistar	M	0, 200 mg/kg	p.o. (gavage)	2 h 36 h				X			
Li 2002	5 rats/group	10 weeks	Wistar/ST	M	0,100, 500 mg/kg/day	p.o. (diet)	3 weeks		X	X				
Watanabe 2000	5 rats/sex/group	7 weeks	Wistar	M/F	0, 100, 400 mg/kg	p.o. (gavage)	94 days	X	X	X	X			
Study in NHP (non-human primates; n = 1)														
Rothwell 2002	4 NHP/sex/group	2 to 6 years	Cynomolgus Macaques	M/F	0, 300, 600, 1200 mg/kg	p.o. (gavage)	52 weeks	X	X	X	X	X	X	X
Studies in humans (n = 7)														
Arioglu 2000	20 patients	6 to 65 years	–	M/F	200 to 600 mg/day	p.o	6 months	X						X
Azziz 2001	305 patients	30.1 +/- 6.0 years 28.9 +/- 5.4 years 29.2 +/- 5.8 years 29.0 ± 5.2 years (mean +/- SD)	–	F	Placebo 150 mg/day 300 mg/day 600 mg/day	p.o	44 weeks	X	X					
Björnsson 2006*	148 cases	Not reported	–	-	Not reported	p.o	Not reported							
Knowler 2005	582+585 patients	51 years (mean)	–	M/F	Placebo, 400 mg/day	p.o	0.9 year (mean) (range 0.5 to 1.5 years)	X	X					
Mavandadi 1999	21 patients	31–64 years	–	M/F	200 to 600 mg/day	p.o	6 months	X	X					
St.Peter 2001	291 patients	53.5 ± 12.8 (mean +/- SD)	–	M/F	Not reported	p.o	412.7 ± 255.6 days (mean +/- SD)	X	X					
Yale 2001	200 patients	59 years (mean)	–	M/F	400 mg/day	p.o	6–12 months	X						
Continued														

(b) Rosiglitazone studies

References	# animals/participants	Age	Strain/species	Sex	Drug dose	Route	Exposure time	ALT	AST	ALP	Total bilirubin	Liver weight, relative	Liver weight, absolute	Histo-pathology
Studies in mice (n = 5)														
Anandharajan 2009	8 mice/group	7–9 weeks	C57BL/6 J	M	0, 10 mg/kg/day	p.o. (gavage)	10 days	X	X				X	
Bedoucha 2001	3 mice/group	13 weeks	C57BL/6 J	M	0, 2.5 mg/kg/day	p.o. (gavage)	28 days		X					X
Jia et al. 2019	4–6 mice/group	6 weeks	BALB/c	F	0, 30, 100, 300 mg/kg	i.p	1 day/single dose	X	X					X
Otake 2011	5 mice/group	7 weeks	ICR	M	0, 30,100 mg/kg/day	p.o. (gavage)	14 days						X	
Zhang et al. 2019	6 mice/group	8 weeks	C57BL/6 J	M	0, 30 mg/kg/day	Gastric intubation	6 days	X	X	X	X	X	X	X
Studies in rats (n = 7)														
Anandharajan 2009	8 rats/group	6 weeks	Wistar	F	0, 40 mg/kg/day	p.o. (gavage)	28 days	X	X				X	
Anwar et al. 2015	6 rats/group	7–8 weeks	Sprague–Dawley	M/F	0, 40 mg/kg/day	Diet	16 weeks	X	X	X	X			X
Dadarkar 2011	3 rats/group	8 weeks	Sprague Dawley	F	0, 300 mg/kg	i.p	1 day/single dose	X	X		X		X	X
Egerod 2009	10 rats, 5 rats, 5 rats	6 weeks	Sprague Dawley	M	0 mg/kg/day, 8 mg/kg/day, 20 mg/kg/day	p.o. (gavage)	7 days					X	X	
Meghani 2012	6 rats/group	Not reported	Wistar	M	0, 80 mg/kg/day	p.o.(gavage)	14 days	X	X	X				X
Schafer et al. 2012	6 rats	10 weeks	ZDF	M	3 mg/kg	p.o. (gavage)	8 weeks	x	X			X		
Spicker et al. 2007	9 rats, 5 rats	5–6 weeks	Sprague Dawley	M	0, 1000 mg/kg/day	p.o. (gavage)	1 day/single dose	X	X					
Studies in humans (n = 11)														
Aramwit 2009	13 patients	54.17 ± 11.42 years (Mean +/- SD. Range 35–85 years)	–	M/F	4 mg/day	p.o	12 weeks	X	X					
Beysen 2008	6 patients	56 +/- 6 years (Mean +/- SD)	–	M/F	16 mg/day for 4 weeks followed by 8 mg/day for 16 weeks	p.o	20 weeks	X	X					
Chalasani 2005	210 patients (Cohort 1)	53 ± 11 years			4.4 +/- 1.8 mg/day									
	628 patients (Cohort 2)	55 +/- 11 years (Mean +/- SD)		M/F	4.4 +/- 2 mg/day (mean +/- SD)	p.o	12 months	X	X		X			
Chiang 2007	78 patients	56.8 +/- 7.2 years (Mean +/- SD)	–	M/F	2 to 8 mg/day	p.o	3–12 months	X	X	X				
Dereli 2005	20 patients	29.4 ± 1.7 years			2 mg/day	p.o	8 months							
	20 patients	31.4 ± 0.9 years (Mean +/- SD)	–	F	4 mg/day			X	X					
Gegick 2001*	77 patients	59 ± 10.4	–	M/F	4 or 8 mg/day	p.o	3.2 months							
Gegick 2004*	49 patients	59.5 ± 10.9 (Mean +/- SD)	–	M/F	4 or 8 mg/day	p.o	12.6 months							
Hussein 2004	96 patients	64.6 +/- 10.3 Mean +/- SD. Range 41–82 years	–	M/F	4 or 8 mg/day	p.o.	2 months		X					

Continued

References	# animals/participants	Age	Strain/species	Sex	Drug dose	Route	Exposure time	ALT	AST	ALP	Total bilirubin	Liver weight, relative	Liver weight, absolute	Histo-pathology
Nolan 2000	93 patients	62.3 +/- 9.5 (Mean =/- SD. Range 39–80 years)	–	M/F	Placebo,	p.o	8 weeks	X						
	95 patients	62.9 ± 10.1 (Mean =/- SD. Range 43–83 years)			4 mg/day									
	90 patients	62.6 ± 9.9 years (Mean +/- SD. Range 40–80 years)			8 mg/day									
	91 patients	63.4 +/- 9.1 (Mean +/- SD. Range 41–79 years)			12 mg/day									
Phillips 2001	173 patients	57.7 ± 9.2 years	–	M/F	Placebo,	p.o	26 weeks	X						
	181 patients	57.5 ± 9.9 years			4 mg o.d									
	186 patients	56.8 ± 9.4 years			2 mg b.i.d									
	181 patients	58.9 ± 9.9 years			8 mg o.d									
	187 patients	56.5 ± 9.7 years (Mean +/- SD)			4 mg b.i.d									
Wong 2005*	52 patients	62.92 ± 7.30 rosiglitazone and 61.58 ± 9.7 control	–	M/F	4 mg/day	p.o	24 weeks							

Table 1. Description of studies and outcomes reported for (a) troglitazone and (b) rosiglitazone. *Narrative review includes no numerical data.

reported for ALT and AST (6 studies, total 68 rats), ALP (3 studies, total 32 rats) and total bilirubin (2 studies, total 18 rats). However, there appeared to be an increase in liver weight after rosiglitazone compared to control (2 studies, total 22 rats).

Also included was a controlled one year-long study on male and female NHPs. The 24 NHPs which were given troglitazone (300, 600 or 1200 mg/kg), had lower levels of ALT, AST, ALP and total bilirubin than the control group. However, liver weights were reported to be significantly higher in the NHPs given troglitazone compared to controls.

We found relatively few published human trials on both drugs. The only study, which measured ALT in 19 patients exposed to troglitazone, found no significant difference compared to placebo at 12 and 26 weeks. There were 5 published human trials of rosiglitazone, which together indicate no significant difference in ALT after drug exposure. There was no significant change in AST in patients in 8 studies with rosiglitazone ($p = 0.06$), except for one study which found a transient increase in AST at earlier time points[31]. We found no studies which reported ALP or bilirubin levels after troglitazone exposure. In the studies where patients were given rosiglitazone there were inconclusive results for ALP and AST (5 studies with 1548 patients) and for ALP (1 study with 156 patients). However, the one study that reported bilirubin levels found that bilirubin levels increased in patients on rosiglitazone compared to placebo (1 study with 156 patients).

Histopathology findings were reported in all studies in narrative form, making it impossible to perform a quantitative analysis. These data are summarized in S4. The sole published NHP study[32] reported a dose–response increase in liver weight in NHPs after troglitazone administration, a two-fold relative liver weight increase compared to controls at the highest dose in both male and female animals. Studies without a placebo control or numerical data for the endpoints were not included in the meta-analysis and are summarized in S5.

GRADE assessment of confidence in evidence. We graded our confidence in the effect estimates (S6a, S6b) as "low" and "very low" for the outcomes measured (ALT, AST, ALP, bilirubin and liver weight) in all species (mice, rats, NHPs and humans) for both troglitazone and rosiglitazone. We have downgraded for RoB in both animal and human studies and in both RCTs and observational studies. Because the focus of this investigation was on drug safety for human patients, the animal studies were downgraded for indirectness. We have downgraded most of the outcomes for imprecision due to wide confidence intervals (CI) and the small number of subjects. We have also downgraded for publication bias, which we strongly suspect since regulatory studies do not appear to be in the public domain.

Evidence stream 2: analysis of in vitro *ToxCast data.* As an initial step in evaluating the in vitro data, we compared the pharmacological activity of troglitazone and rosiglitazone. For this comparison, we included all assays in which both drugs were tested and there was a response (i.e., AC_{50} was not reported as NA, indicating not available, or 1,000,000, indicating no activity within concentrations tested). This workflow resulted in a total set of 437 assays performed on both drugs (Fig. 4A). Troglitazone was active in almost twice as many assays (129) as rosiglitazone (69). The Venn diagram (Fig. 4B) further dissects the data in this set in which both drugs were active (51 tests), as well as tests in which only troglitazone or rosiglitazone were active. Troglitazone not only was active in almost twice as many tests in the ToxCast database, but also uniquely activated 78 tests, compared to detected activity in only 10 tests for rosiglitazone. The full set of "positive" tests, along with identifying information and their AC_{50} values for rosiglitazone vs. troglitazone are available in S7.

Next, we tried to better understand the underlying biological processes represented by the positive assays for each drug using the ToxCast assignments of assays to biological processes. Troglitazone activated more assays across all biological processes represented by the common set of 437 assays for both drugs (Fig. 5). Unsurprisingly, given the desired target of both drugs on PPAR-γ, a nuclear transcription factor, the largest effects for both drugs were seen in assays with transcriptional factor or gene expression regulation targets. However, across these three broad biological processes, troglitazone consistently activated more endpoints, indicating more potential for off-target activity, leading to more potentially undesirable side effects (Fig. 5).

These results indicate that the in vitro/cellular assays-based pharmacological activity of troglitazone is higher across all measured biological processes relative to rosiglitazone. However, the AC_{50} data in in vitro assays do not account for human exposure levels and thus may not be relevant to the in vivo scenario, which usually needs to be addressed in the form of an IVIVE (In Vitro to In Vivo Extrapolation) model. For this reason, we introduced the NAS score and used it (Fig. 6) to put the in vitro results in the context of the human exposure and to stratify which of the assays/molecular targets might be activated more with each drug. Thus, the NAS metric allowed us to stratify the 437 tests based on their "activation potential" in patients administered a clinically relevant dose of either rosiglitazone or troglitazone.

As indicated, all positive tests fall into three distinct clusters—those activated by rosiglitazone alone (top cluster), those activated by troglitazone alone (middle cluster), and those activated by both drugs (bottom cluster). The strongest activated assays for both drugs measure the perturbation of PPAR-γ—the desired drug target. However, troglitazone activates several other assays (either uniquely or to a greater extent relative to rosiglitazone based on the relative NAS score). Notably, molecular targets of several of these assays are implicated in liver injury/repair pathways[33–40] such as MMP1 (Matrix Metalloproteinase-1), NR3C1 (glucocorticoid receptor), NR1H3 (Liver X receptor alpha), NR1H4 (Bile Acid Receptor), TIMP1 (Tissue inhibitor of Metalloproteinase-1), ICAM1 (intracellular adhesion molecule 1), CXCL9 (T-cell chemoattractant/Chemokine ligand 9), IL8 (Interleukin/Chemokine ligand 8), CD38 (multifunctional ectoenzyme CD38 molecule), VDR (Vitamin D receptor), NRF2 (antioxidant nuclear transcription factor 2), and HLA-DR (MHC Class II cell surface receptor). S8 provides a heat map of assay targets that illustrates their (potential) relative activation in tests treated with troglitazone vs. rosiglitazone. In summary, the analyses of in vitro assays from the ToxCast database showed a clear distinction

Figure 2. Risk of Bias assessment of (**a**) animal and (**b**) human studies according to criteria defined in the OHAT Risk of Bial Tool for Human and Animal Studies. + +: There is direct evidence of low risk of bias practices; +: There is indirect evidence of low risk of bias practices OR it is deemed that deviations from the low risk of bias practices for these criteria during the study would not appreciably bias results;—or NR: There is indirect evidence of high risk of bias practices OR there is insufficient information (e.g., not reported or 'NR') provided about relevant risk of bias; There is direct evidence of high risk of bias practices; NA: Question not relevant for study type. Please note that the study of Bedoucha 2001 on mice included both troglitazone and rosiglitazone, and that the study of Anandharajan 2009 on rosiglitazone included both mice and rats.

between rosiglitazone and troglitazone in terms of both the number and type of off-target biological activities, with troglitazone being active in almost twice as many tests as rosiglitazone. Moreover, several tests measuring mediators of liver effects were either uniquely activated or activated potentially to a higher level by troglitazone.

Evidence stream 3: liver-related human ADRs. *Troglitazone.* The WHO Vigibase database, over the 4-year period 1998 to 2001, contained 6021 records of adverse events on troglitazone (liver-related + non-liver-related), with 1348 unique case IDs. 49 of these are of an unknown category because the event terms are under assessment for WHO ADR terminology (WHO-ART) and 204 adverse event terms are not accepted in WHO-ART, corresponding to "general disorder and administration site conditions" (or MedDRA Preferred Term: "Un-evaluable event") for 247 unique cases. This leaves 5768 total reported adverse events in 4 years for 1348 unique cases (6021 − [49 + 204]).

Rosiglitazone. A total of 1,141 adverse events (liver-related + non-liver-related) were reported for rosiglitazone during the 4 years, 2000 to 2003, since first coming on the market, with 280 unique cases. Of these 1141 adverse events, 9 are of an unknown category and 4 are not accepted in WHO-ART (as described for troglitazone), leaving 1128 (1141 − [9 + 4]) total reported adverse events in 4 years for 280 unique cases. From these data we calculated the comparison of fatal liver-related events between the two drugs (Table 3).

In summary, the analysis of the real world evidence data found that in the first 4 years since drug treaent approval there was a fivefold difference between both the total number of reported adverse events, and the number of unique cases with liver-related reported adverse events. Moreover, when comparing the total fatalities caused by and concurrent with the reported liver-related events, we observed an over tenfold difference with troglitazone-related deaths in just the first year of marketing in the US (54 deaths), and an eightfold difference in fatalities during the first four years on the market. These incidents in the first year since release of troglitazone on the US market served as a basis for the US FDA's decision to withdraw the drug's marketing authorization.

Discussion

To our knowledge, this is the first study to combine evidence from systematic review, in vitro and pharmacovigilance data to compare preclinical animal studies with in vitro methods for their ability to detect human ADRs. Our systematic review found that the animal and human studies did not detect a clear liver safety signal using traditional liver safety biomarkers and, therefore, failed to identify the potential of troglitazone to represent a DILI hazard. In contrast, the in vitro data revealed that troglitazone had twice as much off-target activity as rosiglitazone, alerting to a potential for troglitazone ADRs. The pharmacovigilance data found a clear difference between troglitazone and rosiglitazone in terms of liver injury, with a five-fold higher relative frequency of severe liver adverse events and an eight-fold higher relative frequency of liver fatalities in patients treated with troglitazone compared with rosiglitazone within the first four years of market approval. Our findings provide further evidence to support the accuracy of mechanistic in vitro data to predict drug induced ADRs in vivo, shown previously in many studies, both specific to glitazones and more generally[41,42]. As troglitazone and rosiglitazone are both long established drugs with a wealth of clinical safety data (and withdrawal in the case of Troglitazone) there are also substantial examples of their application to mechanistic in vitro approaches in published literature[43–45].

We took measures to reduce the bias in our systematic review by using two independent reviewers to screen the literature, check data extraction and conduct RoB and certainty in the GRADE assessments. Although we searched three global literature databases, it is possible that we may have missed some studies. Those we found were relatively small and typically involved limited numbers of dose groups with few subjects per group. Fully synthesizing the study findings was difficult due to heterogeneity in study designs, outcomes, doses, treatment times and strains used. Reporting was generally poor for animal studies, with incomplete reporting of outcomes and justifications for study design, species choice and power calculations generally missing. Toxicokinetic measurements were typically not reported, so increases in internal exposure with dose were difficult to assess and, thus, comparison of blood concentrations between animal species and between animals and humans were not possible. The latter would have been an elegant way of comparing species differences. A thorough histopathological evaluation of liver tissue would have been useful in the assessment of both drugs, but histopathology data were not consistently reported and were frequently described in a narrative form without underlying data on individual subjects or even groups. The human studies also had significant reporting limitations, with selective outcome reporting in 40% of studies and an absence of randomisation information in two RCTs. Overall, the potential for RoB in the animal and human studies was significant and confidence in the findings using GRADE was 'low' or 'very low', in part due to poor internal validity. Synthesising evidence from the systematic review with other data sources was challenging because of different methodologies, reporting formats and endpoints, for example, none of the ToxCast test targets map to the traditional liver safety tests evaluated in the systematic review. The ToxCast database itself has limitations, including the lack of metabolic capacity in most of the assays, a multi-step

A.

Troglitazone

Reference		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Mice	Bedoucha 2001	NR	NR	NA	NA	++	NR	NR	+	+	–
	Fujimoto 2009	NR	NR	NA	NA	+	NR	++	+		++
	Jia 2009	NR	NR	NA	NA	+	NR	NR	+		–
	Mak 2018	NR	NR	NA	NA	NR	NR	+	NR		++
	Ong 2007	+	NR	NA	NA	++	NR	+	+	+	++
Rats	Boitier 2011	NR	NR	NA	NA	NR	NR	+	NR	++	++
	Cepa 2018	+	NR	NA	NA	+	NR	++	++	-	++
	Cheng 2017	+	NR	NA	NA	+	NR	+	++	+	++
	Hermann 2002	NR	NR	NA	NA	+	NR	++	++	++	++
	Jia 2000	+	NR	NA	NA	-	NR	NR	++	++	
	Kakiuchi-Kiyota 2011	++	NR	NA	NA	+	NR	++	++		++
	Kostrubsky 2001	NR	NR	NA	NA	-	NR	+	NR	+	++
	Li 2002	NR	NR	NA	NA	+	NR	+	NR		+
	Watanabe 2000	NR	NR	NA	NA	+	NR	NR	NR		-
NHPs	Rothwell 2002	NR	NR	NA	NA	+	NR	++	+	-	++

Rosiglitazone

Reference		Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Mice	Anandharajan 2009	++	NR	NA	NA	+	NR	+	NR	+	+
	Bedoucha 2001	NR	NR	NA	NA	++	NR	NR	+	+	–
	Otake 2011	NR	NR	NA	NA	+	NR	NR	NR	+	+
	Zhang 2019	++	NR	NA	NA	++	++	++	NR	+	++
Rats	Anandharajan 2009	++	NR	NA	NA	+	NR	+	+	+	+
	Anwar 2015	+	NR	NA	NA	+	NR	++	+	-	++
	Dadarkar 2011	NR	NR	NA	NA	+	NR	+	+		++
	Egerod 2009	+	NR	NA	NA	+	NR	++	++	+	++
	Meghani 2012	+	NR	NA	NA	-	NR	+	-	+	++
	Schafer 2012	++	NR	NA	NA	+	NR	+	NR	+	++
	Spicker 2007	NR	NR	NA	NA	–	NR	++	NR	NR	–

B.

Troglitazone

Reference	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Arioglu 2000	NA	NA	NA	-	-	NA	+	+	+	-
Azziz 2001	NR	++	++	+	NA	NR	-	+	+	
Björnsson 2006	NA	NA	NA	-	-	NA	NA	NR	++	++
Knowler 2005	NR	NR	+	NA	NA	-	-	+	-	–
Mavandadi 1999	NA	NA	NA	-	NA	–	++	-	+	-
St. Peter 2001	NA	NA	NA	-	NA	–	–	–	–	–
Yale 2001	++	++	++	NA	NA	++	-	+	-	-

Rosiglitazone

Reference	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Aramwit 2009	NA	NA	NA	-	NA	–	+	+	-	+
Beysen 2008	+	-	NA	NA	-	–	++	+	-	+
Chalasani 2005	NA	NA	NA	-	NA	–	+	-	+	+
Chiang 2007	NA	NA	NA	-	NA	–	+	-	+	+
Dereli 2005	++	-	++	NA	NA	–	+	+	-	+
Gegick 2001	NA	NA	+	-	NA	NA	+	-	+	++
Gegick 2004	NA	NA	+	–	NA	NA	+	-	+	++
Hussein 2004	NA	NA	-	-	NA	–	–	+	-	–
Nolan 2000	++	+	++	NA	NA	++	++	+	+	–
Phillips 2001	+	+	++	++	NA	++	++	+	-	–
Wong 2005	++	++	++	NA	NA	–	++	+	-	

Study or Subgroup	Troglitazone			Control			Weight	Std. Mean Difference IV, Random, 95% CI	Std. Mean Difference IV, Random, 95% CI
	Mean	SD	Total	Mean	SD	Total			
2.1.1 ALT mouse									
Fujimoto 2009 35 week old 300 mg/kg 4 weeks	39.6	8.3	9	25.4	4.6	9	10.8%	2.02 [0.83, 3.20]	
Fujimoto 2009 8 week old 300 mg/kg 4 weeks	38.6	20.3	10	27	3.1	10	13.0%	0.77 [-0.15, 1.68]	
Jia 2019b 100 mg/kg 6 h	71.2	27.7272	5	61.9	0.5657	2	7.7%	0.32 [-1.34, 1.97]	
Jia 2019b 30 mg/kg 6h	52.6	6.9318	5	61.9	0.5657	2	6.4%	-1.26 [-3.18, 0.66]	
Jia 2019b 300 mg/kg 6h	410.4	110.909	5	61.9	0.4	1	2.9%	2.51 [-0.73, 5.75]	
Jia 2019c 300 mg/kg 3h	138.1	36.0007	5	80.3	13.64	5	7.8%	1.92 [0.27, 3.56]	
Jia 2019c 300 mg/kg 6h	138.8	43.6033	5	91.4	12.7456	5	8.9%	1.33 [-0.12, 2.78]	
Mak 2018 0.2% in food 1 week	19.04	2.7713	3	21.95	4.0703	3	7.4%	-0.67 [-2.39, 1.06]	
Mak 2018 0.2% in food 2 weeks	19.54	2.8925	3	19.88	1.3856	3	8.0%	-0.12 [-1.72, 1.48]	
Mak 2018 0.2% in food 3 weeks	17.76	1.9226	3	21.77	0.3811	3	3.8%	-2.31 [-5.06, 0.43]	
Mak 2018 0.2% in food 4 weeks	23.66	3.0831	3	20.43	3.4814	3	7.1%	0.79 [-0.99, 2.56]	
Mak 2018 0.2% in food 5 weeks	25.22	2.8925	3	20.88	2.3209	3	5.9%	1.32 [-0.72, 3.37]	
Ong 2007 300 mg/kg /day 4 weeks	99	34	5	86	28	5	10.3%	0.38 [-0.88, 1.63]	
Subtotal (95% CI)			64			54	100.0%	**0.63 [0.02, 1.23]**	
Heterogeneity: Tau² = 0.51; Chi² = 21.65, df = 12 (P = 0.04); I² = 45%									
Test for overall effect: Z = 2.04 (P = 0.04)									
2.1.2 ALT rat									
Cepa 2018 05 mg/kg 0.25h	59	21	9	63	6	4	5.0%	-0.20 [-1.39, 0.98]	
Cepa 2018 05 mg/kg 0.5h	56	13	9	62	10	4	4.8%	-0.46 [-1.65, 0.74]	
Cepa 2018 05 mg/kg 1h	60	21	10	63	6	4	5.1%	-0.15 [-1.31, 1.01]	
Cepa 2018 05 mg/kg 2h	53	11	10	59	10	4	5.0%	-0.52 [-1.70, 0.66]	
Cepa 2018 05 mg/kg 6h	52	15	10	65	19	4	4.8%	-0.76 [-1.96, 0.45]	
Cepa 2018 05 mg/kg one day	73	10	10	78	16	4	5.0%	-0.40 [-1.57, 0.78]	
Cepa 2018 25 mg/kg 0.25h	63	10	9	63	6	3	4.1%	0.00 [-1.31, 1.31]	
Cepa 2018 25 mg/kg 0.5h	58	10	9	62	10	3	4.0%	-0.37 [-1.69, 0.95]	
Cepa 2018 25 mg/kg 1h	62	9	10	63	6	3	4.2%	-0.11 [-1.40, 1.18]	
Cepa 2018 25 mg/kg 2h	61	14	10	59	10	3	4.2%	0.14 [-1.15, 1.43]	
Cepa 2018 25 mg/kg 6h	64	14	10	65	10	3	4.2%	-0.07 [-1.36, 1.22]	
Cepa 2018 25 mg/kg one day	113	32	10	78	16	3	3.7%	1.09 [-0.29, 2.48]	
Cepa 2018 50 mg/kg 0.25h	65	21	10	63	6	3	4.2%	0.10 [-1.19, 1.39]	
Cepa 2018 50 mg/kg 0.5h	64	11	10	62	10	3	4.2%	0.17 [-1.12, 1.46]	
Cepa 2018 50 mg/kg 1h	68	16	10	63	6	3	4.2%	0.32 [-0.98, 1.61]	
Cepa 2018 50 mg/kg 2h	64	14	8	59	10	3	3.9%	0.35 [-0.99, 1.69]	
Cepa 2018 50 mg/kg 6h	107	37	10	65	19	3	3.7%	1.13 [-0.26, 2.53]	
Cepa 2018 50 mg/kg one day	135	111	10	78	16	3	4.1%	0.53 [-0.79, 1.84]	
Cheng 2017 200 mg/kg 7 days	252	14.4	4	64.6	5.98	4	0.1%	14.78 [4.52, 25.04]	
Jia 2000 200 mg/100g food 44 weeks	35.8	8.1	6	37.2	4.2	3	3.7%	-0.17 [-1.56, 1.22]	
Jia 2000 200 mg/100g food 60 weeks	36	3.2	6	37.2	4.2	3	3.6%	-0.30 [-1.70, 1.09]	
Li 2002 100 mg/kg 3 weeks	10.6	1	5	11	1.5	2	2.6%	-0.30 [-1.96, 1.36]	
Li 2002 500 mg/kg 3 weeks	11.8	1.9	5	11	1.5	2	3.4%	0.39 [-1.06, 1.85]	
Watanabe 2000 female 100 mg/kg 12 weeks	72.2	11.3	5	88.8	10.9	2	2.0%	-1.25 [-3.16, 0.67]	
Watanabe 2000 female 400 mg/kg 12 weeks	69.2	11.9	5	88.8	10.9	3	2.4%	-1.47 [-3.23, 0.28]	
Watanabe 2000 male 100 mg/kg 12 weeks	53.4	9.3	5	75.6	19	2	1.7%	-1.57 [-3.63, 0.49]	
Watanabe 2000 male 400 mg/kg 12 weeks	52	4.3	5	75.6	19	3	2.1%	-1.78 [-3.67, 0.10]	
Subtotal (95% CI)			220			85	100.0%	**-0.12 [-0.40, 0.15]**	
Heterogeneity: Tau² = 0.03; Chi² = 27.80, df = 26 (P = 0.37); I² = 6%									
Test for overall effect: Z = 0.87 (P = 0.38)									
2.1.3 ALT non human primate									
I Rothwell 2002 femal 300 mg/kg 52 weeks	47	12.6	4	69	36.8	1	12.7%	-1.27 [-4.05, 1.51]	
Rothwell 2002 female 1200 mg/kg 52 weeks	46	9.2	4	69	36.8	2	26.7%	-0.92 [-2.83, 1.00]	
Rothwell 2002 female 600 mg/kg 52 weeks	39	15.6	4	69	36.8	1	11.7%	-1.40 [-4.29, 1.49]	
Rothwell 2002 male 1200 mg/kg 52 weeks	47	18.8	4	59	7.1	2	30.7%	-0.58 [-2.36, 1.21]	
Rothwell 2002 male 300 mg/kg 52 weeks	39	12	4	59	7.1	1	13.1%	-1.21 [-3.94, 1.52]	
Rothwell 2002 male 600 mg/kg 52 weeks	42	4.4	4	59	7.1	1	5.1%	-2.81 [-7.18, 1.56]	
Subtotal (95% CI)			24			8	100.0%	**-1.05 [-2.04, -0.06]**	
Heterogeneity: Tau² = 0.00; Chi² = 1.00, df = 5 (P = 0.96); I² = 0%									
Test for overall effect: Z = 2.08 (P = 0.04)									
2.1.4 ALT human									
Mavandadi 1999 600 mg 12 weeks	36	10	19	41.5	17.9	9	48.6%	-0.41 [-1.21, 0.39]	
Mavandadi 1999 600 mg 26 weeks	35	8.1	19	41.5	17.9	10	51.4%	-0.51 [-1.29, 0.26]	
Subtotal (95% CI)			38			19	100.0%	**-0.47 [-1.02, 0.09]**	
Heterogeneity: Tau² = 0.00; Chi² = 0.03, df = 1 (P = 0.86); I² = 0%									
Test for overall effect: Z = 1.63 (P = 0.10)									

-4 -2 0 2 4
Lower with T   Lower with C

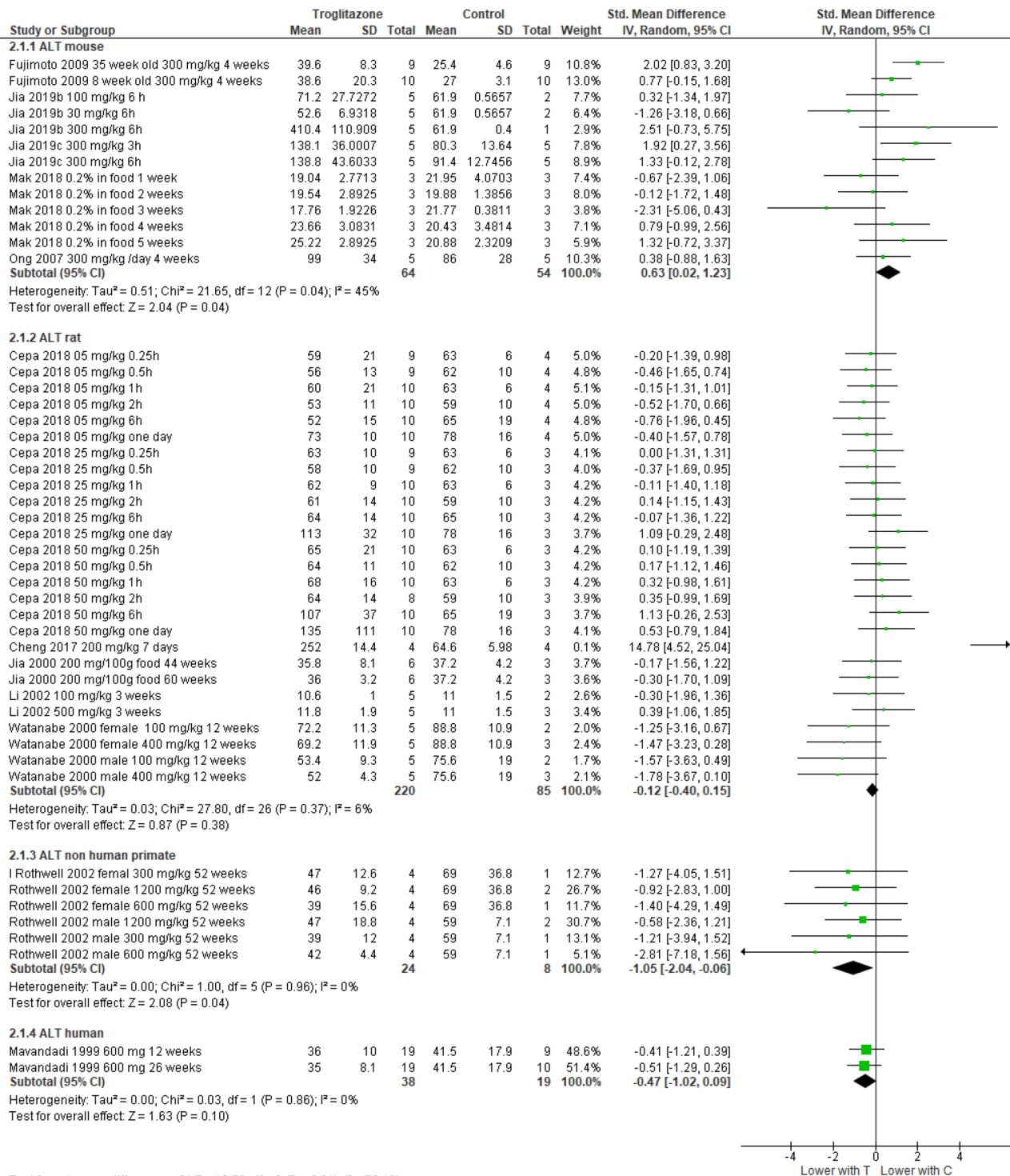Test for subgroup differences: Chi² = 10.76, df = 3 (P = 0.01), I² = 72.1%

Figure 3. Forest plots are presented separately for troglitazone and for rosiglitazone. Each outcome is sorted by species, increasing dose of the drug and increasing follow-up periods.
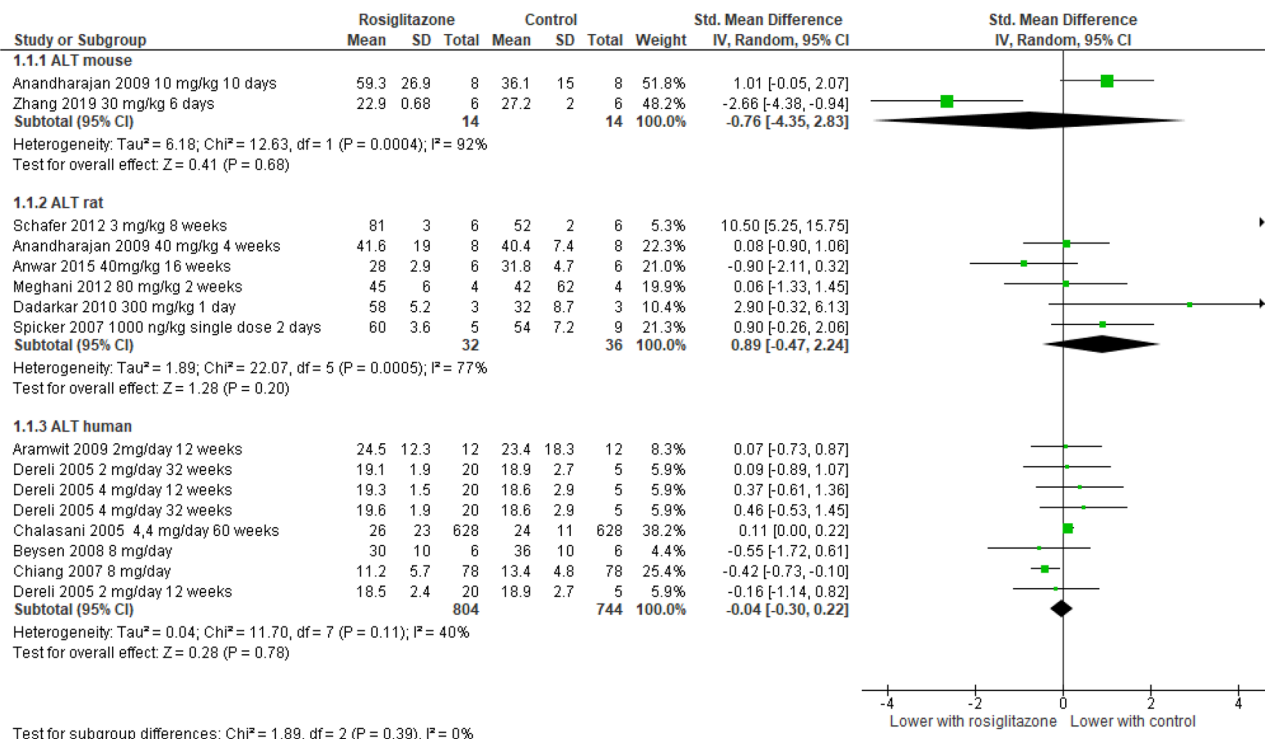
Figure 3. (continued)

process to access the raw data and metadata of individual experiments, and a lack of data on the exposure of cells to the compounds. Thus, there is a potential for bias in ToxCast experiments. Furthermore, not all drugs are tested in all assays, resulting in many missing data points, and a potential bias towards "more popular" or "more toxic" drugs. The NAS-based approach we present here is a way to assess the relevance of in vitro findings in the context of human exposure, but it is based on population C_{max} values. Inter-individual differences in pharmacokinetics and metabolism exist, leading to a range of C_{max} values which are not accounted for here and may be at least partially responsible for the ADRs reported for drugs on the market. Our research objective was limited to investigating the in vitro tests included in the ToxCast database. We are aware of the progress made to date with development of the advanced in vitro models, including organ-on-chip 3-dimentional human tissue models[46,47], and believe that the logical follow-up to this study should be a systematic review of published data on other in vitro test systems, which will undoubtedly expand the biological space and help refine the findings of our study. However, such investigation was outside the scope of this study. In terms of Vigibase, the main weakness is the lack of data on the number of prescriptions issued, making it difficult to compare the incidence of adverse effects between drugs. Hence, we used time on the market since approval in order to compare the number of adverse events between the two drugs. Furthermore, the Vigibase data, although classified using a controlled vocabulary, are descriptive, making them difficult to compare with other evidence streams.

It is important to note that in the troglitazone studies, the elevated liver weights in NHPs, the elevations in ALT, AST, ALP and liver weights in mice and the elevated ALP in rats, were not strong enough to be regarded as warning signals for human DILI risk. Furthermore, the current preclinical testing regime is not sufficiently robust for predicting adverse effects in human populations with low incidences, interpersonal variability, and/or where mechanisms are unknown. The DILI induced by troglitazone appears to be idiosyncratic (i.e. rare, caused by agents that have little or no intrinsic toxic activity, unpredictable, not dose dependent, not reproducible in animal models and with a variable latency period[23]). This might explain the inconsistent DILI signals in animal and human studies. When a drug enters the market, the number of patients treated with the drug increases, raising the chance of detecting idiosyncratic events. Although our analysis of pharmacovigilance data found a clear difference between troglitazone and rosiglitazone liver injury, this was a retrospective analysis; what is needed is a means of *preventing* adverse events. Our study suggests that in vitro data, typically available in the early stages of drug development, may help identify drugs that cause DILI, and can provide insight into the mechanisms of potential adverse effects. The use of such in vitro assays currently is mostly limited to early drug discovery stages. This review provides evidence that mechanistic approaches have a great potential to support regulatory review and approval, thereby supplementing the mandated animal safety data, which are frequently not sufficiently conclusive. The ToxCast in vitro data suggest that if we expand the types of assays acceptable in regulatory submissions to include in vitro human biology-based data, potentially unsafe compounds can be prevented from entering the market and causing human suffering.

All the studies included in the systematic review were published *after* both drugs had already been approved for use. Some of the studies might have been conducted before (but only published after) approval to contribute

Species	Troglitazone***					Rosiglitazone***				
	ALT SMD (95% CI)	AST SMD (95% CI)	ALP SMD (95% CI)	T. Bilirubin SMD (95% CI)	Liver Weight SMD (95% CI)	ALT SMD (95% CI)	AST SMD (95% CI)	ALP SMD (95% CI)	T. Bilirubin SMD (95% CI)	Liver Weight SMD 95% CI)
Mouse	0.63* (0.02, 1.23)	1.15** (0.55, 1.75)	2.98** (2.00, 3.97)	−0.05 (−0.68, 0.59)	1.32** (0.50, 2.14)	−0.76 (−4.35, 2.83)	−0.63 (−1.45, 0.19)	−5.00** (−7.69, −2.31)	−2.38** (−4.01, −0.76)	0.14 (−0.42, 0.71)
Rat	−0.12 (−0.40, 0.15)	0.12 (−0.03, 0.54)	1.06* (−0.02, 2.14)	−0.50 (−1.25, 0.26)	−0.57 (−1.34, 0.21)	1.42 (−0.18, 3.01)	−0.26 (−1.99, 1.46)	1.60 (−0.48, 3.69)	−1.57 (−3.77, 0.64)	1.08* (0.12, 2.04)
Non-human Primate	−1.05* (−2.04, -0.06)	−0.97* (−1.94, -0.01)	−1.10* (−2.12, -0.09)	−2.56** (−4.14, -0.97)	2.13** (0.76, 3.50)	N/A	N/A	N/A	N/A	N/A
Human	−0.47 (−1.02, 0.09)	−0.26 (−0.81, 0.29)	Not estimable	Not estimable	N/A	−0.04 (−0.30, 0.22)	0.30 (−0.01, 0.60)	0.44* (0.12, 0.76)	0.22** (0.11, 0.33)	N/A

Table 2. Summary of forest plot results of five liver-related outcomes. *$P \leq 0.05$; **$P \leq 0.005$. *SMD* standard mean difference; Not estimable = too few studies reporting endpoint to estimate SMD; N/A = no studies reporting endpoint; ***See Table 1 for species-specific dose information.
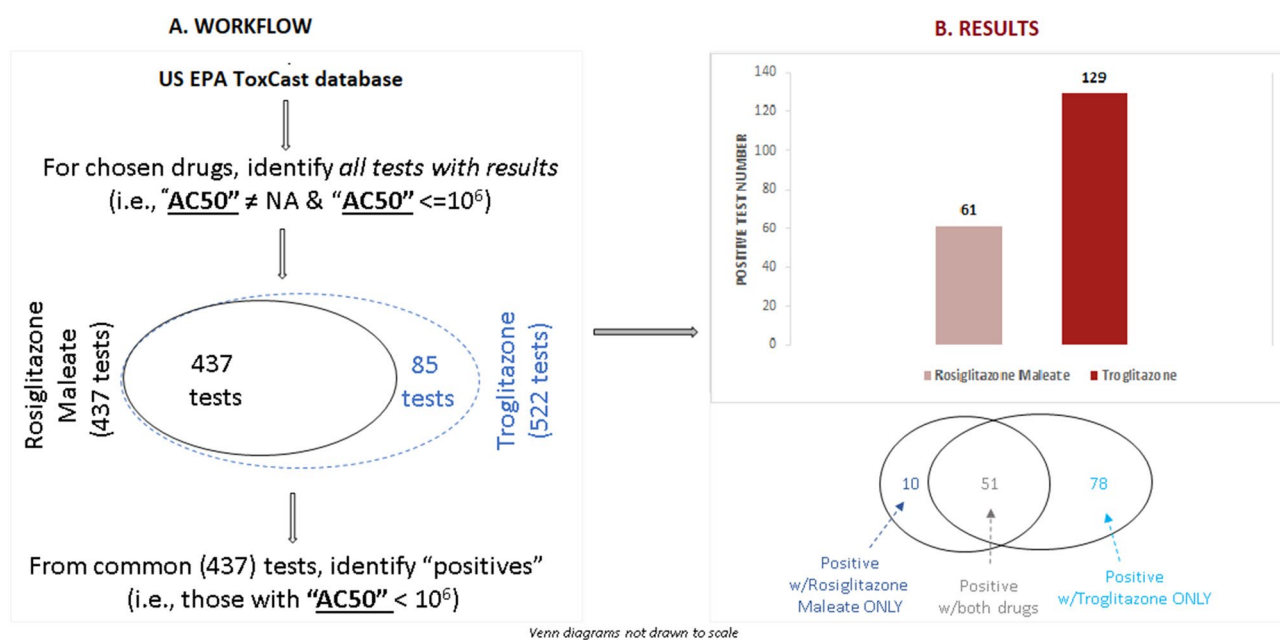
Figure 4. ToxCast database analysis: workflow to identify the tests performed on the two drugs (**A**) and results based on the number of positive tests (**B**).

information to regulatory agencies, but this seems unlikely since the study designs do not comply with Organisation for Economic Co-operation and Development (OECD) guidelines. This means we lack some of the data that informed regulatory decisions. We did not have access to data from the OECD guideline studies conducted under Good Laboratory Practices from the sponsor pharmaceutical companies, since they have not been published and are not disclosed by regulatory agencies. We suspect that this 'hidden evidence' creates significant publication bias. The lack of access to unpublished regulatory studies, individual animal and raw data presents challenges when conducting systematic reviews in the field of toxicology and we strongly recommend that these data are made public. We also recommend that reporting and publication standards need to markedly improve. Data on pharmacokinetics, toxicokinetics and metabolism are often missing or not reported in published pre-clinical safety studies. This makes assessments of the toxicological warning signals and especially interspecies comparison, including comparison to human predicted exposure, very challenging. The reporting of primary data with individual animal data is needed for all outcomes, particularly for histopathology, and would enable more data included in quantitative meta-analyses. There is a need to adapt a tool such as GRADE for uncertainty analysis of in vitro studies, which will help regulators to critically evaluate and accept in vitro data. We also recommend that regulators and industry look beyond traditional safety biomarkers and incorporate more sensitive and human biology-specific biomarkers, which are necessary to study the effects of chemicals on the human body. miRNA and advanced proteomics are promising approaches for discovering human biomarkers, but further work is needed to develop, validate, and use them. There are plenty of potential biomarkers in Tox-Cast (where tests were selected in 2005) and more in the literature published since. A comprehensive map of in vitro tests is needed, preferably mapped to human biomarkers, as an amendment of existing regulations and test guidelines to allow for their use. Mechanistic data are important for understanding toxicological pathways
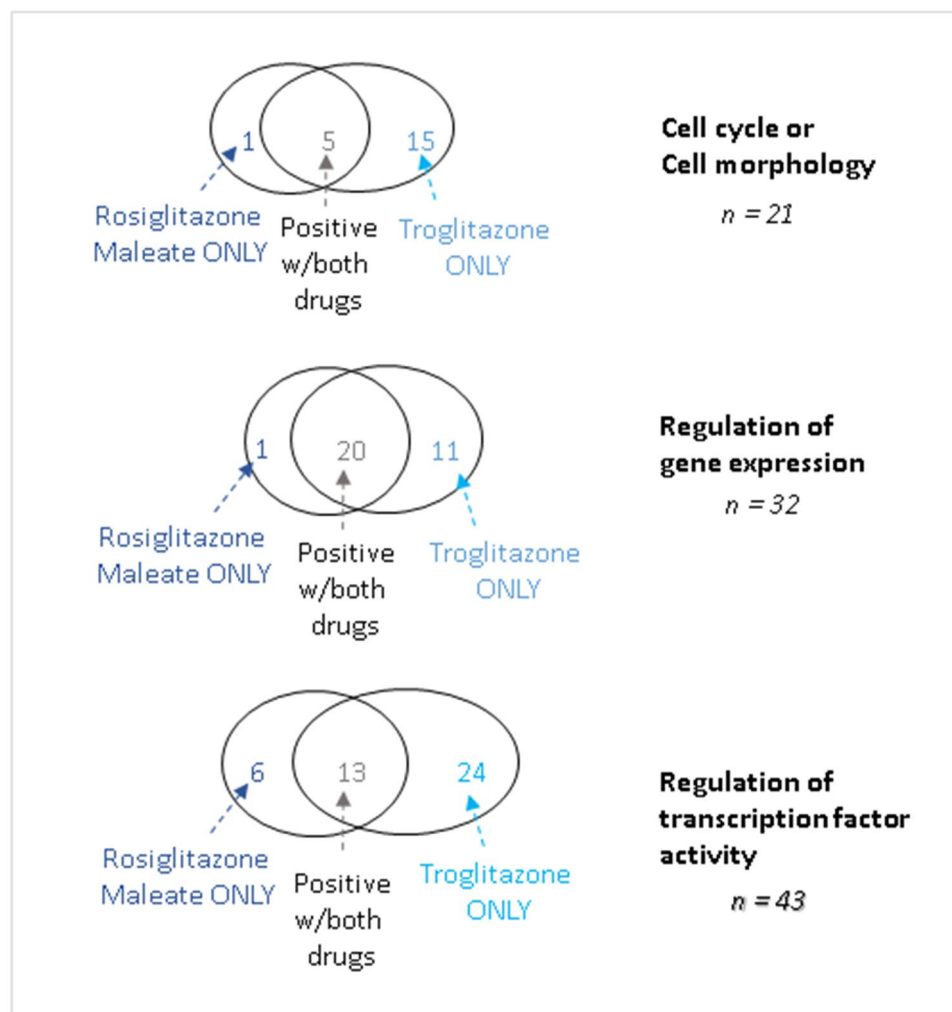
Figure 5. ToxCast database analysis: number of positive tests for troglitazone and rosiglitazone segmented by biological processes.
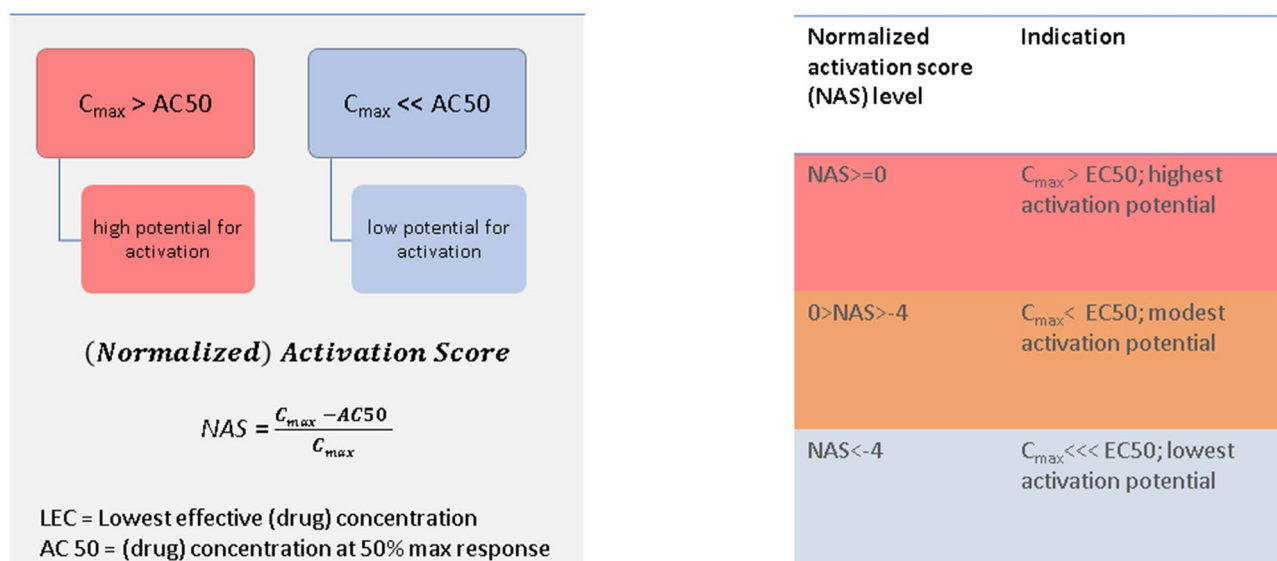
Figure 6. Definitions of Noramlized Activation Score and biological targets activation potentials.

Outcome	Troglitazone[a]	Rosiglitazone[b]
Total reported adverse events (WHO-ART Corrected)	5768	1128
Liver-specific adverse events		
Total Deaths*	221	28
Not recovered/not resolved**	8	36
Recovered/resolved with sequelae	134	4
Recovered/resolved	9	30
Unknown	35	27
Not reported	941	155
Total hepatic adverse events	1348	280

Table 3. Comparison of hepatic and non-hepatic adverse events reported in the first 4 years following market release. *To calculate Total Deaths, MedDRA terms for Death (Event may have contributed), Death (Cause of death other than event), Fatal (Cause of death is event) were added; **WHO controlled vocabulary definitions can be found, e.g. in https://www.who.int/medicines/publications/Pharmaco_TB_web_v2.pdf. [a]Troglitazone adverse events reported from 1998–2001. [b]Rosiglitazone adverse events reported from 2000–2003.

General	Liver disorder, hepatotoxicity, liver injury, hepatotoxic effect, hepatic pain
Liver function biomarkers	Alanine aminotransferase abnormal, alanine aminotransferase increased, aspartate aminotransferase, aspartate aminotransferase abnormal, aspartate aminotransferase increased, bilirubinaemia, bilirubin increased, blood bilirubin abnormal, blood bilirubin increased, hepatic enzyme increased, hepatic enzymes increased, hyper-bilirubinemia, liver function test abnormal, liver function tests abnormal nos*
Medical diagnoses	Hepatic failure, hepatitis granulomatous, acute hepatic failure, hepatitis fulminant, hepatic disease, hepatitis nos, autoimmune hepatitis, hepatitis acute, hepatomegaly, hepatitis chronic active, hepatosplenomegaly, hepatitis cholestatic, jaundice, hepatitis cholestatic, jaundice cholestatic, hepatitis, jaundice nos, ocular icterus, yellow skin, hepatic encephalopathy, hepatic infarction, hepatitis A, hepatitis A antibody positive, hepatitis B antibody
Biliary tract disorder	Biliary cirrhosis, bile duct stone, bile duct stricture, bile duct carcinoma, portal hypertension, hepatorenal syndrome, hypertension portal, gallbladder disorder, cholecystitis, cholecystitis chronic, cholelithiasis, cholestasis intrahepatic, cholestasis, cholangitis, portal vein thrombosis, hepatic cirrhosis, cryptogenic cirrhosis
Histological changes	Biopsy liver abnormal, hepatic necrosis, hepatic steatosis, hepatocellular damage, fatty liver
Liver malignancies	Hepatic neoplasm, hepatic neoplasm malignant, hepatic metastases

Table 4. Liver-related adverse events coded as MedDRA terms in Vigibase were classified into 6 prime categories of important liver endpoints. *nos* not otherwise specified.

leading to adverse events and the OECD Adverse Outcome Pathway (AOP) offers a framework for organising these data[48,49]. More knowledge on the mechanisms of toxicity would reduce the number of in vivo studies and allow for more targeted testing, preventing unnecessary human suffering and death. However, regulators need confidence in NAMs and mechanistic test results, which means that in vitro approaches need to be robustly evaluated for relevance and quality. The use of human biology- based in vitro test systems and in silico predictions in hazard identification and characterization hold great promise but regulatory acceptance of the data generated in these models is essential in order to replace animal experiments in regulatory review processes. To increase our mechanistic understanding of toxicity, we recommend that all compounds intended for human use be tested in all validated ToxCast or similar assays so that more data become available on the relationships between chemical structure and biological effects. Furthermore, the assays in the current ToxCast program could be mapped to the AOP initiatives and should include more assays with potential molecular initiating events and key events[50]. This would increase understanding of the mode of actions of chemicals and identify the various toxicological pathways, including the ones for DILI, enabling an explanation as to why, for example, troglitazone causes more toxicity to the liver than rosiglitazone. Indeed, three-dimensional (3D) hepatic organoid cultures have emerged as promising tools to assess the mechanisms and risks of hepatotoxicity in drug discovery[51], providing improved metabolic activity and an enhanced liver phenotype not achievable with conventional two-dimensional hepatic models. Modelling complex liver diseases including nonalcoholic steatohepatitis (NASH) is now possible using differentiated human pluripotent stem cells to produce functional bile canaliculi systems disrupted by cholestasis-inducing drugs such as troglitazone. Additionally, dysregulation of biliary- and hepatocyte-associated genes, as seen in NASH patient tissue samples, has been observed in these organoids[52]. 3D cultures of liver microtissues have repeatedly outperformed primary human hepatocytes in correctly classifying hepatotoxicants from different pharmacological classes of molecules[51]. In risk assessments of chemicals, authorities like the European Food Safety Authority (EFSA) rely more on epidemiological data than mechanistic data[53]. Associations discovered in these studies should be confirmed by mechanistic data in order to demonstrate their biological plausibility.

In conclusion, we found that neither published animal nor human studies, taken together, were able to accurately predict the potential of troglitazone to represent a DILI hazard in humans, while in vitro data were able to detect the hazard. Our findings indicate that the paired-compound approach to comparing various

tests, pioneered here, could be used to evaluate the relevance and predictivity of in vitro human biology-based approaches, bringing a systematic, transparent, and evidence-based approach to drug development. Adopting such approaches could make new drugs safer and reduce late attrition, preventing unfortunate human ADRs and deaths and making the drug development process more financially sustainable. We also propose that transparent protocol-driven evidence-based approaches should become standard in preclinical research and that this would benefit the pharmaceutical industry, society, and first and foremost, patients.

Methods

We combined three evidence streams: a systematic literature review of published human clinical trials and animal studies (Evidence Stream 1); in vitro data from the US EPA ToxCast database (Evidence Stream 2); and human ADRs from Vigibase, a global pharmacovigilance database of individual case safety reports run by the World Health Organisation (WHO) (Evidence Stream 3).

Evidence stream 1: systematic review of in vivo *studies*.   The aim of the systematic review was to investigate the effect on liver function of in vivo exposures to troglitazone and rosiglitazone in humans and standard preclinical animal models (rats, mice, dogs, and NHPs), as outlined in the registered systematic review protocol CRD42018112353[27].

Search strategy.   On February 7th 2020, PubMed, Embase (Embase.com), and Web of Science (Clarivate Analytics) were searched as outlined in the protocol[27]. The search strategies were developed by a medical librarian (RW) in collaboration with the review team and peer-reviewed by GV. The complete search strategies are presented in S1.

Eligibility criteria.   Populations.   Controlled studies of troglitazone or rosiglitazone in humans and experimental rats, mice, dogs and non-human primates (NHPs) with hepatic endpoints were included. Humans with diseases that are not a primary indication for the drugs of interest were excluded, as were patients with pre-existing liver injury. Genetically modified animal models or disease models were excluded.

Interventions.   Studies were included that indicated the oral, intravenous, or intraperitoneal administration of troglitazone or rosiglitazone. Studies of drug combinations were excluded.

Controls.   Human and animal studies with a control or placebo group, or pre-treatment values, were included.

Outcomes.   Standard clinical and preclinical hepatic effects of troglitazone or rosiglitazone in experimental animals and humans were included. Specifically, plasma levels of liver enzyme tests (alanine transaminase (ALT); aspartate transaminase (AST); alkaline phosphatase (ALP); and bilirubin), histopathology results, absolute and relative liver weight, were included. Mechanistic non-standard observations (i.e. gene expression, proteomics, micro RNAs, and similar observations) were excluded.

Study types.   English, Norwegian, Dutch, Swedish and Danish studies were included. Conference abstracts, narrative reviews, opinion papers, case reports and other publication types where the original outcome data are not reported, were excluded. We also excluded reports of general adverse events (AE) monitoring, AE database analyses and single AE case reports from Evidence Stream 1 for two reasons:

1.   Pharmacovigilance reports do not include controls;
2.   These events are included in the pharmacovigilance reports (Evidence Stream 3), thus including them in Stream 1 would be double counting.

Screening and data extraction.   Two reviewers independently screened the literature for relevant studies, first using titles and abstracts and then full texts. Sysrev online software (Insilica LLC, Bethesda, MD, US) was used to manage the screening process. DistillerSR online software (Evidence Partners, Ottawa, ON, Canada) was used to extract the data from included studies.

Analysis.   For each drug, data from human and animal studies were analyzed separately. For each outcome/endpoint, data from each species, dose, follow-up time, and study design were collected and presented in forest plots and tables. Continuous data are presented as standardized mean difference with 95% confidence intervals. Where we considered the population, dose, design, and comparison similar enough, we conducted a meta-analysis based on a random effect model. The internal validity of both animal and human studies was assessed using the OHAT risk-of-bias (RoB) tool[54]. Two reviewers independently assessed the OHAT criteria for each included study, with disagreements resolved by discussion and with the help of a third reviewer if necessary. Our confidence in the evidence was assessed by two reviewers as well, using the GRADE (Grading of Recommendations Assessment, Development and Evaluation) approach[55].

Evidence stream 2: in vitro mechanistic data.   The in vitro data were obtained from US EPA ToxCast database which is the largest curated collection of > 1100 molecular and cellular assays for up to 10,000 chemicals, including ~ 500 FDA-approved drugs[22]. The US EPA contributes the ToxCast data to a US federal

agency collaboration known as Tox21 (Toxicology in the 21st Century) which pools chemical research, data, and screening tools from several US federal agencies. As part of the EPA's commient to share its chemical data openly and transparently, all ToxCast chemical screening data are publicly available via the ToxCast dashboard, which allows users to search and query the data.

Datasets of troglitazone and rosiglitazone were retrieved from the ToxCast database (invitrodb_v2) and analysis was done according to pre-registered protocol: https://doi.org/10.5281/zenodo.252909. All assays, without pre-selection, in which the drugs were tested in the ToxCast database were used in the analysis. The ToxCast data processing pipeline consists of multiple processing steps resulting in progressively refined estimates of AC_{50} values (compound concentration at 50% of maximal assay activity)[56]. For this analysis, we retrieved Level 5 data containing AC_{50} values from the best performing model used to fit dose–response curves. All AC_{50} values were used as downloaded per description above and no further processing was performed.

Analysis. All assays/targets and their pathways were compared across the troglitazone/rosiglitazone drug pair using normalised activation score (NAS) values. The NAS score was developed here to adjust the activities of the drugs in the in vitro tests (AC_{50}) to human exposure. The NAS value reflects the difference between the average peak concentration that a drug achieves in plasma after its administration to humans at the highest dose (C_{max}) and the AC_{50} value for the assay in an in vitro test. Normalisation of in vitro data to human maximal plasma concentrations was first proposed for use in the normalisation of drug-drug interactions[57] and subsequently was used for normalisation of in vitro data[58]. The human C_{max} used here is taken from the US FDA drug insert for each drug. Troglitazone has a reported C_{max} of 1.61 μg/mL at the standard therapeutic dose of 400 mg/day[59]. Rosiglitazone has a reported C_{max} of 156 ng/mL at the standard therapeutic dose of 2 mg/day[60]. The NAS formula used:

$$NAS = \frac{C_{max} - AC_{50}}{C_{max}}$$

The NAS value was used as a ranking metric to stratify assays using the following logic:

- All assays/targets thereof with NAS value > 0 were ranked to have a putative "higher activation potential" or higher possible off-target/toxic effect, because the C_{max} concentrations in patients are expected to exceed AC_{50} values.
- All assays/targets thereof with NAS values > − 4 but NAS < 0 were ranked to have a putative "modest activation potential" or modest possible off-target effect, because, in this case, drug plasma concentration is equal to and up to fivefold lower than the AC_{50} value.
- All assays/targets thereof with NAS values < − 4 were ranked to have a putative "low activation potential" or unlikely to cause off-target/toxic effects at these concentrations, because drug plasma concentration is more than fivefold lower than the AC_{50} value.

All assays or their affected gene targets/pathways were compared across the drug pair using NAS values either directly or after grouping them within their respective biological processes as represented within these assay descriptions in the ToxCast database. Gene targets/pathways associated with the assays were identified using the ToxCast assay list of targets. Differentially affected gene targets/pathways were identified based on differences in NAS values (e.g. higher for troglitazone vs. rosiglitazone) and literature linking these gene targets/pathways with liver injury or repair pathways. The analysis was performed according to the pre-published protocol[61]. All analysed data files, and the R code used as part of this analysis, are available at: (https://github.com/Sri-Bandhakavi/ToxCast_Rosiglitazone_Troglitazone_EBTC_Analysis).

Evidence stream 3: liver-related human ADRs. Vigibase is the largest continuously-updated database in the world, with over 20 million reports of suspected adverse effects of medicines submitted since 1968 by member countries of the WHO Programme for International Drug MonitoringWHO Programme for International Drug Monitoring. On January 2, 2018 individual cases where liver toxicity was reported as an adverse event were retrieved separately from VigiBase for troglitazone and rosiglitazone. Data was analysed using the general WHO Guideline for using Vigibase data[62]. To accommodate for the differences in total time on the market for the two drugs and latency in reporting, Vigibase ADR data for the first 4 years since each drug's approval by US FDA were used in this analysis (i.e. 1998–2001 for troglitazone and 2000–2003 for rosiglitazone).

As real world evidence, the liver-related adverse events reported in Vigibase are considered the "gold standard" for human safety in our study; these data were analysed separately for troglitazone and rosiglitazone in the following step-wise analysis:

- The number of unique cases was counted. The total liver-related adverse events count was compared with other non-liver-related adverse events for outcomes for the first 4 years after approval in the US, and plotted against the dose and demographic characteristics.
- The liver-related adverse events coded as MedDRA (Medical Dictionary for Regulatory Activities) terms in Vigibase were classified into 6 prime categories of important liver endpoints (Table 4).
- The proportional reporting ratio and probability of occurrence were calculated for 4 years each for rosiglitazone and troglitazone separately[63].

All analyses were performed according to the pre-published protocol http://doi.org/10.5281/zenodo.2528922[64].

Ethics approval. Ethics approval was not required for this study as it involves the analysis of secondary data sources only. The views expressed in this manuscript do not necessarily represent those of the U.S. Food and Drug Administration.

Transparency statement. The lead author (the manuscript's guarantor) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as originally planned and registered have been explained.

Data availability
The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

References
1. US Food and Drug Administration, "Preventable Adverse Drug Reactions: A Focus on Drug Interactions. ADRs: Prevalence and incidence," https://www.fda.gov/drugs/drug-interactions-labeling/preventable-adverse-drug-reactions-focus-drug-interactions (Accessed 24 Mar 2020).
2. Pirmohamed, M. *et al.* Adverse drug reactions as cause of admission to hospital: Prospective analysis of 18,820 patients. *Br. Med. J.* **329**(7456), 15–19 (2004).
3. Bouvy, J. C., De Bruin, M. L. & Koopmanschap, M. A. Epidemiology of adverse drug reactions in Europe: a review of recent observational studies. *Drug Saf.* **38**(5), 437–453 (2015).
4. Gwathmey, J. K., Tsaioun, K. & Hajjar, R. J. Cardionomics: a new integrative approach for screening cardiotoxicity of drug candidates. *Expert Opin. Drug Metab. Toxicol.* **5**(6), 647–660 (2009).
5. Elliott, R., Camacho, E., Campbell, F., Jankovic, D., St James, M. M., Kaltenthaler, E., & Faria, R. Prevalence and economic burden of medication errors in the NHS in England: rapid evidence synthesis and economic analysis of the prevalence and burden of medication error in the UK www.eepru.org.uk/wp-content/uploads/2020/03/medication-error-report-edited-27032020.pdf (Accessed 27 Mar 2020), (2018).
6. Mullin, R. Tufts study finds big rise in cost of drug development. *Chem. Eng. News* http://www.cen.acs.org/articles/92/web/2014/11/Tufts-Study-Finds-Big-Rise.hl (Accessed 24 Mar 2020), (2014).
7. Wouters, O. J., McKee, M. & Luyten, J. Estimated research and development invesent needed to bring a new medicine to market, 2009–2018. *JAMA* **323**(9), 844 (2020).
8. Thomas, D. W., et al. Clinical development success rates 2006–2015. https://www.bio.org/sites/default/files/legacy/bioorg/docs/Clinical%20Development%20Success%20Rates%202006-2015%20-%20BIO,%20Biomedtracker,%20Amplion%202016.pdf (Accessed 27 Mar 2020).
9. Harrison, R. K. Phase II and phase III failures: 2013–2015. *Nat. Rev. Drug Discov.* **15**(12), 817–818 (2016).
10. Pampaloni, F., Reynaud, E. G. & Stelzer, E. H. K. The third dimension bridges the gap between cell culture and live tissue. *Nat. Rev. Mol. Cell Biol.* **8**(10), 839–845 (2007).
11. Ravi, M., Paramesh, V., Kaviya, S. R., Anuradha, E. & Solomon, F. D. P. 3D cell culture systems: advantages and applications. *J. Cell. Physiol.* **230**(1), 16–26 (2015).
12. BioIndustry Association and Medicines Discovery Catapult, State of the Discovery Nation 2018 and the role of the Medicines Discovery Catapult, http://www.md.catapult.org.uk/FlipBuilder/mobile/index.hl (Accessed 24 Mar 20)), (2018).
13. Geerts, H. Of mice and men. Bridging the translational disconnect in CNS drug discovery. *CNS Drugs* **23**(1), 915–926 (2009).
14. Perrin, S. Preclinical research: make mouse studies work. *Nature* **507**, 423–425 (2014).
15. Wall, R. J. & Shani, M. Are animal models as good as we think?. *Theriogenology* **69**(1), 2–9 (2008).
16. Waring, M. J. *et al.* An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nat. Rev. Drug Discov.* **14**(7), 475–486 (2015).
17. Hwang, T. J. *et al.* Failure of investigational drugs in late-stage clinical development and publication of trial results. *JAMA Intern. Med.* **176**(12), 1826 (2016).
18. Alliance for Human Relevant Science, Accelerating the growth of human relevant life sciences in the United Kingdom. A white paper by the alliance for human relevant science, www.humanrelevantscience.org/wp-content/uploads/Accelerating-the-Growth-of-Human-Relevant-Sciences-in-the-UK_2020-final.pdf (Accessed 27 Mar 2020), (2020).
19. Innovate UK, A non-animal technologies roadmap for the UK: advancing predictive biology, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/474558/Roadmap_NonAnimalTech_final_09Nov2015.pdf (Accessed 3 Mar 2020), (2015).
20. Archibald, K., Tsaioun, K., Kenna, J. G. & Pound, P. Better science for safer medicines: the human imperative. *J. R. Soc. Med.* **111**(12), 433–438 (2018).
21. Avila, A. M. *et al.* An FDA/CDER perspective on nonclinical testing strategies: classical toxicology approaches and new approach methodologies (NAMs). *Regul. Toxicol. Pharmacol.* **114**, 104662 (2020).
22. US Environmental Protection Agency, "Toxicity forecasting. Advancing the next generation of chemical evaluation," www.epa.gov/chemical-research/toxicity-forecasting (Accessed Mar 24 2020), (2019).
23. Hoofnagle, J. H. & Björnsson, E. S. Drug-induced liver injury: types and phenotypes. *N. Engl. J. Med.* **381**(3), 264–273 (2019).
24. Watkins, P. B. Drug safety sciences and the bottleneck in drug development. *Clin. Pharmacol. Ther.* **89**(6), 788–790 (2011).
25. Olson, H. *et al.* Concordance of the toxicity of pharmaceuticals in humans and in animals. *Regul. Toxicol. Pharmacol.* **32**(1), 56–67 (2000).
26. Hoffmann, S. *et al.* A primer on systematic reviews in toxicology. *Arch. Toxicol.* **91**(7), 2551–2575 (2017).
27. Tsaioun, K. *et al.* Systematic reviews of the hepatotoxic effects of specified compounds/drugs as observed in humans and experimental animals (rats, mice, dogs and non-human primates). PROSPERO 2018 CRD42018112353. Available from https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42018112353.
28. Scheen, A. J. Thiazolidinediones and liver toxicity. *Diabetes Metab.* **27**(3), 305–313 (2001). PMID: 11431595.
29. Isley, W. L. & Oki, J. C. Hepatotoxicity of thiazolidinediones. *Diabetes Obes. Metab.* **3**, 389–392 (2000).
30. Chen, M. *et al.* DILIrank: the largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. *Drug Discov. Today* **21**(4), 648–653 (2016).

31. Dereli, D., Dereli, T., Bayraktar, F., Ozgen, A. G. & Yilmaz, C. Endocrine and metabolic effects of rosiglitazone in non-obese women with polycystic ovary disease. *Endocr. J.* **52**(3), 299–308 (2005).
32. Rothwell, C., McGuire, E. J., Altrogge, D. M., Masuda, H. & Iglesia, F. A. Chronic toxicity in monkeys with the thiazolidinedione antidiabetic agent troglitazone. *J. Toxicol. Sci.* **27**(1), 35–47 (2002).
33. Duarte, S., Baber, J., Fujii, T. & Coito, A. J. Matrix metalloproteinases in liver injury, repair and fibrosis. *Matrix Biol.* **44–46**, 147–156 (2015).
34. Saiman, Y. & Friedman, S. L. The role of chemokines in acute liver injury. *Front. Physiol.* **3**, 213 (2012).
35. Böhm, F. *et al.* FGF receptors 1 and 2 control chemically induced injury and compound detoxification in regenerating livers of mice. *Gastroenterology* **139**(4), 1385–1396 (2010).
36. Rudraiah, S., Zhang, X. & Wang, L. Nuclear receptors as therapeutic targets in liver disease: are we there yet?. *Annu. Rev. Pharmacol. Toxicol.* **56**(1), 605–626 (2016).
37. Zimmermann, H. W., Trautwein, C. & Tacke, F. Functional role of monocytes and macrophages for the inflammatory response in acute liver injury. *Front. Physiol.* **3**, 56 (2012).
38. Gomez-Ospina, N. *et al.* Mutations in the nuclear bile acid receptor FXR cause progressive familial intrahepatic cholestasis. *Nat. Commun.* **7**(1), 10713 (2016).
39. Jadeja, R. N., Upadhyay, K. K., Devkar, R. V. & Khurana, S. Naturally occurring Nrf2 activators: potential in treaent of liver injury. *Oxid. Med. Cell. Longev.* **2016**, 1–13 (2016).
40. Zúñiga, S., Firrincieli, D., Housset, C. & Chignard, N. Vitamin D and the vitamin D receptor in liver pathophysiology. *Clin. Res. Hepatol. Gastroenterol.* **35**(4), 295–302 (2011).
41. Yu, L. *et al.* Integrating in vitro testing and physiologically-based pharmacokinetic (PBPK) modelling for chemical liver toxicity assessment: a case study of troglitazone. *Environ. Toxicol. Pharmacol.* **74**, 103296 (2020).
42. Yokoyama, Y. *et al.* Comparison of drug metabolism and its related hepatotoxic effects in heparg, cryopreserved human hepatocytes, and HepG2 cell cultures. *Biol. Pharm. Bull.* **41**, 722–732 (2018).
43. Lloyd, S. *et al.* Differential in vitro hepatotoxicity of troglitazone and rosiglitazone among cryopreserved human hepatocytes from 37 donors. *Chem. Biol. Interact.* **142**, 57–71 (2002).
44. Edling, Y., Sivertsson, L. K., Butura, A., Ingelman-Sundberg, M. & Ek, M. Increased sensitivity for troglitazone-induced cytotoxicity using a human in vitro co-culture model. *Toxicol. Vitr.* **23**, 1387–1395 (2009).
45. Usui, T., Hashizume, T., Katsumata, T., Yokoi, T. & Komuro, S. In vitro investigation of the glutathione transferase M1 and T1 null genotypes as risk factors for troglitazone-induced liver injury. *Drug Metab. Dispos.* **39**, 1303–1310 (2011).
46. Khetani, S. R. *et al.* Use of micropatterned cocultures to detect compounds that cause drug-induced liver injury in humans. *Toxicol. Sci. Off. J. Soc. Toxicol.* **132**(1), 107–117 (2013).
47. Bale, S. S. *et al.* In vitro platforms for evaluating liver toxicity. *Exp. Biol. Med.* **239**(9), 1180–1191 (2014).
48. Villeneuve, D. L. *et al..* Adverse outcome pathway (AOP) development I: Strategies and principles. *Toxicol. Sci.* **142**(2), 312–320. https://doi.org/10.1093/toxsci/kfu199 (2014).
49. OECD. Guidance Document for the Use of Adverse Outcome Pathways in Developing Integrated Approaches to Testing and Assessment (IATA), OECD Series on Testing and Assessment, No. 260, OECD Publishing, Paris. https://doi.org/10.1787/44bb06c1-en (2017).
50. Jeong, J. & Choi, J. Development of AOP relevant to microplastics based on toxicity mechanisms of chemical additives using ToxCast and deep learning models combined approach. *Environ. Int.* **137**, 105557 (2020).
51. Proctor, W. R. *et al.* Utility of spherical human liver microtissues for prediction of clinical drug-induced liver injury. *Arch. Toxicol.* **91**, 2849–2863 (2017).
52. Bin Ramli, M. N. *et al.* Human pluripotent stem cell-derived organoids as models of liver disease. *Gastroenterology* **159**(4), 1471–1486 (2020).
53. EFSA Panel on Contaminants in the Food Chain (EFSA CONTAM Panel), Risk to human health related to the presence of 2 perfluoroalkyl substances in food. *EFSA J.* https://doi.org/10.2903/j.efsa.2020.6223
54. Rooney, A. A., Boyles, A. L., Wolfe, M. S., Bucher, J. R. & Thayer, K. A. Systematic review and evidence integration for literature-based environmental health science assessments. *Environ. Health Perspect.* **122**(7), 711–718 (2014).
55. Balshem, H. *et al.* GRADE guidelines: 3. Rating the quality of evidence. *J. Clin. Epidemiol.* **64**(4), 401–406 (2011).
56. Filer, D. L., Kothiya, P., Setzer, R. W., Judson, R. S. & Martin, M. T. tcpl: the ToxCast pipeline for high-throughput screening data. *Bioinformatics* **33**(4), btw680 (2016).
57. Obach, R. S., Walsky, R. L. & Venkatakrishnan, K. Mechanism-based inactivation of human cytochrome P450 enzymes and the prediction of drug-drug interactions. *Drug Metab. Dispos.* **35**, 246–255 (2007).
58. Xu, J. J. *et al.* Cellular imaging predictions of clinical drug-induced liver injury. *Toxicol. Sci.* **105**, 97–105 (2008).
59. US Food and Drug Administration, Application number 020720, S12, S14. (1999).
60. US Food and Drug Administration, Medication Guide (Avandia). (2019).
61. Tsaioun, K. *et al..* Evidence-based methodology for identifying drugs with potential human liver toxicities using US EPA ToxCast data set (Version 6). Zenodo. https://doi.org/10.5281/zenodo.2529091 (2018).
62. W. C. C. for I. D. monitoring Uppsala Monitoring Centre, "Guideline for using VigiBase data in studies," 2018. Available: www.who-umc.org. (Accessed: 05 Feb 2021).
63. Evans, S. J. W., Waller, P. C. & Davis, S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiol. Drug Saf.* **10**(6), 483–486 (2001).
64. Tsaioun, K, et al. (2018, December 30). Evidence-based methodology for identifying drugs with potential human liver toxicities using WHO Vigibase data (Version 1). Zenodo. http://doi.org/10.5281/zenodo.2528922

Acknowledgements

Author contributions

The Evidence-based Toxicology Collaboration (EBTC) Working Group at Johns Hopkins Bloomberg School of Public Health contributed throughout the project, acting at various times as a wider stakeholder group. K.T. and H.D. planned the study and contributed equally to the conduct and coordination of the project. The specific contributions are listed as follows: G.E.V. conducted the meta-analysis and risk-of-bias and GRADE assessments of human studies; S.E.F. conducted the risk-of-bias and GRADE assessments of animal studies; S.B. conducted the US EPA ToxCast data analysis; J.M. conducted the WHO Vigibase adverse events data analysis; R.R. provided

input in protocols development, screening of the literature and manuscript drafting; C.H.C.L. provided support in animal data meta-analysis in the systematic review; B.K. conducted the risk-of-bias assessments of animal studies in the systematic review, data extraction and quality control, and PRISMA adherence; R.A.W. guided the working group in developing a search strategy and conducted literature searches for the systematic review; and M.C. provided input during planning stages and stakeholder feedback at various stages of the project as well as in developing the manuscript. P.P. assembled and drafted revisions of the manuscript, which all authors contributed to. The majority of the co-authors participated in screening the literature for the systematic review. All authors have read and approved the final version of the manuscript. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-85708-2.

Correspondence and requests for materials should be addressed to K.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.