International Journal of Epidemiology, 2018, 1–13
doi: 10.1093/ije/dyy239
Original article

Original article

How often can meta-analyses of individual-level data individualize treatment? A meta-epidemiologic study

Ewoud Schuit [1,2,3,4]* Alvin H Li[1,5] and John PA Ioannidis[1,2]

[1]Departments of Medicine, of Health Research and Policy, of Biomedical Data Science and of Statistics, Stanford University, Stanford, CA, USA, [2]Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA, USA, [3]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands, [4]Cochrane Netherlands, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands and [5]Clinical Epidemiology Program, The Ottawa Hospital Research Institute, Ottawa, Canada

*Corresponding author. Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Huispost nr. STR 6.131, PO Box 85500, 3508 GA Utrecht, The Netherlands. E-mail: e.schuit@umcutrecht.nl

Editorial decision 27 September 2018; Accepted 8 October 2018

Abstract

Background: One of the claimed main advantages of individual participant data meta-analysis (IPDMA) is that it allows assessment of subgroup effects based on individual-level participant characteristics, and eventually stratified medicine. In this study, we evaluated the conduct and results of subgroup analyses in IPDMA.

Methods: We searched PubMed, EMBASE and the Cochrane Library from inception to 31 December 2014. We included papers if they described an IPDMA based on randomized clinical trials that investigated a therapeutic intervention on human subjects and in which the meta-analysis was preceded by a systematic literature search. We extracted data items related to subgroup analysis and subgroup differences (subgroup–treatment interaction $p < 0.05$).

Results: Overall, 327 IPDMAs were eligible. A statistically significant subgroup–treatment interaction for the primary outcome was reported in 102 (36.6%) of 279 IPDMAs that reported at least one subgroup analysis. This corresponded to 187 different statistically significant subgroup–treatment interactions: 124 for an individual-level subgrouping variable (in 76 IPDMAs) and 63 for a group-level subgrouping variable (in 36 IPDMAs). Of the 187, only 7 (3.7%; 6 individual and 1 group-level subgrouping variables) had a large difference between strata (standardized effect difference $d \geq 0.8$). Among the 124 individual-level statistically significant subgroup differences, the IPDMA authors claimed that 42 (in 21 IPDMAs) should lead to treating the subgroups differently. None of these 42 had $d \geq 0.8$.

Conclusions: Availability of individual-level data provides statistically significant interactions for relative treatment effects in about a third of IPDMAs. A modest number of these interactions may offer opportunities for stratified medicine decisions.

Key words: individual participant data meta-analysis, individual patient data meta-analysis, IPDMA, subgroup analysis, differential treatment effect, aggregate data meta-analysis

Key Messages

- About 85% of individual participant data meta-analyses (IPDMAs) reported subgroup analyses, of which 37% had subgroup–treatment interaction p-values < 0.05.
- Two-thirds of significant subgroup–treatment interactions were found for individual-level subgrouping variables and one-third for group-level subgrouping variables.
- Only 4% of subgroup differences had a large difference in standardized treatment effect between strata of the subgroup ($d \geq 0.8$).
- In one-third of significant individual-level subgroup–treatment interactions, the authors claimed that this finding should lead to treating subgroups differently.

Introduction

Available evidence on treatment effectiveness can be summarized by a systematic review followed by quantitative evidence synthesis by a meta-analysis. Such meta-analyses can be performed either using published or aggregate data (AD) or individual participant data (IPD). IPD meta-analysis (IPDMA) is considered the 'gold-standard' of meta-analysis, since IPD allows e.g. standardizing outcome definitions and analyses across studies, potentially including longer follow-up times.[1–3] Despite this golden label, the treatment estimates themselves are generally comparable between the two types of meta-analysis, especially when reported outcome data and treatment and outcome definitions are uniform across studies.[4,5] A main advantage of IPDMA over AD meta-analysis (ADMA), and as such an important reason why IPDMAs are conducted, is to assess subgroup effects, i.e. to identify differential treatment effects according to subgroups, specifically those that are based on patient-level characteristics (e.g. gender, age, comorbidities, etc).[1,3,6] Such individual-level characteristics typically would not be possible to evaluate for their impact on the treatment effect based on published or aggregate group data. Group data allow the evaluation of subgroup effects only for variables that take the same value in all patients randomized in each trial arm (e.g. treatment dose). In theory, the conduct of subgroup analyses addresses the quest of modern individualized (aka stratified, personalized or precision) medicine, i.e. making different treatment decisions for individual patients based on different expected treatment effects for such specific patients. Currently, it is unclear how often meta-analyses of individual-level data facilitate such individualized treatment decisions.

In this study, we evaluated the subgroup analyses reported in IPDMAs, specifically by assessing the type of subgrouping variable (group- or individual-level), and the statistical significance and magnitude of these subgroup effects in terms of subgroup–treatment interaction and subgroup-specific treatment-effect estimates. Finally, we tried to assess whether the most prominent subgroup differences observed are considered by their authors to offer support for stratified treatment decisions.

Methods

The study is a survey of IPDMAs, i.e. a meta-epidemiological assessment. We report our study according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines for the reporting items where this is appropriate.[7] No formal protocol exists for this study.

IPDMA search and selection strategy

Three searches were performed. First, we screened for eligibility all IPDMAs that were included in a previously published systematic review of IPDMAs.[8] Second, we updated the search strategy of that review to 31 December 2014 by searching PubMed, EMBASE and the Cochrane Library. Since we noticed some IPDMAs to be missing from that review, especially those published by trialists' collaboration groups and those led by some prominent, highly experienced UK groups performing IPDMAs, we supplemented the search strategy by searching PubMed also using the following search terms: 'trialists' OR ((Stewart L[au] OR Peto R[au] OR Collins R[au]) AND overview)'. No language restrictions were used.

We included studies if they described an IPDMA based on randomized clinical trials (RCTs) that investigated an intervention on human subjects and in which the meta-analysis

was preceded by a systematic literature search. For example, we excluded studies if they only used trials from a pharmaceutical company drug database or when they combined data from a select number of RCTs without systematically reviewing the literature. These papers were excluded because they did not aim to perform a meta-analysis based on all available evidence. We included IPDMAs that referred to a publication or protocol in which a systematic search strategy was presented. For updated IPDMAs (e.g. Cochrane Reviews), we included only the most recent paper. If duplicates existed that presented the exact same study (e.g. in a specialist journal and in the Cochrane Database of Systematic Reviews), we only included the first publication.

One author (E.S.) performed study selection, first based on title only, then based on abstract. We excluded studies only if it was very clear that they did not fulfil the inclusion criteria. When unclear, we considered the study eligible for full-text assessment. We performed full-text assessment alongside data extraction and in duplicate (E.S., A.H.L.). Any disagreement was solved by a third reviewer (J.P.A.I.).

Data extraction and statistical analysis

Data were extracted for the following types of categories from each paper: general information (e.g. country, overarching disease) and subgroup analysis (e.g. conduct of subgroup analysis, identification of statistically significant subgroup–treatment interactions; see below). We pilot tested data extraction on 20 randomly selected papers and made adjustments after discussion. Similarly to full-text assessment, we performed data extraction in duplicate (E.S. and A.H.L.) and resolved discrepancies by discussion. A third author (J.P.A.I.) arbitrated the remaining disagreements.

For each IPDMA with a statistically significant subgroup–treatment interaction ($p_{\text{interaction}} < 0.05$), we extracted the total number of statistically significant subgroup–treatment interactions. We considered only first-level subgroup analyses on the primary outcome and ignored second- and higher-level subgroup analyses (i.e. subgroups within subgroups) and subgroup analyses on secondary outcomes. We considered an outcome to be a primary outcome when labelled as such by the authors of the IPDMA. When multiple outcomes were labelled as the primary outcome(s), we considered results from subgroup analyses for each of these outcomes. Additionally, we did not consider statistically significant subgroup–treatment interactions on the absolute risk scale. For each statistically significant interaction, we collected the subgrouping variable itself (e.g. gender), whether the subgrouping variable was on the group level (i.e. similar for all participants within the trial arm, e.g. control treatment) or on the individual level (i.e. information that may differ on the level of individual participant, e.g. blood pressure), the *p*-value for interaction and the treatment effects within strata of the subgroup (i.e. separately for females and males). When the treatment effects in each of the strata of the subgroup were not clearly reported in the paper, we either used a digitizer to extract the effect estimate from a plot (WebPlotDigitizer: http://arohatgi.info/WebPlotDigitizer/app/) or we contacted the IPDMA principal investigator to provide additional information.

Next, we calculated the relative treatment effects between subgrouping strata using the subgroup-specific treatment effects, e.g. odds ratio (OR) in females vs odds ratio in males. We used the smallest treatment effect as the reference, resulting in relative effects >1 for relative measures of association and >0 for continuous measures, i.e. (standardized) mean differences. For example, if the OR of mortality with treatment vs placebo was 2.0 in females and 0.5 in males, the relative treatment effect for mortality of females vs males would be $2.0/0.5 = 4$. Relative treatment effects calculated from relative measures of association [e.g. OR, risk ratio (RR) and hazard ratio (HR)] were transformed into a standardized effect using the Chinn transformation: Ln(relative measure)/1.81.[9] So, in the example of females vs males, the standardized effect would equal $\text{Ln}(\text{OR females/OR males})/1.81 = \text{Ln}(4)/1.81 = 0.77$. Mean differences were transformed to standardized effects where possible by dividing the differences into mean differences between subgroups by the pooled standard deviation of both subgroups. If the information reported was not sufficient to determine a standardized effect, we refrained from calculating a standardized effect. This happened in three of the six IPDMAs that reported mean differences, representing five of their nine significant subgroup–treatment interactions.

Additionally, as subgroup analysis in IPDMA allows assessment of treatment effectiveness based on individual participant characteristics, we also extracted whether the authors of the IPDMAs made any claim for clinical relevance of the identified subgroup effect, i.e. treated different subgroups differently (e.g. treated only one subgroup). We also examined in more detail the most prominent subgroup differences, i.e. those where the standardized effect differed by $d \geq 0.8$ across subgroups, corresponding to a ratio of relative measures of association of 4.25 or higher. Such standardized effect differences are considered to be large[10] and we chose to apply this threshold to differences in standardized effects as well. RRs and HRs may deviate less from the null than ORs, and thus assuming them interchangeable with OR may result in somewhat larger *d* estimates than the true values.

Statistical analysis was descriptive, with continuous variables being presented as medians with interquartile ranges (IQRs) and categorical or dichotomous variables as a number and percentage of the total number of included

IPDMAs or number of (statistically significant) subgroup analyses.

Results

Included studies

From database inception to 31 December 2014, we identified 726 publications (Figure 1). Of these, 399 were excluded because the paper did not perform or did not clearly report or refer to a systematic literature search ($n = 194$), paper did not actually present IPDMA (e.g. meta-analysis based on published data, $n = 39$), an update was available ($n = 33$), included observational studies ($n = 26$), full text was not available ($n = 26$), was a duplicate publication ($n = 24$), presented no comparison of treatments ($n = 11$) or other reasons ($n = 46$). The other reasons included no intention to collect IPD from all eligible trials, paper presented conference abstract, case report/series, methodological work, prognostic work, single randomized–controlled trial, protocol for review of IPDMAs, commentary, editorial, no reported meta-analysis of interventions. One publication presented an analysis that was considered a subgroup analysis of one of the included papers and, as a paper, was excluded, but results of the subgroup analysis were considered. Eventually, data were extracted for 327 IPDMAs (see Supplementary data at *IJE* online for a list of IPDMAs).

General information

Of the 327 IPDMAs, 107 (32.7%) had been published after 2010. Many IPDMAs were published in a high-impact general medical or specialty journal (*Lancet* $n = 54$, 16.5%; *Journal of Clinical Oncology* $n = 26$, 8.0%; *BMJ* $n = 19$, 5.8%; *European Heart Journal* $n = 8$, 2.4%; *NEJM* $n = 5$, 1.5%; *JAMA* $n = 3$, 0.9%; other $n = 212$, 64.8%). The overarching disease fields that were investigated in the IPDMA were most often cancer (32.4%) and cardiovascular disease (30.3%), followed by infectious diseases (7.3%), neurological disease (4.3%), obstetrics and gynaecology-related disease (4.3%), musculoskeletal disease (4.3%), mental disease (3.1%) and other disease (11.3%). Studied interventions included drugs ($n = 174$, 53.2%), biologics ($n = 23$, 7.0%), surgical or other procedure ($n = 24$, 7.3%), radiotherapy ($n = 12$, 3.7%) and other ($n = 61$, 18.7%). The majority of IPDMAs had one primary outcome ($n = 186$, 56.9%), whereas 69 (21.1%) had two and 71 (21.7%) had three or more primary outcomes. The primary outcome was unclear for one IPDMA. The outcomes that were most often investigated were mortality ($n = 159$, 48.6%), a composite outcome including

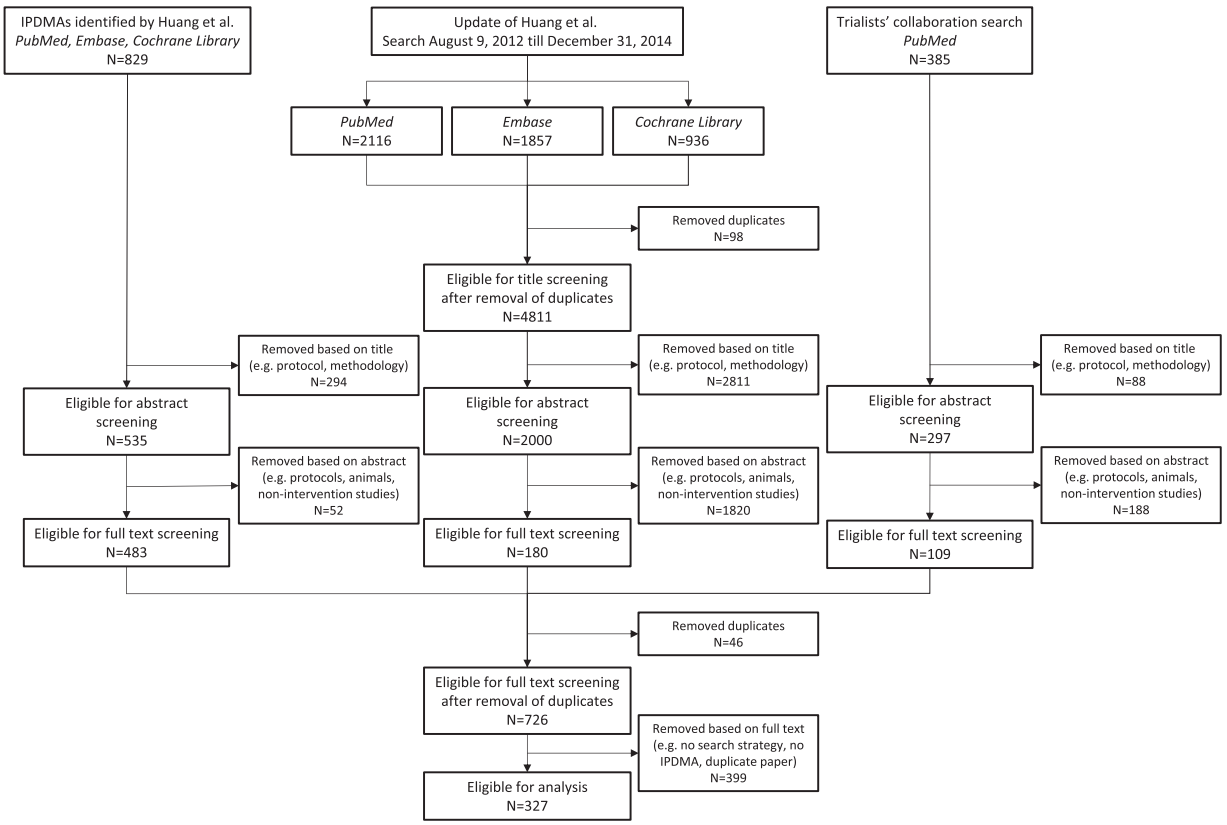

Figure 1. Flow chart of study identification, presented separately for the three searches.

mortality ($n = 62$, 19.0%) and recurrence, progression or disease-free survival ($n = 58$, 17.7%). The IPDMAs included IPD from a median of eight trials (IQR 5–14), corresponding to a median of 2376 (IQR 965–6358) participants per IPDMA.

Subgroup analysis

Subgroup analysis was reported in the majority of IPDMAs ($N = 279$, 85%) and 102/279 (36.6%) reported one ($n = 54$, 53%), two ($n = 30$, 29%) or between three and eight ($n = 18$, 18%) statistically significant subgroup–treatment interactions for the primary outcome. Sixty-six out of 102 (64.7%) IPDMAs reported statistically significant subgroup–treatment interactions for individual-level subgrouping variables only, 26 (25.5%) IPDMAs for group-level subgrouping variables and 10 (9.8%) IPDMAs for both individual and group-level subgrouping variables. In total, these 102 IPDMAs combined reported 187 statistically significant subgroup–treatment interactions: 124 (66.3%) for an individual-level subgrouping variable and 63 (33.7%) for a group-level subgrouping variable. Of the 187 significant subgroup–treatment interactions found, 45 (24%) reported an OR, 65 (35%) an RR, 62 (33%) an HR, 9 (4.8%) a standardized mean difference and 6 (3.2%) a mean difference. IPDMAs that reported no, one, two, or three or more significant subgroup–treatment interactions showed to have included an increasing number of studies [median (IQR) of 6 (4–12), 11 (6–15), 11 (7–16) and 12 (6–24), respectively] and participants [median (IQR) of 1984 (868–6028), 3229 (1719–7353), 3332 (1968–8559), 5407 (2387–22 635), respectively].

Subgrouping variables and outcomes

As shown in Table 1, the most common individual-level subgrouping variables for which a statistically significant interaction with the treatment was reported were cancer-related factors (e.g. nodal status or performance status) ($n = 25$, 20.2%), age ($n = 20$, 16.1%) and cardiovascular disease-related factors (e.g. stenosis group or seizure type) ($n = 17$, 13.7%). Group-level subgrouping variables mostly included subgroups related to type of treatment (e.g. type of chemotherapy or combination vs single agent) ($n = 16$, 25.4%), dosage ($n = 13$, 20.6%) and additional treatment (e.g. background treatment or some other intervention performed during treatment) ($n = 7$, 11.1%). Statistically significant subgroup–treatment interactions were most often reported for the outcomes of mortality ($n = 49$, 39.5% for individual-level subgrouping variables and $n = 24$, 38.1% for group-level subgrouping variables), a composite outcome including mortality ($n = 23$, 18.6%; $n = 7$, 11.1%) or recurrence of disease ($n = 11$, 8.9%; $n = 18$, 28.6%), regardless of whether the investigated subgroup was an individual-level or group-level variable, respectively.

Subgroup effects

The distribution of the subgroup–treatment interaction p-values and the differences in standardized effects can be found in Figures 2 and 3, respectively. Both figures indicate mostly modest p-values and small subgroup effects. The median p-value was 0.02 (IQR 0.007–0.034) and 0.008 (IQR $2e^{-3}$–0.02) and the median effect size difference was 0.15 (IQR 0.09–0.37) and 0.17 (IQR 0.12–0.21), which is slightly less than 0.2, a small effect, for significant individual-level and group-level subgrouping variables, respectively. For seven (3.7%) subgroup effects (six for an individual- and one for a group-level subgrouping variable), the difference between strata of the subgroup was ≥ 0.8, indicating a large effect. The outcomes for these seven large subgroup effects were mortality ($n = 2$),[11,12] treatment failure at 12 months, defined as persisting symptoms and signs of otitis media ($n = 1$),[13] parasitological failure by day 14 ($n = 1$),[14] allograft survival at 2 years ($n = 1$) and 5 years ($n = 1$)[15] and quality of life ($n = 1$).[16] For 17 subgroup analyses (8.9%; 13 individual-level and 4 group-level), we were unable to determine relative effect sizes because treatment effects were not presented for each stratum of the subgroup.

Claims of clinical relevance

Authors claimed clinical relevance of the subgroup effect by suggesting different treatments for different strata of the subgroups in 42 (33.9%) out of 124 individual-level subgroups with a statistically significant subgroup–treatment interaction. Of these 42, the outcomes were mortality ($n = 17$, 40.5%), a composite outcome including mortality ($n = 9$, 21.4%), pain ($n = 6$, 14.3%) and other outcomes ($n = 10$, 23.8%). More details are presented in Table 2. For 66 (53.2%) subgroup analyses, no such claim was made and, for 16 (12.9%), the authors suggested further research was needed.

For two out of the six individual-level subgroup variables with a difference in standardized treatment effects across strata ≥ 0.8, the authors suggested further research, whereas the other four made no claim about the clinical relevance of the identified subgroup effect. The two IPDMAs that found a difference ≥ 0.8 for mortality did not make any claim[11] or suggested future research.[12]

Discussion

Our empirical evaluation of 327 IPDMAs shows that approximately one-third reported a statistically significant subgroup–treatment interaction and approximately a

Table 1. Characteristics of statistically significant subgroup analyses (p-value for interaction <0.05; $N = 187$)

	Individual-level subgrouping variable ($N = 124$)	Group-level subgrouping variable ($N = 63$)
Subgrouping variable		
Cancer-related factor	25 (20.2)	N.P.
Age	20 (16.1)	N.P.
Cardiovascular disease-related factor	17 (13.7)	N.P.
Baseline risk or severity	12 (9.7)	N.P.
Gender	8 (6.5)	N.P.
Diabetes	7 (5.7)	N.P.
Race	3 (2.4)	N.P.
Dosage	N.P.	13 (20.6)
Type of treatment	N.P.	16 (25.4)
Additional treatment	N.P.	7 (11.1)
Trial type/type of comparison	N.P.	5 (7.9)
Control treatment	N.P.	2 (3.2)
Treatment duration	N.P.	3 (4.8)
Other	32 (25.8)	17 (27.0)
Types of outcome		
Mortality	49 (39.5)	24 (38.1)
Composite outcome including mortality	23 (18.6)	7 (11.1)
Recurrence, progression or disease-free survival	11 (8.9)	18 (28.6)
Cardiovascular disease	11 (8.9)	N.P.
Pain	6 (4.8)	3 (4.8)
Treatment/device failure	4 (3.2)	2 (3.2)
Depression	4 (3.2)	N.P.
Other	16 (12.9)	9 (14.3)
Subgroup effects		
p-value interaction		
Median (IQR)	0.02 (0.007–0.034)	0.008 (0.002–0.02)
Min–max	1e-6–0.05	1e-6–0.05
<0.0001	2 (1.6)	9 (14.3)
Difference in standardized treatment effects		
Median (IQR)	0.15 (0.09–0.37)	0.17 (0.12–0.21)
Min–max	0–1.32	0–0.26
$d \geq 0.8$	6 (4.8)	1 (1.6)
p-value unclear or difference not available	16 (12.9)	6 (9.5)

N.P., not present; IQR, interquartile range.

quarter reported such interactions that could not have been examined based on group-level published information. In one-third of subgroup analyses with a statistically significant subgroup–treatment interaction, the authors claimed the subgroup effect to be clinically relevant by suggesting treatment or withholding treatment of patients based on the subgroup effect. However, none of such claims was made for the subgroups with a large difference in the standardized treatment effects reported in subgroup strata.

Some potential limitations should be discussed. First, we may have missed some eligible IPDMAs despite our use of multiple, complementary search strategies. Considering the large number of IPDMAs included in our review, we suspect this will not have influenced the overall findings of our study. Second, we did not extract information on the number of subgroup analyses performed. Such information would have helped in estimating the proportion of subgroup analyses that show a statistically significant subgroup–treatment interaction. However, in our pilot data extraction, we noticed that reporting of the planned subgroup analyses, i.e. the outcome and the subgrouping variables of interest, was often not clear, which is why we focused only on reported subgroups with a statistically significant subgroup–treatment interaction. We cannot verify which of these statistically significant interactions pertain to pre-planned analyses. Similar reporting issues were reported in a recent review that studied current reporting and statistical practice in IPD systematic reviews.[6] Third, we applied the Chinn transformation to RR and HR,

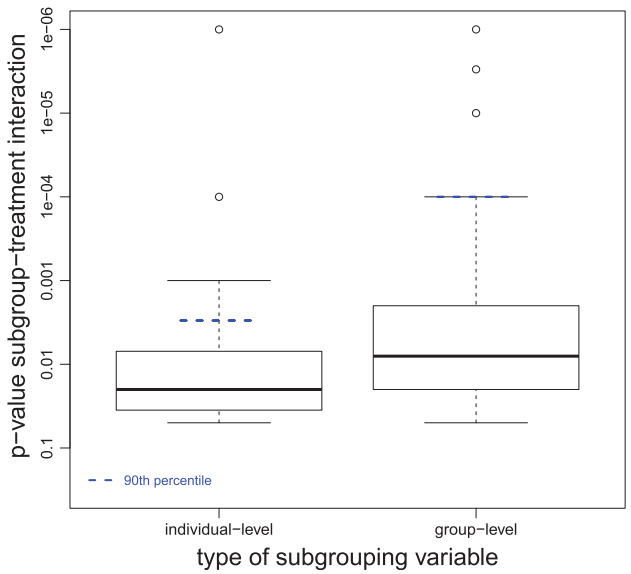

Figure 2. Boxplots of the *p*-value of subgroup–treatment interactions on the Log10 scale, reported in subgroup analyses.

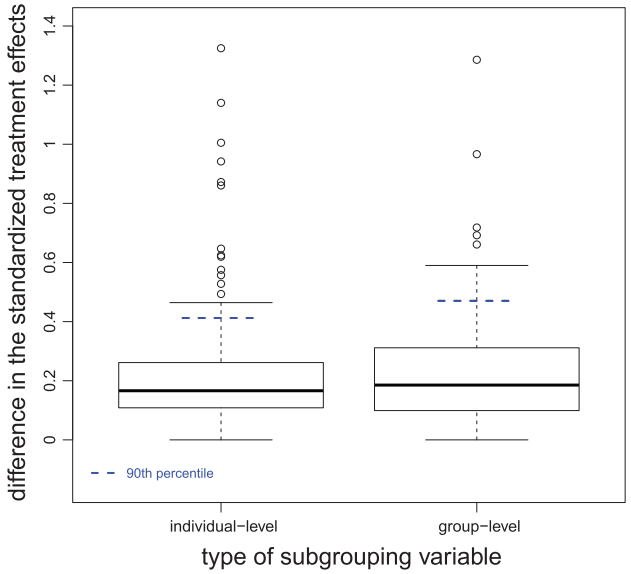

Figure 3. Boxplots of the difference in the standardized treatment effects reported in subgroup strata.

whereas it was specifically designed for the OR.[9] As we discussed in the 'Methods' section, our estimates of differences between subgroups through the *d* metric may be, on average, larger than the true differences, if any. Therefore, the percentage of IPDMAs with truly $d \geq 0.8$ may be even less than the 3.7% that we observed. We should stress that the magnitude of the difference, regardless of how it is measured and expressed, is only one among many factors that might be used in making decisions about personalized treatment. Treating patients in different strata differently cannot depend on any single number, nor on arbitrary levels of statistical significance. This is why we also strongly focused on the interpretation of the IPDMAs' authors to understand how often they hint at different treatment management for different subgroups. This needs to be decided on a case-by-case basis, considering efficacy, safety, cost and many other circumstances. The proportion of IPDMAs with such differential treatment recommendations was nevertheless still very modest, accounting for only 7% (21/327) of the examined IPDMAs.

A previous empirical evaluation (whose search results we also included in our searches) has reported on the distribution and epidemiological characteristics of 829 IPDMAs[8] and compared results from overall and subgroup analyses of a subset of 204 IPDMAs and their matching ADMA.[4] Whereas we only focused on IPDMAs that included RCTs that assessed a therapeutic intervention, Huang *et al.* included IPDMAs that assessed all types of research questions (i.e. therapeutic, prognostic, diagnostic, etc.) and included data from RCTs, observational studies and other study designs, which is why their number of identified IPDMAs was larger than in our study. Huang *et al.* reported that 59.3% of IPDMAs reported subgroup analyses, which is lower than the 85.3% in our study. This difference is probably due to our stricter inclusion criteria of only IPDMAs that combined randomized controlled trials. These IPDMAs are more likely to assess subgroup effects than diagnostic or prognostic IPDMAs. They also reported that 44 out of 544 subgroup analyses (8.1%) reported a statistically significant subgroup–treatment interaction, which is substantially lower than the 36.6% we found in our study. The main explanation for this difference is that we did not assess the number of subgroup analyses conducted, but used the—obviously much smaller—number of IPDMAs as a denominator instead. If we assume that all of the 44 statistically significant subgroup–treatment interactions identified by Huang *et al.* were reported in unique IPDMAs, that would correspond to 44/121 (36.4%)—a proportion close to what we found in our study. Huang *et al.* did not assess the outcomes investigated, the distribution of the interaction *p*-value or the relative treatment effects across subgroups or the claims made as to their clinical relevance.

The availability of individual-level information yielded statistically significant subgroup differences in about a quarter of the IPDMAs. However, only approximately 1 in 25 of all subgroups with a statistically significant subgroup–treatment interaction showed a large relative standardized treatment effect across strata of the subgroup. The authors of the IPDMAs made a modest number of claims for treating differentially people in different subgroups based on these observed interactions. This evidence may thus offer some opportunities for practising stratified medicine that would not have been available from

Table 2. Overview of individual participant data meta-analyses that claim differential treatment effects based on subgroup analysis ($N = 21$)

Study	Intervention	Disease	Outcome	Inferences and suggestions of authors based on treatment effects and p-value for subgroup–treatment interaction
Rovers *et al.*[17]	Ventilation tubes	OME	Hearing level	Treat children attending day-care not those not attending day-care (MD 7.0 vs 0.90, $p = 0.02$)
Daniels *et al.*[18]	LUNA	Chronic pelvic pain	VAS score	Avoid LUNA in women with visible pathology compared with those without (3 months: MD 0.81 vs –0.08, $p = 0.01$, 6 months: MD 0.61 vs –0.28, $p = 0.01$, and 12 months: MD 0.21 vs –0.68, $p = 0.01$)
Jorgensen *et al.*[19]	Cervical insufficiency	Cerclage	Pregnancy loss or death before discharge	Avoid cerclage in multiples, not in singletons ($p = 0.03$)
			Baby healthy at discharge (all babies)	Avoid cerclage in multiples, not in singletons ($p = 0.04$)
Mbuagbaw *et al.*[20]	Text messaging	HIV	Adherence	Use text messaging in those with at least primary level of education not in those with lower level of education (ROR 1.51, $p < 0.05$)
Emberson *et al.*[21]	Alteplase	AIS	Good stroke outcome	Treat ASAP (OR 1.75 vs 1.26 vs 1.15 for ≤3 vs 3–4.5 vs >4.5 h, $p = 0.04$)
Solomon *et al.*[22]	Celecoxib	CVD	CV death, MI, stroke, HF or thromboembolic event	Treat low-risk individuals not high-risk individuals (HR 0.98 vs 1.28 vs 2.18 for low vs medium vs high baseline CV risk, $p = 0.02$)
[23]	ACE-inhibitor	AMI	Mortality	Avoid ACE-inhibitors in those aged ≥75, but not in others (OR 1.04 vs 0.89 vs 0.82 vs 0.94 for ≥75 vs 65–74 vs 55–64 vs <55, $p = 0.009$)
			Mortality	Treat those with a heart rate ≥100 bpm not those with bpm < 100 (OR 0.83 vs 0.89 vs 1.01 for ≥100 bpm, 80–100 bpm, <80 bpm, $p = 0.006$)
			Mortality	Treat those with an anterior infarct not those with an infarct at another location (OR 0.87 vs 0.98, $p = 0.01$)
Turnbull *et al.*[24]	ARB-based regimen	CVD	Stroke	Treat non-diabetics not diabetics (RR 0.96 vs 0.74, $p = 0.05$)
	ARB-based regimen		HF	Treat diabetics not non-diabetics (RR 0.70 vs 1.13, $p = 0.002$)
	More intensive treatment		MACE	Treat diabetics not non-diabetics (RR 0.75 vs 1.01, $p = 0.002$)
	ACE-inhibitor		CV death	Treat diabetics not non-diabetics (RR 0.67 vs 0.86, $p = 0.002$)
	More intensive treatment		CV death	Treat diabetics not non-diabetics (RR 0.67 vs 1.30, $p = 0.002$)
	ACE-inhibitor		Total mortality	Treat diabetics not non-diabetics (RR 0.76 vs 0.94, $p = 0.002$)
Bonati *et al.*[25]	Stenting	Symptomatic carotid stenosis	Any stroke or death within 120 days of randomization	Avoid stenting in patients ≥70 years old not in those <70 (RR 2.04 vs 1.00, $p = 0.02$)

(Continued)

Table 2. Continued

Study	Intervention	Disease	Outcome	Inferences and suggestions of authors based on treatment effects and *p*-value for subgroup–treatment interaction
Mihaylova *et al.*[26]	More statins	High cholesterol level	Major coronary event	Treat those with a baseline risk <10% (RR 0.57 vs 0.61 vs 0.77 vs 0.77 vs 0.78 for baseline risk <5 vs 5–10 vs 10–20 vs 20–30 vs ≥30%)
			Coronary vascularization	Treat those with a baseline risk <10% (RR 0.52 vs 0.63 vs 0.75 vs 0.79 vs 0.76 for baseline risk <5 vs 5–10 vs 10–20 vs 20–30 vs ≥30%)
			Major vascular event	Treat those with a baseline risk <10% (RR 0.62 vs 0.69 vs 0.79 vs 0.81 vs 0.79 for baseline risk <5 vs 5–10 vs 10–20 vs 20–30 vs ≥30%)
Fournier *et al.*[27]	Anti-depressants	Depression	Change in depression	Treat those with very severe baseline severity of depression not those with lower severity (SMD 0.47 vs 0.17 vs 0.11 for very severe vs severe vs mild-to-moderate, $p = 0.002$)
Greb *et al.*[28]	High-dose chemotherapy	Aggressive NHL	Overall survival	Treat poor-risk patients not good-risk patients (HR 0.95 vs 1.46, $p = 0.032$)
Hacke *et al.*[29]	Intravenous rt-PA	Stroke	3-month favourable outcomes	Treat ASAP (OR 2.81 vs 1.55 vs 1.40 vs 1.15 for 0–90 vs 91–180 vs 181–270 vs 271–360 min, $p = 0.005$)
Hills *et al.*[30]	GO + chemotherapy	AML	Survival	Treat those with favourable or intermediate not those with adverse characteristics (OR 0.50 vs 0.86 vs 1.03, $p = 0.008$)
			Survival	Treat those with favourable or intermediate not those with adverse characteristics (OR 0.47 vs 0.84 vs 0.99, $p = 0.006$)
Hlatky *et al.*[31]	CABG, PCI	Multivessel disease	Mortality	Treat diabetics not non-diabetics (HR 0.70 vs 0.98, $p = 0.014$)
			Mortality	Treat patients ≥65 years not those <55 or between 55 and 64 years (HR 0.82 vs 1.25 vs 0.90, $p = 0.002$)
Kotecha *et al.*[32]	β blocker	HF + AF	All-cause mortality	Avoid β blockers in AF patients not in those with sinus rhythm (HR 0.97 vs 0.73, $p = 0.002$)
[33]	PORT	NSCLC	Survival	Avoid PORT in patients with early-stage I or II completely resected NSCLC (HR 1.42 vs 1.26 vs 0.99 for 1 vs 2 vs 3, $p = 0.004$)
Quan *et al.*[34]	Pharmacologic treatment	Hypertension	Fatal CV event	Treat AA women, some White women may benefit (RR 0.57 vs 0.92, $p = 0.05$)
			Fatal and non-fatal CV event	Treat AA women, some White women may benefit (RR 0.54 vs 0.76, $p = 0.04$)
			All-cause mortality	Treat AA women, some White women may benefit (RR 0.59 vs 0.98, $p = 0.003$)

(Continued)

Table 2. Continued

Study	Intervention	Disease	Outcome	Inferences and suggestions of authors based on treatment effects and *p*-value for subgroup–treatment interaction
Rerkasem and Rothwell[35]	Endarterectomy	Symptomatic carotid stenosis	Any first stroke or operative death	Treat individuals with 70% or more stenosis without near occlusion not those with near occlusion (RR 0.53 vs 0.95, $p = 0.04$)
			Outcome[a]	Treat individuals with 70% or more stenosis without near occlusion not those with near occlusion (RR 0.40 vs 1.04, $p = 0.005$)
			Outcome[b]	Treat individuals with 70% or more stenosis without near occlusion not those with near occlusion (RR 0.40 vs 1.29, $p = 0.02$)
			Outcome[a]	Treat within 2 weeks of the last symptoms (RR 0.42 vs 0.57 vs 0.69 vs 0.96 for endarterectomy <2 vs 2–4 vs 4–12 vs >12 weeks, $p = 0.006$)
Rovers *et al.*[36]	Antibiotics	AOM	Pain, fever or both at 3–7 days	Treat children with bilateral AOM not without bilateral AOM (RR 0.72 vs 0.92, $p = 0.021$)
			Pain, fever or both at 3–7 days	Treat children <2 years old with bilateral AOM not in others (RR 0.64 vs 0.92 vs 0.84 vs 0.92, $p = 0.022$)
			Pain, fever or both at 3–7 days	Treat children with otorrhea not without otorrhea (RR 0.52 vs 0.80, $p = 0.039$)
Troughton *et al.*[37]	NP-guided treatment	CHF	Survival	Treat patients <75 years of age, not those ≥75 (HR 0.62 vs 0.98, $p = 0.028$)

MD, mean difference; SMD, standardized mean difference; OR, odds ratio; HR, hazard ratio; RR, risk ratio; ROR, relative odds ratio; *p*, *p*-value; OME, otitis media with effusion; VAS, visual analog scale; LUNA, laparoscopic uterosacral nerve ablation; HIV, human immunodeficiency virus; AIS, acute ischaemic stroke; ASAP, as soon as possible; CVD, cardiovascular disease; MI, myocardial infarction; HF, heart failure; AMI, acute myocardial infarction; MACE, major cardiovascular events; rt-PA, recombinant tissue plasminogen activator; GO, gemtuzumab ozogamicin; AML, acute myeloid leukaemia; ACE, angiotensin-converting enzyme; NHL, non-Hodgkin lymphoma; bpm, beats per minute; ARB, angiotensin receptor blocker; NIHSS, National Institutes of Health Stroke Scale; CABG, coronary artery bypass surgery; PCI, percutaneous coronary interventions; AF, atrial fibrillation; PORT, postoperative radiotherapy; NSCLC, non-small-cell lung cancer; AA, African-American; CV, cardiovascular; AOM, acute otitis media; CHF, chronic heart failure; NP, natriuretic peptide.

[a]Ipsilateral ischaemic stroke in the territory of the symptomatic carotid artery and any stroke or death within 30 days of trial surgery.

[b]Ipsilateral disabling or fatal ischaemic stroke in the territory of the symptomatic carotid artery and any disabling stroke or death that occurred within 30 days of trial surgery.

published group data. However, we should caution that subgroup analyses in a meta-analytic setting are mostly secondary, exploratory analyses, even when pre-specified in a protocol, because the trial data have already been collected and patterns of subgroup differences may have already been hinted upon, potentially influencing the type I and II errors of the analysis.[38] Currently, it is unclear whether any large and/or seemingly clinically meaningful subgroup differences in IPDMAs have resulted in initiation of new studies specifically focusing on verifying the subgroup effect. In the absence of prospective validation, incorporation of such stratified into clinical guidelines should be cautious. Currently, if anything, there is underutilization of overall evidence from systematic reviews and IPDMAs in clinical guidelines[39] and this is likely to be worse for evidence from subgroup analyses. IPDMAs may offer some additional advantages, e.g. more uniformity amongst outcome definitions and statistical analyses performed,[1–3] compared with AD meta-analyses. As such, sharing of IPD from randomized clinical trials—as suggested by several parties[40–42]—is still warranted and not only for the conduct of IPDMA, but also for sake of transparency and the opportunity of reproducibility.[43,44]

Based on the results of our survey, IPDMA may offer some opportunities for stratified medicine (based on different relative treatment effects) in the minority of clinical topics where such analyses are performed. We should also acknowledge that, even when the relative treatment effects are the same across subgroups, the absolute risks involved and thus also the absolute magnitude of the treatment effects (e.g. absolute risk difference) may differ greatly between subgroups. This may offer additional opportunities for stratified medicine. The yield of opportunities may also improve with the routine consideration of more variables, which depends on the collection and sharing of more rich datasets from randomized trials. Currently, many investigators who contribute their RCTs to IPDMAs provide only a very limited data set. Facilitation of data sharing may enhance the quality and range of IPDMAs in the future. More expanded use of primary raw data from trials makes meta-analyses more reliable.[45–47]

At the same time, all subgroup analyses need to be seen with caution, as they may not always reflect genuine differences. Claims for subgroup differences have been refuted repeatedly in single trials.[48–52] This may be true also for IPDMAs, although these meta-analyses operate with more data than single trials and thus may have less uncertain results. In contrast to single trials, subgroup differences in IPDMAs are derived from multiple studies and thus have some in-built validation. However, given the much larger power, some of the detected statistically significant signals may be of small magnitude and thus questionable clinical value. Regardless, some of this information may be genuine input to individualize treatment. Optimizing stratified treatment effects may actually depend on a combination of multiple subgrouping variables into risk scores. Currently available methods for improved individualized treatment decisions through therapeutic and prediction research evidence are probably under-utilized.[53] Therefore, it will be interesting to see whether involving multivariable prediction models into future subgroup analyses in IPDMA can improve the reliability of this information. Access to IPD allows validation and adjustment of existing prediction models or, if no such models are available, development of new ones.

Supplementary data

Supplementary data are available at *IJE* online.

Funding

This work was supported by the Netherlands Organisation for Scientific Research (project 825.14.001 to E.S.); the Laura and John Arnold Foundation (METRICS to J.P.A.I.); Sue and Bob O'Donnell (to J.P.A.I.). Canadian Institute for Health Research Doctoral Scholarship with a Michael Smith Foreign Study Supplement (to A.L.). The funding sources had no role in the study design, data collection, analysis, preparation of the manuscript or decision to submit the manuscript for publication.

All authors had full access to all of the data (including statistical reports and tables) in the study and can take responsibility for the integrity of the data and the accuracy of the data analysis.

Acknowledgements

J.P.A.I. and E.S. initiated the study. E.S., A.L. and J.P.A.I. designed the study. E.S. performed the initial literature search. E.S. and A.L. performed full-text assessment and data extraction. E.S. analysed the data in close collaboration with A.L. and J.P.A.I. Interpretation of data was performed by all authors. E.S. wrote the first draft. All authors edited consecutive versions of manuscript. All authors agreed with submission of the final manuscript and to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. E.S. is the guarantor.

Conflict of interest: All authors have completed the ICMJE uniform disclosure at www.icmje.org/coi_disclosure.pdf and declare: no support from any organization for the submitted work, no association to organizations that might have an interest in the submitted work, authors indicate no other relationships or activities that could appear to have influenced the submitted work.

References

1. Debray TP, Moons KG, van VG *et al*. Get real in individual participant data (IPD) meta-analysis: a review of the methodology. *Res Synth Methods* 2015;**6**:293–309.
2. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ* 2010; **340**:c221.

3. Stewart LA, Tierney JF. To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. *Eval Health Prof* 2002;**25**:76–97.
4. Huang Y, Tang J, Tam WW *et al.* Comparing the overall result and interaction in aggregate data meta-analysis and individual patient data meta-analysis. *Medicine (Baltimore)* 2016;**95**:e3312.
5. Williamson PR, Marson AG, Tudur C, Hutton JL, Chadwick D. Individual patient data meta-analysis of randomized anti-epileptic drug monotherapy trials. *J Eval Clin Pract* 2000;**6**:205–14.
6. Simmonds M, Stewart G, Stewart L. A decade of individual participant data meta-analyses: a review of current practice. *Contemp Clin Trials* 2015;**45**:76–83.
7. Liberati A, Altman DG, Tetzlaff J *et al*. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Med* 2009;**6**:e1000100.
8. Huang Y, Mao C, Yuan J *et al*. Distribution and epidemiological characteristics of published individual patient data meta-analyses. *PLoS One* 2014;**9**:e100151.
9. Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. *Stat Med* 2000;**19**:3127–31.
10. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. Mahwah, NJ: Lawrence Erlbaum Associates, 1988.
11. Goodacre S, Stevens JW, Pandor A *et al*. Prehospital noninvasive ventilation for acute respiratory failure: systematic review, network meta-analysis, and individual patient data meta-analysis. *Acad Emerg Med* 2014;**21**:960–70.
12. Valgimigli M, Sabate M, Kaiser C *et al*. Effects of cobalt-chromium everolimus eluting stents or bare metal stent on fatal and non-fatal cardiovascular events: patient level meta-analysis. *BMJ* 2014;**349**:g6427.
13. Boonacker CW, Rovers MM, Browning GG, Hoes AW, Schilder AG, Burton MJ. Adenoidectomy with or without grommets for children with otitis media: an individual patient data meta-analysis. *Health Technol Assess* 2014;**18**:1–118.
14. Adjuik M, Babiker A, Garner P, Olliaro P, Taylor W, White N. Artesunate combinations for treatment of malaria: meta-analysis. *Lancet* 2004;**363**:9–17.
15. Szczech LA, Berlin JA, Feldman HI, Anti-Lymphocyte Antibody Induction Therapy Study Group. The effect of antilymphocyte induction therapy on renal allograft survival: a meta-analysis of individual patient-level data. *Ann Intern Med* 1998;**128**:817–26.
16. Giallauria F, Vigorito C, Piepoli MF, Stewart Coats AJ. Effects of cardiac contractility modulation by non-excitatory electrical stimulation on exercise capacity and quality of life: an individual patient's data meta-analysis of randomized controlled trials. *Int J Cardiol* 2014;**175**:352–57.
17. Rovers MM, Black N, Browning GG, Maw R, Zielhuis GA, Haggard MP. Grommets in otitis media with effusion: an individual patient data meta-analysis. *Arch Dis Child* 2005;**90**:480–85.
18. Daniels JP, Middleton L, Xiong T *et al*. Individual patient data meta-analysis of randomized evidence to assess the effectiveness of laparoscopic uterosacral nerve ablation in chronic pelvic pain. *Hum Reprod Update* 2010;**16**:568–76.
19. Jorgensen AL, Alfirevic Z, Tudur Smith C, Williamson PR. Cervical stitch (cerclage) for preventing pregnancy loss: individual patient data meta-analysis. *BJOG* 2007;**114**:1460–76.
20. Mbuagbaw L, van der Kop ML, Lester RT *et al*. Mobile phone text messages for improving adherence to antiretroviral therapy (ART): an individual patient data meta-analysis of randomised trials. *BMJ Open* 2013;**3**:e003950.
21. Emberson J, Lees KR, Lyden P *et al*. Effect of treatment delay, age, and stroke severity on the effects of intravenous thrombolysis with alteplase for acute ischaemic stroke: a meta-analysis of individual patient data from randomised trials. *Lancet* 2014;**384**:1929–35.
22. Solomon SD, Wittes J, Finn PV *et al*. Cardiovascular risk of celecoxib in 6 randomized placebo-controlled trials: the cross trial safety analysis. *Circulation* 2008;**117**:2104–13.
23. ACE Inhibitor Myocardial Infarction Collaborative Group. Indications for ACE inhibitors in the early treatment of acute myocardial infarction: systematic overview of individual data from 100,000 patients in randomized trials. *Circulation* 1998;**97**:2202–12.
24. Turnbull F, Neal B, Algert C *et al*. Effects of different blood pressure-lowering regimens on major cardiovascular events in individuals with and without diabetes mellitus: results of prospectively designed overviews of randomized trials. *Arch Intern Med* 2005;**165**:1410–19.
25. Bonati LH, Dobson J, Algra A *et al*. Short-term outcome after stenting versus endarterectomy for symptomatic carotid stenosis: a preplanned meta-analysis of individual patient data. *Lancet* 2010;**376**:1062–73.
26. Mihaylova B, Emberson J, Blackwell L *et al*. The effects of lowering LDL cholesterol with statin therapy in people at low risk of vascular disease: meta-analysis of individual data from 27 randomised trials. *Lancet* 2012;**380**:581–90.
27. Fournier JC, DeRubeis RJ, Hollon SD *et al*. Antidepressant drug effects and depression severity: a patient-level meta-analysis. *JAMA* 2010;**303**:47–53.
28. Greb A, Bohlius J, Schiefer D, Schwarzer G, Schulz H, Engert A. High-dose chemotherapy with autologous stem cell transplantation in the first line treatment of aggressive non-Hodgkin lymphoma (NHL) in adults. *Cochrane Database Syst Rev* 2008;**1**:CD004024.
29. Hacke W, Donnan G, Fieschi C *et al*. Association of outcome with early stroke treatment: pooled analysis of ATLANTIS, ECASS, and NINDS rt-PA stroke trials. *Lancet* 2004;**363**:768–74.
30. Hills RK, Castaigne S, Appelbaum FR *et al*. Addition of gemtuzumab ozogamicin to induction chemotherapy in adult patients with acute myeloid leukaemia: a meta-analysis of individual patient data from randomised controlled trials. *Lancet Oncol* 2014;**15**:986–96.
31. Hlatky MA, Boothroyd DB, Bravata DM *et al*. Coronary artery bypass surgery compared with percutaneous coronary interventions for multivessel disease: a collaborative analysis of individual patient data from ten randomised trials. *Lancet* 2009;**373**:1190–97.
32. Kotecha D, Holmes J, Krum H *et al*. Efficacy of beta blockers in patients with heart failure plus atrial fibrillation: an individual-patient data meta-analysis. *Lancet* 2014;**384**:2235–43.
33. PORT Meta-analysis Trialists Group. Postoperative radiotherapy for non-small cell lung cancer. *Cochrane Database Syst Rev* 2005;**2**:CD002142.
34. Quan A, Kerlikowske K, Gueyffier F, Boissel JP. Efficacy of treating hypertension in women. *J Gen Intern Med* 1999;**14**:718–29.
35. Rerkasem K, Rothwell PM. Carotid endarterectomy for symptomatic carotid stenosis. *Cochrane Database Syst Rev* 2011;**4**:CD001081.

36. Rovers MM, Glasziou P, Appelman CL *et al*. Antibiotics for acute otitis media: a meta-analysis with individual patient data. *Lancet* 2006;**368**:1429–35.
37. Troughton RW, Frampton CM, Brunner-La Rocca HP *et al*. Effect of B-type natriuretic peptide-guided treatment of chronic heart failure on total mortality and hospitalization: an individual patient meta-analysis. *EurHeart J* 2014;**35**:1559–67.
38. Schuit E, Roes KC, Mol BW, Kwee A, Moons KG, Groenwold RH. Meta-analyses triggered by previous (false-)significant findings: problems and solutions. *Syst Rev* 2015;**4**:57.
39. Vale CL, Rydzewska LH, Rovers MM, Emberson JR, Gueyffier F, Stewart LA. Uptake of systematic reviews and meta-analyses based on individual participant data in clinical practice guidelines: descriptive study. *BMJ* 2015;**350**:h1088.
40. NIH Data Sharing Policy and Implementation Guidance. 2003. https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm (15 June 2017, date last accessed).
41. MRC Policy and Guidance on Sharing of Research Data from Population and Patient Studies. 2011. https://mrc.ukri.org/Utilities/Documentrecord/index.htm?d=MRC008302 (15 June 2017, date last accessed).
42. Drazen JM. Sharing individual patient data from clinical trials. *N Engl J Med* 2015;**372**:201–02.
43. Nevitt SJ, Marson AG, Davie B, Reynolds S, Williams L, Smith CT. Exploring changes over time and characteristics associated with data retrieval across individual participant data meta-analyses: systematic review. *BMJ* 2017;**357**:j1390.
44. Taichman DB, Sahni P, Pinborg A *et al*. Data sharing statements for clinical trials. *BMJ* 2017;**357**:j2372.
45. Ioannidis JP. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *Milbank Q* 2016;**94**:485–514.
46. Ioannidis JP. Meta-analyses can be credible and useful: a new standard. *JAMA Psychiatry* 2017;**74**:311–12.
47. Ioannidis JP. Next-generation systematic reviews: prospective meta-analysis, individual-level data, networks and umbrella reviews. *Br J Sports Med* 2017;**51**:1456–58.
48. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet* 2000;**355**:1064–69.
49. Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. *JAMA* 2007;**298**:1209–12.
50. Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet* 2005;**365**:176–86.
51. Wallach JD, Sullivan PG, Trepanowski JF, Sainani KL, Steyerberg EW, Ioannidis JP. Evaluation of evidence of statistical support and corroboration of subgroup claims in randomized clinical trials. *JAMA Intern Med* 2017;**177**:554–60.
52. Wallach JD, Sullivan PG, Trepanowski JF, Steyerberg EW, Ioannidis JP. Sex based subgroup differences in randomized controlled trials: empirical evidence from Cochrane meta-analyses. *BMJ* 2016;**355**:i5826.
53. Varadhan R, Segal JB, Boyd CM, Wu AW, Weiss CO. A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research. *J Clin Epidemiol* 2013;**66**:818–25.