BMJ Open

Evaluation of person-level heterogeneity of treatment effects in published multiperson N-of-1 studies: systematic review and reanalysis

Gowri Raman,[1] Ethan M Balk,[2] Lana Lai,[3] Jennifer Shi,[4] Jeffrey Chan,[5] Jennifer S Lutz,[3] Robert W Dubois,[6] Richard L Kravitz,[7] David M Kent[3]

To cite: Raman G, Balk EM, Lai L, *et al*. Evaluation of person-level heterogeneity of treatment effects in published multiperson N-of-1 studies: systematic review and reanalysis. *BMJ Open* 2018;**8**:e017641. doi:10.1136/bmjopen-2017-017641

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2017-017641).

Received 8 May 2017
Revised 22 February 2018
Accepted 13 April 2018

For numbered affiliations see end of article.

Correspondence to
Dr David M Kent;
dkent1@tuftsmedicalcenter.org

ABSTRACT

Objective Individual patients with the same condition may respond differently to similar treatments. Our aim is to summarise the reporting of person-level heterogeneity of treatment effects (HTE) in multiperson N-of-1 studies and to examine the evidence for person-level HTE through reanalysis.

Study design Systematic review and reanalysis of multiperson N-of-1 studies.

Data sources Medline, Cochrane Controlled Trials, EMBASE, Web of Science and review of references through August 2017 for N-of-1 studies published in English.

Study selection N-of-1 studies of pharmacological interventions with at least two subjects.

Data synthesis Citation screening and data extractions were performed in duplicate. We performed statistical reanalysis testing for person-level HTE on all studies presenting person-level data.

Results We identified 62 multiperson N-of-1 studies with at least two subjects. Statistical tests examining HTE were described in only 13 (21%), of which only two (3%) tested person-level HTE. Only 25 studies (40%) provided person-level data sufficient to reanalyse person-level HTE. Reanalysis using a fixed effect linear model identified statistically significant person-level HTE in 8 of the 13 studies (62%) reporting person-level treatment effects and in 8 of the 14 studies (57%) reporting person-level outcomes.

Conclusions Our analysis suggests that person-level HTE is common and often substantial. Reviewed studies had incomplete information on person-level treatment effects and their variation. Improved assessment and reporting of person-level treatment effects in multiperson N-of-1 studies are needed.

Strengths and limitations of this study

- ► Our analysis suggests that person-level heterogeneity of treatment effects (HTE) is common and often substantial.
- ► Our analysis was limited by the paucity of N-of-1 studies in the literature and by the low statistical power in the available studies.
- ► Multiperson N-of-1 studies are the best design to estimate individual patient treatment effects and compare the variation in effects between individuals to variation within individuals across different periods.

INTRODUCTION

Clinicians commonly observe that individual patients given the same treatment for the same condition appear to respond differently from one another. This observation, combined with our understanding of the complex mechanisms of diseases and therapies and the potential importance of myriad patient-specific factors (eg, age, sex, illness severity, comorbidities, co-treatments and molecular differences influencing pharmacokinetics and dynamics), has led to a widely held assumption that the observed variation in treatment response seen between individuals is not merely random, but stable and potentially predictable. This assumption underpins the field of personalised medicine, which aims to determine the best treatment for an individual patient, as opposed to treating all patients with the intervention found to be most effective for the 'average' patient.

Nevertheless, statistical analyses aimed at discovering heterogeneity of treatment effects (HTE) among groups of individuals (eg, subgroup analyses of parallel arm randomised trials) typically fail to find compelling and reliable evidence for the presence of such heterogeneity. For example, statistically significant differences in treatment effects between men and women are often reported, but a systematic review indicates that the frequency of these interactions across studies suggests that the vast majority occur by chance.[1] Similarly, the field of pharmacogenetics, also built on the assumption of stable variation in treatment responses, has

largely failed to live up to its promise to broadly improve the targeting of drugs—particularly outside the special case of oncology (where studies generally depend on the subclassification of tumour tissue not on variation in germ line polymorphisms).[2 3] This failure to find reproducible HTE has supported the contrarian notion that true individual effects may be a 'myth', an overinterpretation of random noise.[4]

To distinguish between these two possibilities, Kalow *et al*[5] have suggested that carefully designed series of N-of-1 studies could be performed for those chronic conditions amenable to this design (ie, where the disease process is relatively stable over time, treatment effects are transient and outcomes vary and are observable over time). By estimating individual patient treatment effects and comparing the variation in effects *between* individuals to variation *within* individuals across different periods, it is possible to determine the non-random component of heterogeneity in individual treatment effects—even if one is unable to identify the variables that predict this variation (ie, even in the absence of group-level HTE, such as men vs women or old vs young).

A recent review summarised N-of-1 studies reported in the literature—including multiperson N-of-1 studies—but did not examine whether and how these studies provide information on person-level HTE. Therefore, our objectives are (1) to summarise the conduct and reporting of assessments of variation in person-level treatment effects from N-of-1 studies and (2) to extract, reanalyse and report the results from the subset of studies that provided adequate data in their published reports to examine the extent of the evidence for person-level HTE (ie, participant-level outcomes or effects).[6]

METHODS

This review was conducted in accordance with the highest standards for conducting systematic reviews.[7 8] We defined N-of-1 studies as crossover trials in which each patient receives two or more treatments in a predefined, often randomised, sequence.

Data sources and searches

We used two separate searches because N-of-1 studies can be indexed differently: (1) a search in Medline, Cochrane Central and EMBASE using terms related to repeated crossover studies (for publications indexed from inception to 17 August 2017) and (2) a Medline, Cochrane Central, EMBASE and Web of Science search using terms that are related to N-of-1 (for publications indexed from 2011 to 17 August 2017). For N-of-1 studies indexed before 2011, we used studies included in a prior published systematic review by Gabler *et al*.[6] Our searches combined terms and Medical Subject Headings for N-of-1, single-subject, single-patient, randomised trials, crossover, multiperiod crossover and rotated or repeated period crossover (see online Supplementary appendix tables 1 and 2 for detailed search terms). The searches were not restricted by disease, condition, organ system or treatment.

Study selection

We selected eligible multiperson N-of-1 studies to describe the frequency of reporting of individual outcomes and effects and of documented HTE in these studies. We required a minimum of two individual subjects per study for evaluation of HTE. We excluded studies that included non-pharmacological interventions, reviews, abstracts and protocols. We included studies with placebo or 'no treatment' interventions. Citations were double screened by reviewers using an open-source, online software Abstrackr (http://abstrackr.cebm.brown.edu/). Full-text articles of potentially relevant studies were again double screened for eligibility.

Person-level outcomes were defined as outcomes for each person at each point in time when they were measured, reported in tables, text or graphs. Person-level treatment effect was defined as contrasts of outcomes in individuals on one treatment versus the comparator. Person-level HTE was defined as quantified variation in the person-level treatment effects, whereas HTE more broadly includes any type of subgroup analysis (eg, males vs females; older vs younger) as outlined in figure 1.

Data extraction and quality assessment

One of the four reviewers extracted data from each publication; a second reviewer verified all numerical information and basic descriptors of the study design and analysis. Operational definitions for extraction items were discussed in weekly project meetings and discrepancies between extractors were resolved by consensus with senior authors (DK, GR, EB). From each study, we extracted bibliographic information, details related to study design (number of patients enrolled, selection criteria, interventions evaluated, randomisation methods, outcomes assessed, follow-up duration), information on patient characteristics and person-level measurements of outcomes or estimates of person-level treatment effects (with corresponding measures of their uncertainty). When necessary, we extracted data by digitising the graphs and the values were estimated using Engauge Digitizer V.2.14 (http://digitizer.sourceforge.net/). We assessed the methodological quality of each study based on predefined criteria, in accordance with the Agency for Healthcare Research and Quality suggested methods and the Cochrane risk of bias for clinical trials.[9 10]

We generated graphs showing the trajectory of response for each patient in each study and compared them against the published information. We also generated scatterplots of measurements over time for studies that did not present their data in graphical format to help us identify aberrant data points (eg, errors in data extraction). We verified potentially aberrant data points by re-examining the published data and made corrections, when needed.

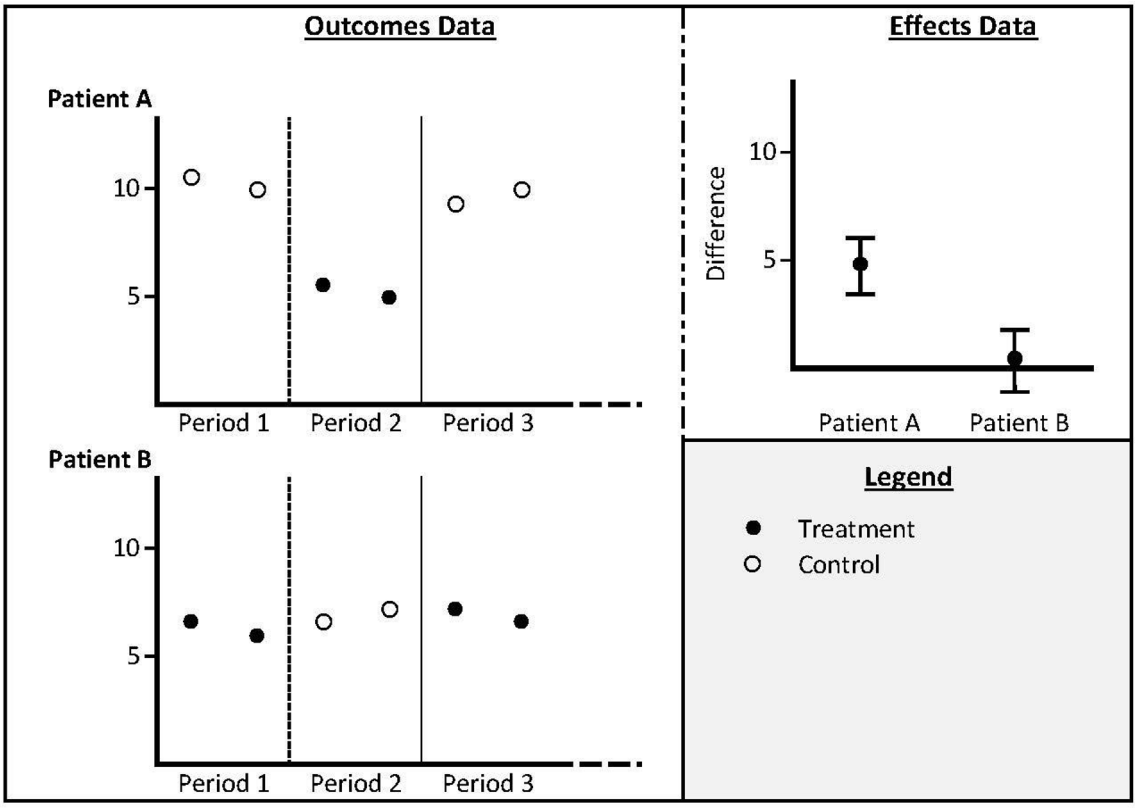

Figure 1 Schematic description of person-level outcomes (outcomes for each patient during each treatment period); person-level effects (contrasts of the outcomes for each patient in one treatment condition vs another) and person-heterogeneity of treatment effects (between patient contrasts of effects).

Data synthesis and analyses

We examined the degree to which studies reported person-level data. This was described using the following items for each reported outcome: (1) qualitative descriptions of HTE (eg, 'there were eight responders and four non-responders'); (2) details of person-level outcomes (ie, outcomes with each treatment within each period); (3) details of person-level treatment effect (ie, a point estimate of contrasts of outcomes in individuals on one treatment vs the comparator); (4) reporting of person-level statistical effect estimate (eg, SD, exact p values or CIs for treatment effects within individuals); (5) description of statistical tests examining HTE (ie, tests evaluating the contrast of treatment effects between individuals or groups in the study) and (6) claims of HTE. Note that qualitative descriptions of HTE for item 1 would include any description that implied that treatment effects varied, whereas item six required a more definite study conclusion (eg, 'our results demonstrate significant variation across individuals in response to treatment X'), whether or not these conclusions were based on robust statistical tests.

Statistical HTE analysis of extracted study results

We performed statistical analysis testing for person-level HTE on all studies presenting person-level data. We used a consistent analytic strategy across studies, to the extent permitted by the reporting in published papers. Our strategy was different for studies that reported person-level outcome measurements and those that reported estimates of person-level treatment effects with their sampling variances (or adequate information to approximately calculate these statistics).

For studies that only reported (or allowed the calculation of) *estimates of person-level treatment effects*, we obtained an average effect using a fixed effect inverse variance model and estimated the variance of the person-level treatment effects using DerSimonian and Laird method of moments estimator.[11 12] In addition to a fixed effect model, we also obtained an average effect using a random-effects model. Finally, we tested the hypothesis that all person-level treatment effects were equal using Cochran's χ^2 test and quantified the proportion of observed variation due to 'true' person-level effect heterogeneity with the I^2 statistic.[13]

For studies that reported *person-level outcomes*, we developed a linear model (for continuous outcomes) or generalised linear model (for binary or count outcomes) using the outcome of interest as the response, the intervention(s) as a covariate and indicator variables for different study participants.[14] This model estimates a common treatment effect across participants. We also derived a similar model with treatment-by-participant interactions. This model allows each patient to have a different effect. The statistical significance of person-level HTE was assessed by a likelihood ratio test comparing the two

models. In addition to a fixed effect model, we also fit a hierarchical linear or generalised linear mixed model with a random intercept and a random slope (for the treatment effect) to estimate the average treatment effect across all patients (assuming person-level HTE). We tested the hypothesis that all person-level treatment effects were equal and quantified the proportion of observed variation due to 'true' person-level effect heterogeneity with the I^2 statistic.[13] For modelling within-patient variance, we used a common variance with an uncorrelated covariance structure, as was used in a prior N-of-1 study.[14] Person-level treatment effect was assumed to be equal across time periods. For the treatment effect, we used more than one random slope when more than two treatments were compared.

Patient and public involvement

Patients and the public were not involved in the design or analysis of this study.

RESULTS

The searches for repeated crossover studies identified 11 891 citations and those for N-of-1 studies identified 3819 citations (indexed from 2011 onwards). Of these, we retrieved 407 full-text articles for review plus 100 N-of-1 trial articles (indexed before 2011) from an existing systematic review.[5] On full-text screening, 62 studies (58 multiperson N-of-1 studies and four repeated period crossover studies) met eligibility criteria (online supplementary appendix tables 3) and are reported multiperson N-of-1 studies throughout the article. An outline of the search and study selection flow is provided in figure 2.

Description of studies

Table 1 summarises the 62 multiperson N-of-1 studies that were published between 1986 and 2017 reporting a total of 1974 patients. The most common clinical domains in the multiperson N-of-1 studies were neurology (16%), arthritis/rheumatology (10%) and psychiatry (9%). Most studies were described as 'double blind' but details about the methods for blinding were often unclear; similarly studies often provided unclear information about the generation of the randomisation sequence and allocation concealment (online supplementary appendix tables 4). Among the studies, 93% compared a pair of treatment strategies, 5% compared three strategies and 2% compared four strategies. Studies had between three and 16 treatment periods and obtained an average

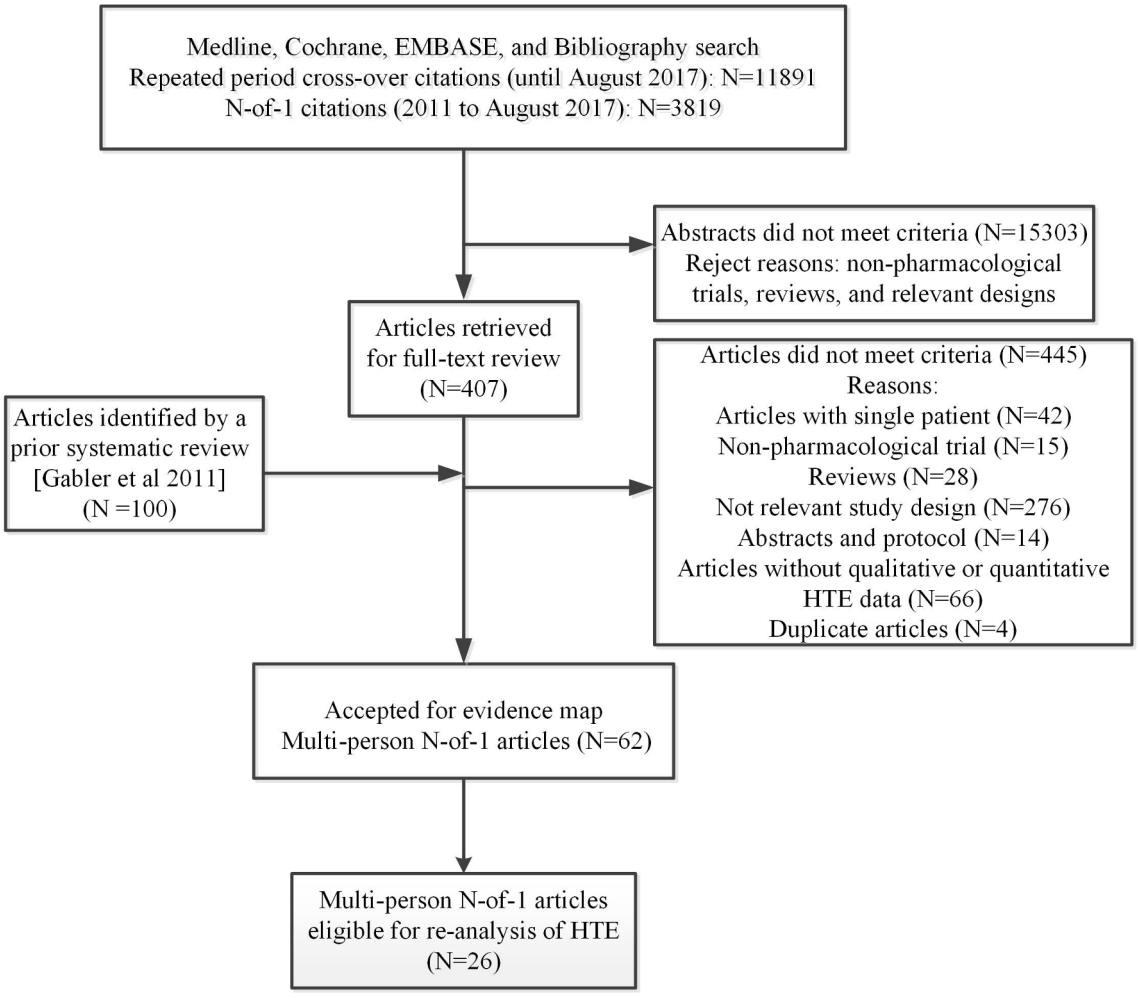

Figure 2 Study flow diagram represents the flow of eligible studies included in this review.

Table 1 Evidence map of multiperson N-of-1 and repeated period crossover studies

Description	Multi-person N-of-1 studies (n=62)
Publication years	1979–2017
Subjects	Total N (median, IQR)
Enrolled	2153 (16, 9–42)
Completed	1705 (12, 7–32)
Intervention and comparisons	
Head-to-head active drugs	10
Placebo	47
Active drug and placebo	1
Population	
Paediatric	12
Adults	50
Major systems studied	
Arthritis/rheumatology	10
Cardiovascular	3
Gastrointestinal	7
Hypertension	1
Psychiatry	9
Neurology	16
Respiratory	9
Miscellaneous*	7
Top 5 disease conditions	
ADHD	6
Angina	3
Chronic pain	5
GORD	5
Obstructive airway	6
Osteoarthritis	6

*Sleep disorders, allergy, cancer, muscular, vascular (for multiperson N-of-1); pain, urology, GYN, Heme/Onc, allergy, dermatology, drug abuse, endocrine, lipids, nephrology, ophthalmology, respiratory (for repeated crossover studies).
ADHD, attention-deficit hyperactivity disorder; GORD, gastro-oesophageal regurgitation disorder; n, number of participants.

of 1–42 outcome measurements per period. Across reported outcomes, 89% of the assessed outcomes were patient reported and 11% were investigator assessed.

Reporting person-level outcomes, effects and HTE

While most studies (92%) had some qualitative acknowledgement that the treatment effects appeared to vary across individuals, formal reporting at the participant level was variable (table 2). Person-level outcomes under each treatment were reported in 52% of multiperson N-of-1 studies. Person-level treatment effects with quantitative data (comparing outcomes on each treatment) for each individual who completed the trial was available in 32%; and details on the statistical evaluation of these effects (as SD or exact pvalues or confidence intervals) were available in 13 (21%) multiperson N-of-1 studies. Only five (8%) studies described statistical tests examining any HTE. However, only two studies (3%) reported person-level HTE, whereas the others examined group-level HTE using conventional subgroup analysis based on observable characteristics.

Table 2 Survey of HTE assessment in multiperson N-of-1 studies

HTE reporting	Multiperson N-of-1 studies (n=62)
Qualitative description	92%
Person-level outcomes	52%
Person-level treatment effects	32%
Statistical analysis of person-level effects (eg, p values)	21%
Any statistical test for HTE	8%*
Claims of heterogeneity	15%

*Only two studies reported person-level HTE, the remaining three studies reported group level effect.
HTE, heterogeneity of treatment effects.

Reanalysis of person-level data

Of the 62 studies, there were 36 studies that provided person-level data, either as outcomes in each treatment period or as person-level treatment effects (table 3). Of these, only 25 studies provided person-level data sufficient to support re-analysis: 14 studies provided person-level outcomes; 13 studies provided person-level treatment effects (two studies provided both). The remaining 11 studies reported either medians or means without data on variance or did not provide sufficient information on completers, so they could not be reanalysed for treatment effect or HTE.

Of 13 studies (with 27 unique comparisons) that reported analysable person-level treatment effect data (table 3), 10 studies had a placebo comparator and three studies had an active comparator. The sample size ranged from 7 to 68; average crossover periods ranged from 6 to 16 days and average outcome measures per period ranged from 1 to 21. The average treatment duration ranged from 14 to 336 days.

There were 14 studies (with 27 unique comparisons) that reported analysable person-level outcome data (table 3), including two studies also reporting person-level treatment effects. Of these, 11 compared the intervention with placebo and three studies compared two active interventions. The sample size ranged from 2 to 22; the average number of crossover periods ranged from 3 to 10 and the average number of outcome measures per period ranged from 1 to 42. The average treatment duration ranged from 9 to 210 days.

Table 3 Characteristics of studies reporting person-level data

Author, Year	Disease	Number enrolled (analysed)	Intervention	Comparator	Cross-over periods	Total intervention duration	Outcome measures per period
Studies with reanalysable person-level outcomes							
Camfield, 1996	Mental retardation with fragmented sleep	6 (6)	Melatonin	Placebo	7	10 weeks	14
Hinderer, 1990	Traumatic spinal cord injury	5 (5)	Baclofen	Placebo	3	9 weeks	2
Langer, 1993	Gastro-oesophageal reflux	2 (2)	Cisapride	Placebo	3	6 weeks	5
Lashner, 1990	Ulcerative colitis	7 (6)	Nicotine	Placebo	4	8 weeks	1
Maier, 1994	Chronic depression	10 (9)	Sulpiride	Placebo	4	28 weeks	42
Mandelcorn, 2004	Brain injury	4 (4)	Ondansetron	Placebo	4	5 weeks	1
McQuay, 1994	Neuropathic pain	19 (19)	Dextromethorphan	Placebo	5	20 days	1
Miyazaki, 1995	Unstable angina	22 (22)	Isosorbide dinitrate	Isosorbide dinitrate: intermittent injection	3	9 days	6
Nathan, 2006	Paediatric brain tumour	12 (7)	Ondansetron and metopimazine	Ondansetron and placebo	Unclear	189 days	Unclear
Parodi, 1979	Unstable angina	12 (12)	Verapamil	Placebo	4	10 days	Unclear
Parodi, 1986	Unstable angina	10 (10)	Verapamil	Propranolol, placebo	8	18 days	Unclear
Tison, 2012	Levodopa-induced dyskinesia in patients with Parkinson's disease	10 (10)	Simvastatin	Placebo	6	96 days	1
Studies with re-analyzable person-level treatment effects							
Emmanuel, 2012	Chronic intestinal pseudo-obstruction	7 (4)	Prucalopride	Placebo	16	48 weeks	21
Haas, 2004	Chronic tension type and migraine headache	39 (16)	Dextroamphetamine	Equi-stimulatory caffeine	8	20 days	20
Jaeschke, 1991	Fibromyalgia	22 (23)	Amitriptyline	Placebo	6	12 weeks	2
Johannessen, 1992	Dyspepsia	68 (46)	Cimetidine	Placebo	12	184 days	15
Lipka, 2017	Autoimmune myasthenia gravis	4 (4)	Ephedrine	Placebo	4	6 weeks	1
Mahon, 1996	Irreversible chronic airflow limitation	16 (14)	Theophylline	Placebo	8	73 days	1
March, 1994	Osteoarthritis	25 (15)	Diclofenac	Paracetamol	6	12 weeks	14
Patel, 1991	Non-reversible chronic airflow limitation	26 (18)	Ipratropium bromide/ theophylline /salbutamol /beclomethasone	Placebo	6	6 weeks	Unclear
Wallace, 1994	Attention deficit hyperactivity disorder	11 (7)	Methylphenidate	Placebo	14	14 days	1

Continued

Table 3 Continued

Author, Year	Disease	Number enrolled (analysed)	Intervention	Comparator	Cross-over periods	Total intervention duration	Outcome measures per period
Woodfield, 2005	Skeletal muscle cramps	13	Quinine	Placebo	6	14 weeks	2
Zucker, 2006	Fibromyalgia	58	Amitriptyline and placebo	Amitriptyline and fluoxetine combination	6	36 weeks	1
Study with both person-level data							
Pereira, 1995	Atrial fibrillation/ deep venous thrombosis	7	Generic warfarin	Coumadin	10	30 weeks	2
Joy, 2014	Statin-related myalgia	8 (7)	Statin	Placebo	6	33 weeks	3
Study with insufficiently reported person-level data							
Person-level outcome data							
Denburg, 1994	Systemic lupus erythematosus	10	Prednisone	Placebo	6	30 weeks	1
Mitchel, 2015	Fatigue in advanced cancer	43 (33)	Methylphenidate	Placebo	6	18 days	6
Nikles, 2000	Osteoarthritis	14	Ibuprofen	Paracetamol; placebo	6	12 weeks	14
Nikles, 2015	Dry mouth in advanced cancer	17 (4)	Pilocarpine	Placebo	6	18 days	6
Nikles, 2017	Acquired brain injury	53 (38)	Nervous system stimulants	Placebo	6	18 days	6
Reitberg, 2002	Allergic rhinitis	36	Loratadine and chlorpheniramine maleate	loratadine with placebo	8	32 days	4
Sheather-Reid, 1998	Chronic pain	8	Ibuprofen/codeine	Placebo	6	12 weeks	14
Person-level treatment effects							
Huber, 2007	Juvenile idiopathic arthritis	6	Amitriptyline	Placebo	6	17 weeks	12
Privitera, 1994	Partial seizure	16	Dezinamide	Placebo	6	35 weeks	6
Wegman, 2003	Osteoarthritis	13	Paracetamol	Non-steroidal anti-inflammatory drugs	10	20 weeks	14
Wegman, 2005	Regular temazepam users	15	Temazepam	Placebo	10	10 weeks	7

Reanalysis of studies reporting estimates of person-level treatment effects

Thirteen studies (including 27 comparisons, due to multiple outcomes in some studies) reported estimates of person-level treatment effects sufficient to analyse (online supplementary appendix figures 1–16 display graphs of the person-level treatment effect data). Average fixed effect estimates for each analysis are shown in table 4; random-effects estimates were generally similar (online supplementary appendix tables 5). In 8 of the 13 studies (62%) and 15 of the 27 total unique comparisons (56%), we found evidence of statistically significant HTE for at least one outcome (table 4). Generally, the magnitude in the variation of individual patient effects (as seen in the range) was very large compared with the average effects. Most studies (64%) showed person-level effects that differed qualitatively from one another. Most of the variation in the observed individual effects was attributable to 'true' (non-random) heterogeneity of person-level effects; 11 of 27 analyses had I^2 >80%.

Table 4 Analysis results of studies reporting person-level treatment effects

			Main effect	Person-level heterogeneity of treatment effect (HTE)		
Author, year	Outcome	Range of the scales (severity)	Treatment effect (CI)	P for HTE*	Treatment effect range	I^2 % (CI)
Emmanuel, 2012	Bloating	0–4 (0=absent to 4=worst)	−0.344 (−0.619 to −0.069)	<0.001	−1.1 to −0.1	94 (88 to 97)
	Pain	0–4 (0=absent to 4=worst)	−0.440 (−0.771 to −0.110)	<0.001	−0.2 to −1.4	96 (92 to 98)
Haas, 2004	Chronic tension-type headache grade	0–3 (0=none to 3=severe)	0.772 (0.454 to 1.090)	<0.001	0.04 to 1.9	84 (76 to 90)
	Chronic migraine headache grade	0–3 (0=none to 3=severe)	0.542 (0.354 to 0.731)	0.067	0.2 to 0.83	37 (0 to 65)
Jaeschke, 1991	Seven-point symptom scale	1–7 (higher scores represent better function)	0.427 (0.210 to 0.645)	<0.001	−1.02 to 3.18	85 (79 to 89)
	Tender point changes count	Number of tender points	1.320 (0.404 to 2.236)	<0.001	−4.33 to 9.0	72 (57 to 82)
Johannessen, 1992	Six-point symptom scale	0–6 (0=NR to 6=NR)	0.698 (0.466 to 0.931)	<0.001	−1.67 to 3.17	66 (53 to 75)
Joy, 2014	VAS myalgia score	0–100 mm (0=none to 100=worst)	0.119 (−2.283 to 2.521)	0.996	−8.10 to 9.45	0 (0 to 68)
	Symptom-specific VAS	0–100 mm (0=none to 100=worst)	1.937 (0.179 to 3.696)	0.797	−8.0 to 18.05	0 (0 to 68)
	Pain severity score	0–10 (0=none to 10=worst)	0.086 (−0.215 to 0.387)	0.986	0.0 to 1.0	0 (0 to 68)
	Pain interference score	0–10 (0=none to 10=worst)	−0.016 (−0.095 to 0.064)	0.917	−0.02 to 0.75	0 (0 to 68)
Lipka, 2017	Quantitative myasthenia gravis score	0–3 (0=none to 3=severe)	1.006 (0.215 to 1.797)	0.803	0.67 to 1.67	0 (0 to 85)
	Myasthenia gravis (MG) composite	0–50	2.891 (0.348 to 5.433)	0.177	−1.05 to 5.12	39 (0 to 80)
	MG-activities of daily living	0–24	1.099 (−0.277 to 2.474)	0.047	0.03 to 3.0	62 (0 to 87)
	VAS score	0–10 (0=none to 100=worst)	1.275 (−0.115 to 2.665)	0.190	−0.01 to 3.02	37 (0 to 78)
Mahon, 1996	Dyspnoea in Likert Scale	1–7 (1=extremely short of breath to 7=no shortness)	0.125 (−0.181 to 0.430)	<0.001	−0.57 to 0.89	78 (58 to 88)
March, 1994	Mean pain score on VAS	5 point Likert scale (0–100 mm)	−7.093 (−11.939 to −2.248)	<0.001	−33.8 to 4.1	98 (97 to 98)
	Mean stiffness score on VAS	5 point Likert scale (0–100 mm)	−5.992 (−11.280 to −0.704)	<0.001	−36 to 10.7	97 (96 to 98)
Patel, 1991†	Four-item symptom questionnaire (all compared with placebo)	1–7 (1=extremely short of breath to 7=no shortness of breath)	0.340 (0.253 to 0.422)	<0.001	−0.34 to 3.1	91 (87 to 94)
	Four-item symptom questionnaire (use of ipratropium bromide)		0.675 (0.264 to 1.085)	<0.001	−0.22 to 3.1	87 (78 to 92)
	Four-item symptom questionnaire (use of salbutamol)		0.865 (0.042 to 1.687)	<0.001	0.46 to 1.3	94 (NA)
	Four-item symptom questionnaire (use of theophylline)		0.025 (−0.434 to 0.484)	0.172	−0.34 to 0.18	30 (0 to 93)
Pereira, 1995	INR (diff)	Target INR range of 2.0–3.0	0.027 (−0.155 to 0.209)	0.477	−0.28 to 0.37	0 (0 to 75)
Wallace, 1994	Conners 15-item rating scale scores	0–3 (NR)	0.759 (0.341 to 1.178)	0.747	0.42 to 1.22	0 (0 to 79)
Woodfield, 2005	Changes in the number of cramps	Number—mean difference	−18.823 (−28.527 to −9.120)	<0.001	−77 to −2	92 (87 to 95)
	Total days with cramps	days	−6.181 (−9.798 to −2.563)	<0.001	−13 to −1	94 (90 to 96)
Zucker, 2006	FIQ	0–100 (0=best to 100=worst)	−5.019 (−8.784 to −1.254)	0.999	−32.0 to 0.98	0 (0 to 37)

*The significance of person-level HTE was assessed by Cochran's X^2-based test.
†One subject had beclomethasone.
FIQ, Fibromyalgia Impact Questionnaire; INR, international normalised ratio; NA, not applicable; NR, not reported; VAS, Visual Analogue Scale.

Reanalysis of studies reporting person-level outcome measurements

Because some of the 14 studies providing analysable outcome data had multiple outcomes (or multiple outcomes scales), there were a total of 27 comparisons with analysable data. (The online supplementary appendix figures 17–42 displays graphs of the person level outcome results.) Average fixed effect estimates for each analysis are shown in table 5; random effects estimates were generally similar (online supplementary appendix tables 6). In eight of the 14 studies (57%) (17 of the 27 unique comparisons (63%)), there was statistically significant person-level HTE for at least one outcome. Again, the variation in individual effects was often large compared

Table 5 Studies reporting person-level outcomes

Author, year	Outcome	Definition/range of the scales (severity)	Main effect	Person-level heterogeneity of treatment effect (HTE)		
			Fixed treatment effect	P for person treatment interaction*	Treatment effect range (lower range (CI)–upper range (CI))	I^2 % (CI)
Camfield, 1996	Nights without awakening	Between 10:00 PM and 7:00 AM per day	0.865 (0.215 to 1.516)	0.456	0.12 to 2.0	0 (0 to 79)
Hinderer, 1990	Anxiety	Beck Inventory-A anxiety scale 0–3 (0=never, 3=almost all the time)	0.000 (0.000 to 0.000)	<0.001	−6.38 to 0.000	91 (81 to 95)
Joy, 2014	Myalgia score	Visual Analogue Score for myalgia (0=none to 100=worst)	3.3812 (−2.668 to 9.430)	0.565	−11.66 to 60.79	0 (0 to 68)
Langer, 1993	Vomiting	Number of episodes	−1.204 (−2.494 to 0.086)	0.136	−1.34 to 0.17	87 (NA)*
Lashner, 1990	Symptom score: abdominal pain	Symptom scores 0–100 (0=best, 100=worst)	−3.615 (−16.982 to 9.751)	0.007	−35.0 to 15.0	37 (0 to 73)
	Symptom score: bowel movements/day		−0.538 (−1.215 to 0.138)	0.001	−3.0 to 1.0	56.6 (0 to 81)
	Symptom score: consistency of bowel movements		7.000 (−7.551 to 21.551)	0.013	−25.5 to 33.0	28 (0 to 69)
	Symptom score: haematochezia		2.308 (−17.210 to 21.826)	0.003	−38.0 to 47.5	47 (0 to 78)
	Symptom score: general sense of well-being		−6.538 (−25.352 to 12.275)	0.008	−43.0 to 35.0	35 (0 to 73)
Maier, 1994	SCL-90 subscales: depressed mood	Self-rating inventory to measure the effects of drug	−3.536 (−6.718 to −0.354)	<0.001	−17.8 to 2.74	58 (12 to 80)
	SCL-90 subscales: anxiety		−3.753 (−6.582 to −0.924)	<0.001	−17.4 to 2.5	66 (30 to 83)
	SCL-90 subscales: somatisation		−1.419 (−4.316 to 1.478)	0.869	−6.0 to 2.7	0 (0 to 65)
Mandelcorn, 2004	Self-assessment score	0–5 (0=worst, 5=best)	−2.052 (−8.865 to 4.761)	0.05	−7.7 to 4.9	0 (0 to 85)
	Lower extremity ataxia	Fugl-Meyer: three point (0 cannot be performed to 2 can be fully performed)	12.494 (−3.155 to 28.142)	0.025	−6.42 to 36.76	35 (0 to 77)
	Truncal ataxia	AMTI force plate: NR Berg Balance Scale 0–56, with a higher score indicating a better performance	1.196 (−2.866 to 5.257)	0.690	−0.52 to 2.20	0 (0 to 85)
	Upper extremity ataxia	Purdue Pegboard Test: pegs inserted into the board with each hand in 30 s Minnesota Placing Test: reach out, grasp, and place blocks in a specific order	−0.498 (−3.546 to 2.550)	0.382	−3.68 to 1.42	0 (0 to 85)
McQuay, 1994	VAS pain Intensity	0–100 (0=no pain, 100=worst possible pain)	−1.094 (−5.572 to 3.383)	0.004	−8.0 to 10.1	0 (0 to 49)
	VAS relief Intensity	0–100 (0=no relief, 100=complete pain relief)	−3.913 (−11.729 to 3.903)	0.038	−28.4 to 5.15	0 (0 to 49)
Miyazaki, 1995	Incidence of angina	Either ST segment elevation or depression at rest	0.496 (−0.206 to 1.199)	0.125	−16.19 to 17.11	0 (0 to 60)

Continued

Table 5 Continued

Author, year	Outcome	Definition/range of the scales (severity)	Main effect: Fixed treatment effect	Person-level heterogeneity of treatment effect (HTE): P for person treatment interaction*	Treatment effect range (lower range (CI)–upper range (CI))	I^2 % (CI)
Nathan, 2006	Emetic episodes per day	Complete response (0 episodes/day), major response (1–2 episodes/day) or failure (>2 episodes/day)	−0.095 (−0.514 to 0.325)	0.001	−16.5 to 2.08	59 (6 to 82)
Parodi, 1979	Ischaemic attacks	ST elevation or depression (details NR)	−1.544 (−1.838 to −1.251)	0.007	−16.21 to −0.34	48 (0 to 73)
Parodi, 1986	Asymptomatic ST elevation (after verapamil)	0.1 mV of ST segment elevation measured 20 ms after the J point	−1.637 (−1.994 to −1.279)	0.110	−2.37 to −1.30	6 (0 to 65)
	Asymptomatic ST depression (after verapamil)	More than 0.2 mV of ST segment depression measured 80 ms after the J point	−1.083 (−1.903 to −0.262)	0.401	−17.42 to −0.90	0 (0 to 62)
	Symptomatic ST elevation (after verapamil)		−1.580 (−1.906 to −1.254)	<0.001	−15.40 to −1.45	0 (0 to 62)
	Symptomatic ST depression (after verapamil)		−0.990 (−1.411 to −0.569)	0.002	−2.53 to −0.52	6 (0 to 64)
	Asymptomatic ST elevation (after propranolol)		0.100 (−0.086 to 0.286)	0.006	−0.77 to 1.38	62 (25 to 81)
	Asymptomatic ST depression (after propranolol)		0.339 (−0.168 to 0.845)	0.964	−18.3 to 0.83	0 (0 to 62)
	Symptomatic ST elevation (after propranolol)		−0.002 (−0.177 to 0.173)	0.063	−14.9 to 0.68	46 (0 to 74)
	Symptomatic ST depression (after propranolol)		−0.374 (−0.709 to −0.039)	0.023	−17.1 to −0.73	4 (0 to 64)
Pereira, 1995	INR	Target INR range of 2.0–3.0	−0.126 (−0.312 to 0.060)	0.433	−0.42 to 0.16	0 (0 to 71)
Tison, 2012	Troublesome dyskinesia	7 points scale (1=extremely uncomfortable, 7=not at all uncomfortable)	0.167 (−0.449 to 0.783)	0.593	−0.67 to 1.83	0 (0 to 62)

*The significance of person-level HTE was assessed by a likelihood ratio test comparing the two models—model with common treatment effect and model with treatment-by-participant interactions.
INR, international normalised ratio; NR, not reported; SCL, Symptom Checklist.

with the average effect. However, given the lower number of participants per study and periods per participant and also different analytic approach, estimates of I^2[22] were much less precise in these studies.

DISCUSSION

This review documents that multiperson N-of-1 studies rarely examine HTE. Only 8% of 62 multiperson N-of-1 studies described statistical tests examining HTE, but these generally involved comparisons of treatment effects among groups of patients (eg, based on age or sex) rather than across individuals. Only two studies in the whole of the literature tested for person-level HTE.[15 16] Nevertheless, analysable person-level results are sometimes reported in multiperson N-of-1 studies, as outcomes or as treatment effects, suitable for the analysis of person-level HTE. Our reanalyses of the totality of available data from these studies (n=25) suggested the presence of substantial non-random variation in treatment effects across individuals in most studies. This was evident when considering statistical tests for the variation of treatment effects among patients and also by qualitative assessment of the magnitude of effect variation. This represents the first broad empirical examination with reanalysis of person-level HTE across multiperson

N-of-1 studies, and it provides some general support for the a priori assumption of individual patient variation in treatment response that broadly motivates personalised medicine.

In contrast to parallel-group studies that establish efficacy in a group of patients with a common condition, N-of-1 studies establish the effects of an intervention in an individual.[17] In this respect, N-of-1 studies can be thought of as adjuncts to clinical care, where the goal is to select the right treatment for a particular patient, rather than as a research tool, where the goal is to create new generalisable knowledge.[18 19] Indeed, the results of traditional N-of-1 studies may be generalisable only to the future treatment response of the patient in the trial, not to other patients. Nevertheless, using Bayesian meta-analytic techniques, Zucker *et al* showed how the average treatment effect at the population level can also be estimated by combining multiperson N-of-1 studies testing similar interventions in similar patients with the same outcome measures.[14] Similar Bayesian methods have also been suggested for analysis of group-level HTE.[20]

Herein, we demonstrate yet a new application of N-of-1 studies, to explore person-level HTE. This application has important research and clinical implications, even when the determinants of HTE remain unidentified. It is particularly of interest that there was apparent variation in the *degree* of person-level HTE found across conditions and treatments. Since the degree of variation across individuals sets the upper bound for the amount of HTE that might be explainable by observable characteristics, such as clinical or genomic variables, searching for subgroup effects in the absence of person-level HTE is a futile exercise.[4 21 22]

An interesting example of how person-level HTE can vary across different conditions comes from the study of Johannessen *et al* (figure 3).[15] These investigators conducted N-of-1 patient studies comparing cimetidine to placebo for patients presenting with dyspeptic symptoms and reported person-level effects by subgroups of disease categories. Among 46 trial completers, cimetidine had a significant effect for most patients (57%), as it did at the aggregate level. However, not only was there substantial person-level HTE, but person-level HTE varied across conditions, being much more pronounced in non-ulcer dyspepsia (I^2=75%) compared with peptic ulcer disease (I^2=35%) (figure 3)—despite the very similar overall effects seen in these two conditions.

Finding variation in person-level response in multiperson N-of-1 studies identifies those conditions for which N-of-1 studies are likely to be clinically relevant. For condition-treatment combinations shown to have low person-level HTE, single subject studies are highly unlikely to be clinically informative, and the average results from trials (ie, 'one-size-fits-all' effects) are more apt to be applicable to individuals.[23 24] On the other hand, N-of-1 studies may be highly clinically informative for condition-treatments with a high degree of person-level HTE. These conditions would also be potentially higher yield for examining predictors of HTE (genomic or otherwise).

Our findings also have implications for clinical practice and formulary design. For conditions marked by high person-level HTE, even when trials show that one treatment is better on average than others, having a variety of medication options would be useful to optimise outcomes across all patients, particularly for chronic conditions such as those studied here where empiric trials of alternative medications to find the best treatment for an individual might be feasible. For example, the study by March *et al*[25] shows that while patients with osteoarthritis on average had less pain and less stiffness with diclofenac, some patients had improved symptoms on paracetemol. This person-level HTE may not be detectable in conventional parallel-arm trials employing conventional subgroup analysis.[21]

While more studies combining N-of-1 studies are needed to understand the extent of person-level HTE, future studies need to apply greater methodological rigour to improve the state-of-the-science on evaluation of individual treatment effects.[26] While the recently published Consolidated Standards of Reporting Trials Extension for N-of-1 trials may help improve reporting, a tabulation of all information (possibly electronically available) appears the most straightforward way to facilitate the clinical interpretation of these studies.[27] Such reporting allows the inspection of trajectories over time and may reveal patterns that are not captured by regression models. Complete reporting would also facilitate the development and evaluation of methods for the analysis of single subject experiments, particularly its use to better understand the extent and importance of person-level HTE.

The limitations of this review reflect, to a large extent, the limitations of the data in primary studies. Many conditions are not amenable to the N-of-1 design (eg, because treatment effects are cumulative or because outcomes are observed only once). Further, even for conditions and treatment that are potentially amenable to this design, many important disease categories lacked published N-of-1 studies. We relied on published studies only and our analytic cohort may be an underestimation of the true prevalence of these studies—particularly for N-of-1 studies, which may frequently be conducted without the intention of future publication.

In addition, our conclusions regarding the ubiquity of HTE in the data we reanalysed should be interpreted in the context of several important limitations. First, there were only a limited number of available studies that reported data sufficient to analyse, and therefore we present only a very partial picture of the full scope of interindividual variation in effects across clinical conditions. Furthermore, among the studies that did have data, only fairly small number of patients were observed over a small number of treatment periods and we frequently had to rely on data summaries provided by the authors (eg, person-level treatment effects and their

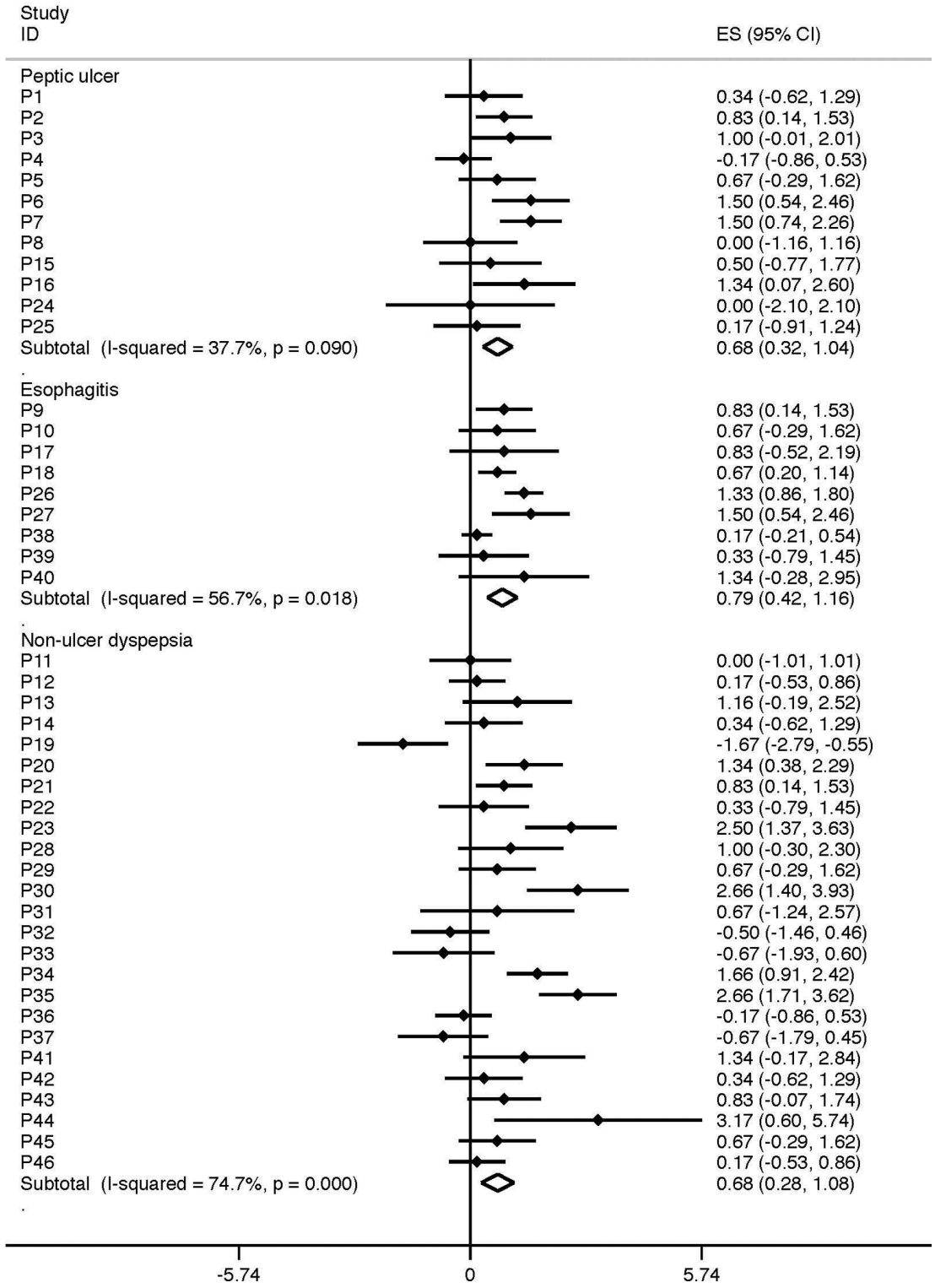

Figure 3 Person-level variation across different disease conditions. This figure depicts the results of 46 different N-of-1 trials of cimetidine as reported by Johanessen *et al*.[12] The effect of cimetidine versus placebo was measured in each subject across 12 crossover periods over the span of 184 days. While cimetidine had a similar average effect regardless of the index condition, there was far greater consistency of effect in patients with peptic ulcer disease and much more variation in effect among patients with non-ulcer dyspepsia.

sampling variance); these data limitations precluded the use of more complex models, for example, models that account for period effects or other effects of time on the outcome.[3]

Our review has demonstrated that HTE remains almost totally unexplored in multiperson N-of-1 studies, which are uniquely capable of exploring variations in individual (person-level) treatment effects. Our reanalysis of the

data from these studies represents the first systematic attempt to obtain empirical support for the a priori argument that treatment effects vary across individual patients, an assumption which underpins all efforts to personalise treatment selection. In this sample, person-level HTE appears to be common and large enough to be clinically meaningful; the degree of person-level HTE appears to vary across conditions and outcomes. Thus, multiperson N-of-1 studies are an under-utilised tool to identify where person-level HTE may be substantial and where efforts to find molecular or clinical predictors of response heterogeneity should be focused. In such conditions, parallel arm studies might yield results that are over-generalised for patient level decision-making.

Author affiliations
[1]Center for Clinical Evidence Synthesis, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center/Tufts University School of Medicine, Boston, MA, USA
[2]Center for Evidence Synthesis in Health, School of Public Health, Brown University, Providence, Rhode Island, USA
[3]Predictive Analytics and Comparative Effectiveness (PACE) Center, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center/Tufts University School of Medicine, Boston, Massachusetts, USA
[4]Center for the Evaluation of Value and Risk in Health, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center/Tufts University School of Medicine, Boston, MA, USA
[5]VA Boston Healthcare System, Center for Healthcare Organization and Implementation Research (CHOIR), Boston, Massachusetts, USA
[6]National Pharmaceutical Council, Washington, District of Columbia, USA
[7]Department of Internal Medicine, University of California, Davis, San Francisco, California, USA

Acknowledgements We would like to acknowledge Issa Dahabreh, MD, MS, Assistant Professor of Health Services, Policy and Practice, Assistant Professor of Epidemiology, Brown University, for statistical advice. We would like to acknowledge Tatum Williamson, MS, Research Assistant, Predictive Analytics and Comparative Effectiveness Center, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, for assistance with updating literature.

Contributors GR and DMK made substantial contributions to the conception or design of the work; the acquisition, analysis or interpretation of data for the work; responsible for drafting the work or revising it critically for important intellectual content and have made an agreement 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. All authors have given final approval of the version to be published.

Funding This work was supported by the National Pharmaceutical Council. Additional support was provided by the Patient-Centered Outcomes Research Institute (PCORI) Award (Predictive Analytics Resource Center (SA.Tufts.PARC.OCSO.2018.01.25) and the National Institutes of Health (3UL1TR001079-04S1).

Disclaimer All statements in this report, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology Committee.

Competing interests None declared.

Patient consent Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data are available.

REFERENCES

1. Wallach JD, Sullivan PG, Trepanowski JF, *et al*. Sex based subgroup differences in randomized controlled trials: empirical evidence from Cochrane meta-analyses. *BMJ* 2016;355:i5826.
2. Hertz DL, McLeod HL. Use of pharmacogenetics for predicting cancer prognosis and treatment exposure, response and toxicity. *J Hum Genet* 2013;58:346–52.
3. Kitsios GD, Kent DM. Personalised medicine: not just in our genes. *BMJ* 2012;344:e2161.
4. Senn S. Individual response to treatment: is it a valid assumption? *BMJ* 2004;329:966–8.
5. Kalow W, Tang BK, Endrenyi L. Hypothesis: comparisons of inter- and intra-individual variations can substitute for twin studies in drug research. *Pharmacogenetics* 1998;8:283–9.
6. Gabler NB, Duan N, Vohra S, *et al*. N-of-1 trials in the medical literature: a systematic review. *Med Care* 2011;49:761–8.
7. Institute of Medicine (US) Committee on Standards for Systematic Reviews of Comparative Effectiveness Research. *Finding What Works in Health Care: Standards for Systematic Reviews*, 2011.
8. Moher D, Liberati A, Tetzlaff J, *et al*. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med* 2009;151:264.
9. AHRQ. Methods guide for effectiveness and comparative effectiveness review. AHRQ Publication No 10(11)-EHC063-EF Chapters 2011. http://effectivehealthcare.ahrq.gov/
10. Higgins JP, Altman DG, Gøtzsche PC, *et al*. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011;343:d5928.
11. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;7:177–88.
12. Schmidt FL, Oh IS, Hayes TL. Fixed- versus random-effects models in meta-analysis: model properties and an empirical comparison of differences in results. *Br J Math Stat Psychol* 2009;62(Pt 1):97–128.
13. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002;21:1539–58.
14. Zucker DR, Ruthazer R, Schmid CH. Individual (N-of-1) trials can be combined to give population comparative treatment effect estimates: methodologic considerations. *J Clin Epidemiol* 2010;63:1312–23.
15. Johannessen T, Petersen H, Kristensen P, *et al*. Cimetidine on-demand in dyspepsia. Experience with randomized controlled single-subject trials. *Scand J Gastroenterol* 1992;27:189–95.
16. Pereira JA, Holbrook AM, Dolovich L, *et al*. Are brand-name and generic warfarin interchangeable? Multiple n-of-1 randomized, crossover trials. *Ann Pharmacother* 2005;39:1188–93.
17. Guyatt GH, Heyting A, Jaeschke R, *et al*. N of 1 randomized trials for investigating new drugs. *Control Clin Trials* 1990;11:88–100.
18. Guyatt G, Sackett D, Taylor DW, *et al*. Determining optimal therapy – randomized trials in individual patients. *N Engl J Med* 1986;314:889–92.
19. Guyatt G, Sackett D, Adachi J, *et al*. A clinician's guide for conducting randomized trials in individual patients. *CMAJ* 1988;139:497–503.
20. Henderson NC, Louis TA, Wang C, *et al*. Bayesian analysis of heterogeneous treatment effects for patient-centered outcomes research. *Health Serv Outcomes Res Methodol* 2016;16:213–33.
21. Dahabreh IJ, Hayward R, Kent DM. Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. *Int J Epidemiol* 2016;45:dyw125.
22. Senn S, Rolfe K, Julious SA. Investigating variability in patient response to treatment--a case study from a replicate cross-over study. *Stat Methods Med Res* 2011;20:657–66.
23. Simon G. Choosing a first-line antidepressant: equal on average does not mean equal for everyone. *JAMA* 2001;286:3003–4.
24. Simon GE, Psaty BM, Hrachovec JB, *et al*. Principles for evidence-based drug formulary policy. *J Gen Intern Med* 2005;20:964–8.
25. March L, Irwig L, Schwarz J, *et al*. n of 1 trials comparing a non-steroidal anti-inflammatory drug with paracetamol in osteoarthritis. *BMJ* 1994;309:1041–4.
26. DEcIDE Methods Center N-of-1 Guidance Panel. In: Kravitz RL, Duan N, eds. *Design and implementation of N-of-1 trials: a user's guide. AHRQ Publication No. 13(14)-EHC122-EF*. Rockville, MD: Agency for Healthcare Research and Quality, 2014.
27. Vohra S, Shamseer L, Sampson M, *et al*. CONSORT extension for reporting N-of-1 trials (CENT) 2015 statement. *BMJ* 2015;350:h1738.