Supplementary Material for Bayesian model-based outlier detection in network meta-analysis

Silvia Metelli*, Dimitris Mavridis, Perrine Créquit, Anna Chaimani
*email: silvia.metelli@u-paris.fr

1 Additional details and results from the simulation study

In this section, we report additional details and results for the extensive simulation study described in Section 6 of the main document. The number of studies per comparison was set to 10 for all comparisons in an ideally balanced design and ranged from 1 to 9 to reflect values more often encountered in practice in three different unbalanced designs. The amount of heterogeneity and number of induced outliers also varied across scenarios. In total, we explored 32 different scenarios. Below, we provide additional details for each method analysed and compared in simulations.

1.1 Interpretation of Bayes Factors

Table 1 reports a commonly used scale for interpretation of Bayes Factors, proposed in a seminal paper by Kass and Raftery (1995).

$B_{1:0}$	log10($B_{1:0}$)	Evidence in favour of outlying behaviour
$1 \leq B_{1:0} < 3.2$	$0 \leq log_1 0(B_{1:0})) < 0.5$	Weak
$3.2 \leq B_{1:0} < 10$	$0.5 \leq log_1 0(B_{1:0}) < 1$	Moderate
$10 \leq B_{1:0} < 100$	$1 \leq log_1 0(B_{1:0}) < 2$	Strong
$B_{1:0} \geq 100$	$log_1 0(B_{1:0}) \geq 2$	Decisive

Table 1: Kass and Raftery table for interpretation of Bayes Factors evidence.

1.2 Gelman's discrepancy measure

For reference, in (1) we report the typical choice of discrepancy measure, namely the well-known omnibus goodness of fit discrepancy measure Gelman's discrepancy measure that was used in Zhang et al. (2015). Suppose $\boldsymbol{y}$ is a vector of observed data (continuous or binary), and $y_{i,k}$ the observed data in study i and arm k. The measure is then given by

$$f_i^G = \frac{\left|y_{i,k} - E\left(Y_{i,k}|\theta, \tau^2\right)\right|^2}{Var\left(Y_{i,k}|\theta, \tau^2\right)}. \tag{1}$$

In the following sections, posterior predictive p-values based on Gelman's discrepancy will be denoted p_G.

1.3 Conditional Predictive Ordinate (CPO)

The CPO expresses the posterior probability of observing the value (or set of values) of y_i when the model is fitted to all data except y_i, with a larger value indicating a better fit of the model to y_i and very low CPO values suggest that y_i is an outlier and an influential point. The CPO is connected with the frequentist studentised residual test for outlier detection, as it is known that data points with large studentised residuals have small CPOs and will be detected as outliers. For interpretation, it is usual to plot the Inverse-CPOs (ICPOs), where a value larger than 40 can be considered as possible outlier, and higher than 70 as extreme values (**?**). For study i, the general formula for CPO can be expressed as

$$\mathrm{CPO}_i = P\left(y_i|\boldsymbol{y}_{-i}\right) = \int P\left(y_i|\boldsymbol{y}_{-i},\boldsymbol{\theta}\right) P\left(\boldsymbol{\theta}|\boldsymbol{y}_{-i}\ \right) d\boldsymbol{\theta}\ = \left(\int \frac{1}{P\left(y_i|\boldsymbol{y}_{-i},\boldsymbol{\theta}\right)} P\left(\boldsymbol{\theta}|\boldsymbol{y}_{-i}\ \right) d\boldsymbol{\theta}\right)^{-1}$$

where $\boldsymbol{y}_{-i}$ is the vector of general observed data (binary or continuous) after having removed the i-th observation, and $\boldsymbol{\theta}$ represents the parameter collection. Thus, the CPO can be estimated by taking the inverse of the posterior mean of the inverse density function of y_i. In words, CPO estimates the probability of observing y_i in the future, after having already observed $\boldsymbol{y}_{-i}$. Estimation of the CPO is available via Integrated Nested Laplace Approximation (INLA), using 'nmaINLA' R package (Gelfand, 1995; Rue et al., 2009). As CPOs tend to confound outliers with influential values, we choose a conservative threshold for detection, equal to CPO ≥ 70.

1.4 Forward Search Algorithm

As briefly described in the main paper, outlier detection in network meta-analysis via FS algorithm operates by choosing a small starting subset of studies considered to be outlier-free, called the basic set and sequentially adding the remaining studies according to some measure of how close the study is to the evolving set, under some posited model. Abrupt changes in one or more of the monitored measures may indicate outlyingness. Here, we monitored mean Cook's distance as diagnostic measure: values falling well above the threshold of 1 correspond to all or some of the artificially generated outliers. FS search for NMA is available via the 'NMAoutlier' R package (Petropoulou et al., 2021).

1.5 Additional results

In the main test we report results for the unbalanced but faily-connected scenario, while here we report additional results for the remaining scenarios analysed in the simulation study. Specifically, Table 2, Table 3 and Table 4 respectively report results for the case of a balanced fairly connected network, an unbalanced but well-connected network and finally, an unbalanced and poorly connected network.

Notation is used as follows. BF: Bayes Factor test; LR: Likelihood Ratio test as in Noma et al. 2020 (bootstrapped p-values reported); $\boldsymbol{p}_{\mathrm{L}}$, $\boldsymbol{p}_{\mathrm{SDO}}$ and $\boldsymbol{p}_{\mathrm{G}}$ posterior predictive p-values under under likelihood-based discrepancy in (9), Stahel-Donoho outlyingness discrepancy in (10), and Gelman's Omnibus χ^2 as in Zhang et al. 2015, CPO: conditional predictive ordinate values; FS: Forward search algorithm as in Mavridis et al 2017 (Cook's distance reported); τ^2: heterogeneity (thresholds for detection: $BF > 3.2$ for Bayes factors; $p < 0.05$ for p-values, $CPO > 40$ for conditional predictive ordinates and Cook's distance > 1 for FS algorithm).

Table 2: Mean Bayes factors and mean posterior predictive p-values for the induced outliers of 1000 simulated data sets for the balanced design with a fairly connected network of 100 studies (scenarios 1-8 in Table 1 in main text). At varying scenarios, either a single outlier or three outliers (outlier 1, outlier 2, outlier 3) are induced in the network.

τ^2	**Induced outliers**	**BF**	p_L	p_{SDO}	**LR**	p_G	**CPO**	**FS**
0	a single outlier	873.4	<0.001	0.001	0.02	<0.001	99	4.2
	outlier 1	1113.1	0.01	0.05	0.01	< 0.001	95	3.1
	outlier 2	128.2	0.01	0.001	0.02	0.01	110	2.3
	outlier 3	582.3	0.001	0.01	0.02	0.05	100	2.4
0.032	a single outlier	1540.1	0.002	< 0.0001	0.01	0.01	132	3.0
	outlier 1	987.1	0.05	0.01	0.01	0.07	72	2.3
	outlier 2	222.1	0.001	0.001	0.01	0.02	89	2.3
	outlier 3	332.1	0.01	0.01	0.03	0.06	88	2.7
0.096	a single outlier	87.1	0.05	0.05	0.02	0.07	35	1.0
	outlier 1	44.1	0.05	0.05	0.03	0.13	31	1.0
	outlier 2	18.7	0.03	0.05	0.04	0.05	39	0.6
	outlier 3	11.8	0.04	0.04	0.04	0.10	27	0.7
0.287	a single outlier	7.4	0.04	0.08	0.06	0.07	30	0.3
	outlier 1	3.5	0.12	0.12	0.13	0.40	22	0.9
	outlier 2	2.3	0.10	0.12	0.10	0.25	44	0.6
	outlier 3	0.99	0.13	0.15	0.12	0.31	23	0.5

Table 3: Mean Bayes factors and mean posterior predictive p-values for the induced outliers of 1000 simulated data sets for the unbalanced design with a well-connected network of 35 studies (scenarios 9-16 in Table 1 in main text). At varying scenarios, either a single outlier or three outliers (outlier 1, outlier 2, outlier 3) are induced in the network.

τ^2	**Induced outliers**	**BF**	p_L	p_{SDO}	**LR**	p_G	**CPO**	**FS**
0	a single outlier	845.6	<0.001	0.02	0.001	0.03	90	2.5
	outlier 1	580.0	0.01	0.05	0.01	0.04	71	2.2
	outlier 2	235.4	0.01	0.001	0.02	0.03	100	2.2
	outlier 3	184.1	0.001	0.01	0.01	0.04	88	2.5
0.032	a single outlier	1540.1	0.002	< 0.001	0.01	0.01	132	3.0
	outlier 1	117.1	0.05	0.01	0.04	0.07	63	1.8
	outlier 2	82.4	0.001	<0.001	0.02	0.05	65	2.7
	outlier 3	22.3	0.01	0.01	0.08	0.20	79	1.7
0.096	a single outlier	11.1	0.05	0.05	0.04	0.07	35	1.0
	outlier 1	7.1	0.06	0.06	0.07	0.20	30	0.8
	outlier 2	3.7	0.03	0.05	0.05	0.25	35	0.6
	outlier 3	2.6	0.04	0.05	0.06	0.13	41	0.6
0.287	a single outlier	3.5	0.04	0.08	0.06	0.07	30	0.3
	outlier 1	1.9	0.22	0.12	0.10	0.40	12	0.6
	outlier 2	2.1	0.10	0.12	0.10	0.25	30	0.6
	outlier 3	1.1	0.17	0.15	0.12	0.31	11	0.5

Table 4: Mean Bayes factors and mean posterior predictive p-values for the induced outliers of 1000 simulated data sets for the unbalanced design with a poorly connected network of 15 studies (scenarios 25-32 in Table 1 in main text). At varying scenarios, either a single outlier or three outliers (outlier 1, outlier 2, outlier 3) are induced in the network.

τ^2	Induced outliers	BF	p_L	p_{SDO}	LR	p_G	CPO	FS
0	a single outlier	111.0	0.01	0.01	0.03	0.01	81	1.7
	outlier 1	153.6	0.01	0.01	0.03	0.04	78	1.2
	outlier 2	118.2	0.03	0.03	0.03	0.08	85	1.1
	outlier 3	14.6	0.03	0.05	0.05	0.05	72	0.7
0.032	a single outlier	28.1	0.10	0.10	0.01	0.22	132	3.0
	outlier 1	188.2	0.07	0.05	0.02	0.27	60	1.5
	outlier 2	234.1	0.04	0.05	0.01	0.30	71	1.2
	outlier 3	12.1	0.03	0.05	0.08	0.43	72	1.9
0.096	a single outlier	8.6	0.10	0.10	0.02	0.34	35	1.2
	outlier 1	9.1	0.10	0.13	0.10	0.50	32	0.6
	outlier 2	2.6	0.08	0.06	0.20	0.51	25	0.7
	outlier 3	2.2	0.10	0.03	0.19	0.60	44	0.7
0.287	a single outlier	2.5	0.04	0.12	0.06	0.30	30	0.3
	outlier 1	2.1	0.22	0.32	0.12	0.40	22	0.5
	outlier 2	1.8	0.13	0.10	0.22	0.54	12	0.5
	outlier 3	1.3	0.19	0.25	0.15	0.51	9	0.5

2 Additional details and results for the down-weighting scheme

In this section we provide additional details which can constitute useful insight for appropriately choosing the hyperparameters of the beta distribution for the down-weighting scheme presented in Section 4 of the main paper and we report additional results for the relative bias of network meta-analysis estimates with and without down-weighting for the remaining simulated scenarios with three outliers. More specifically, in Figure 1 we show a number of different Beta distributions which correspond to applying less and more severe down-grading while in Figure 2, Figure 3 and Figure 4 we show the relative bias of the estimates under the balanced scenario, unbalanced scenario with well-connected network and unbalanced scenario with poorly-connected network.

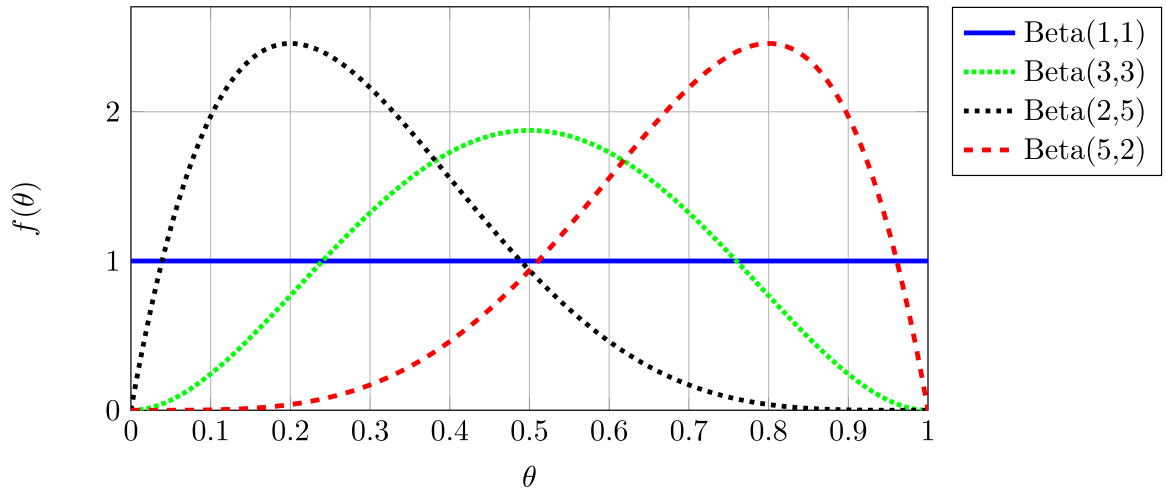

Figure 1: Examples of possible Beta distributions to be used in the down-weighting scheme.

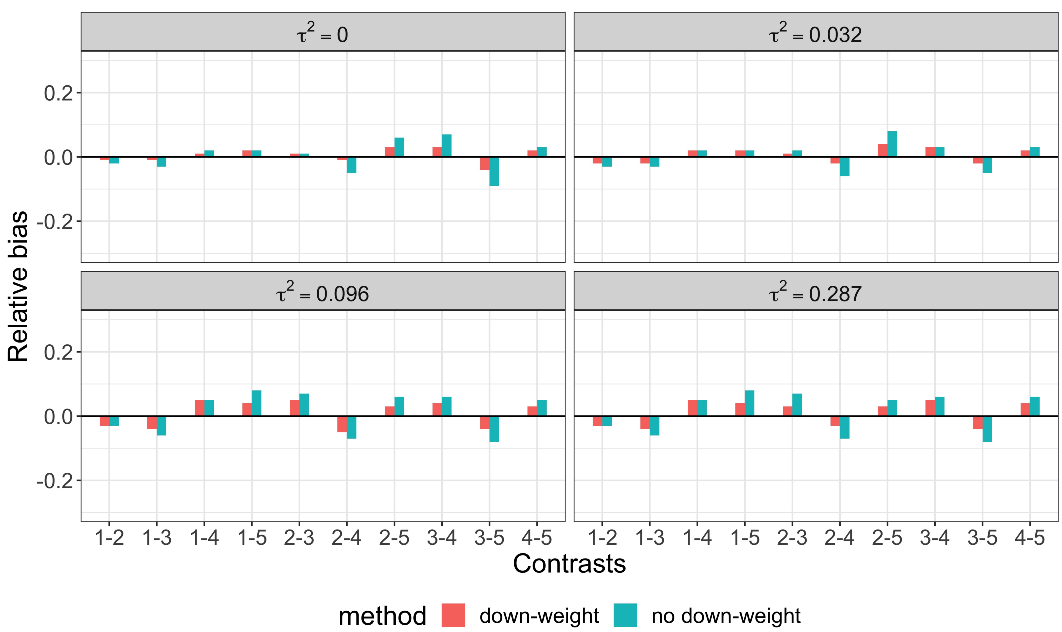

Figure 2: Relative bias plot at varying heterogeneity for the balanced design.

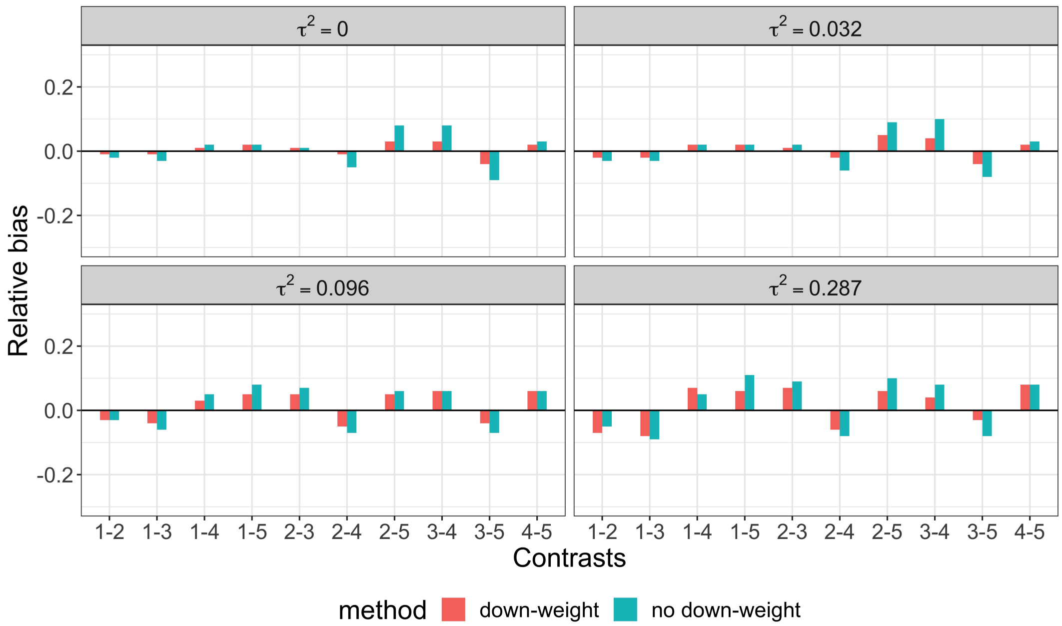

Figure 3: Relative bias plot at varying heterogeneity for the unbalanced, well-connected design.

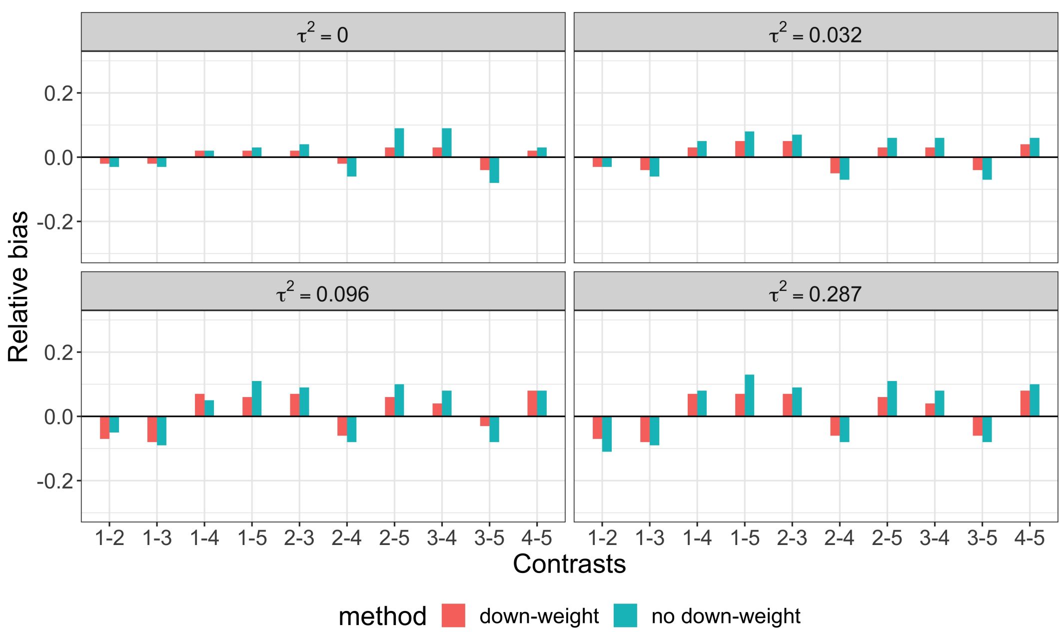

Figure 4: Relative bias plot at varying heterogeneity for the unbalanced, fairly-connected design.

3 Additional details and results from real-world data

In this section, we report additional details and results from the exemplar data sets introduced in Section 2 in the main text. Figure 5 reports the full network of second-line treatments for second-line non-small cell lung cancer, which we further grouped in the main text at treatment class level to facilitate visualisation. Figure 6 reports additional results from posterior predictive assessments under the Stahel Dohono outlyingness (SDO) measure for both the lung cancer data and the smoking cessation data. Finally, Table 5 reports the full contribution matrix for the lung cancer data.

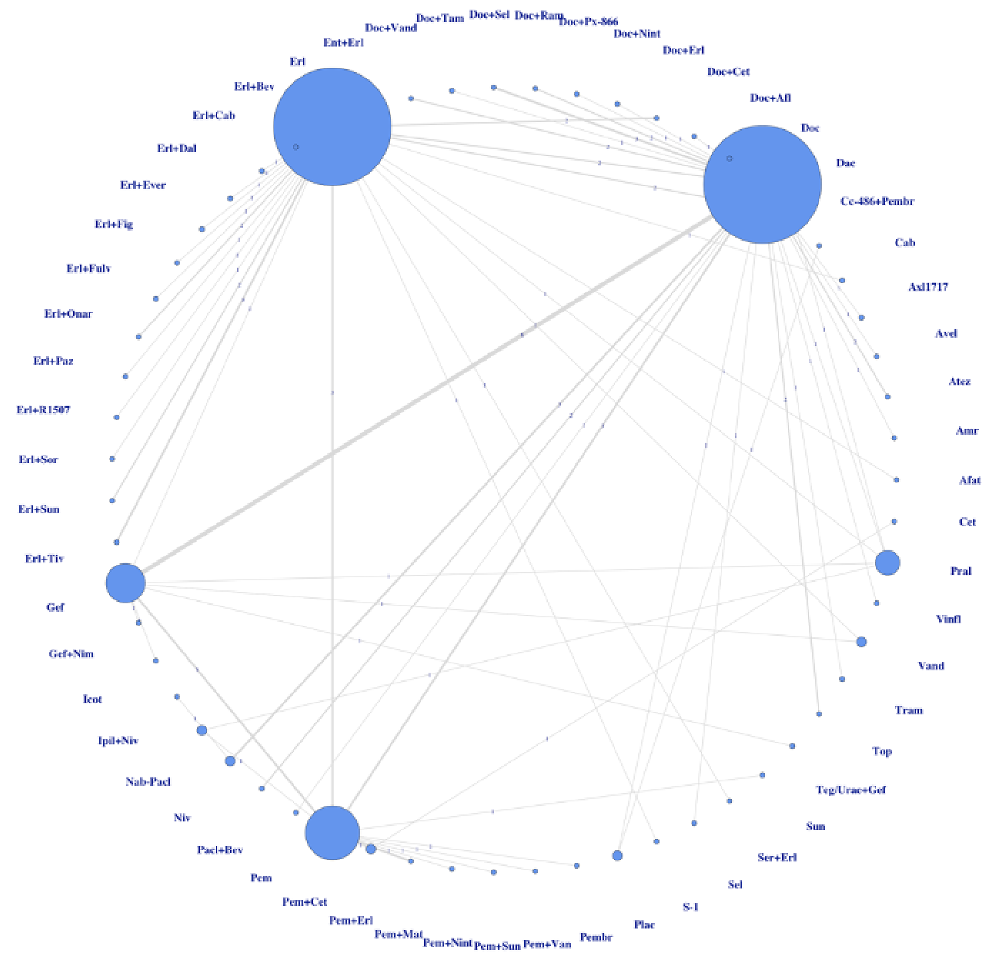

Figure 5: Full network of second-line non-small cell lung cancer treatments at drug-level. Each node represents a different drug and an edge between two drugs exists if they have been compared in at least one study. Edge weight is proportional to the number of studies comparing each two drugs.

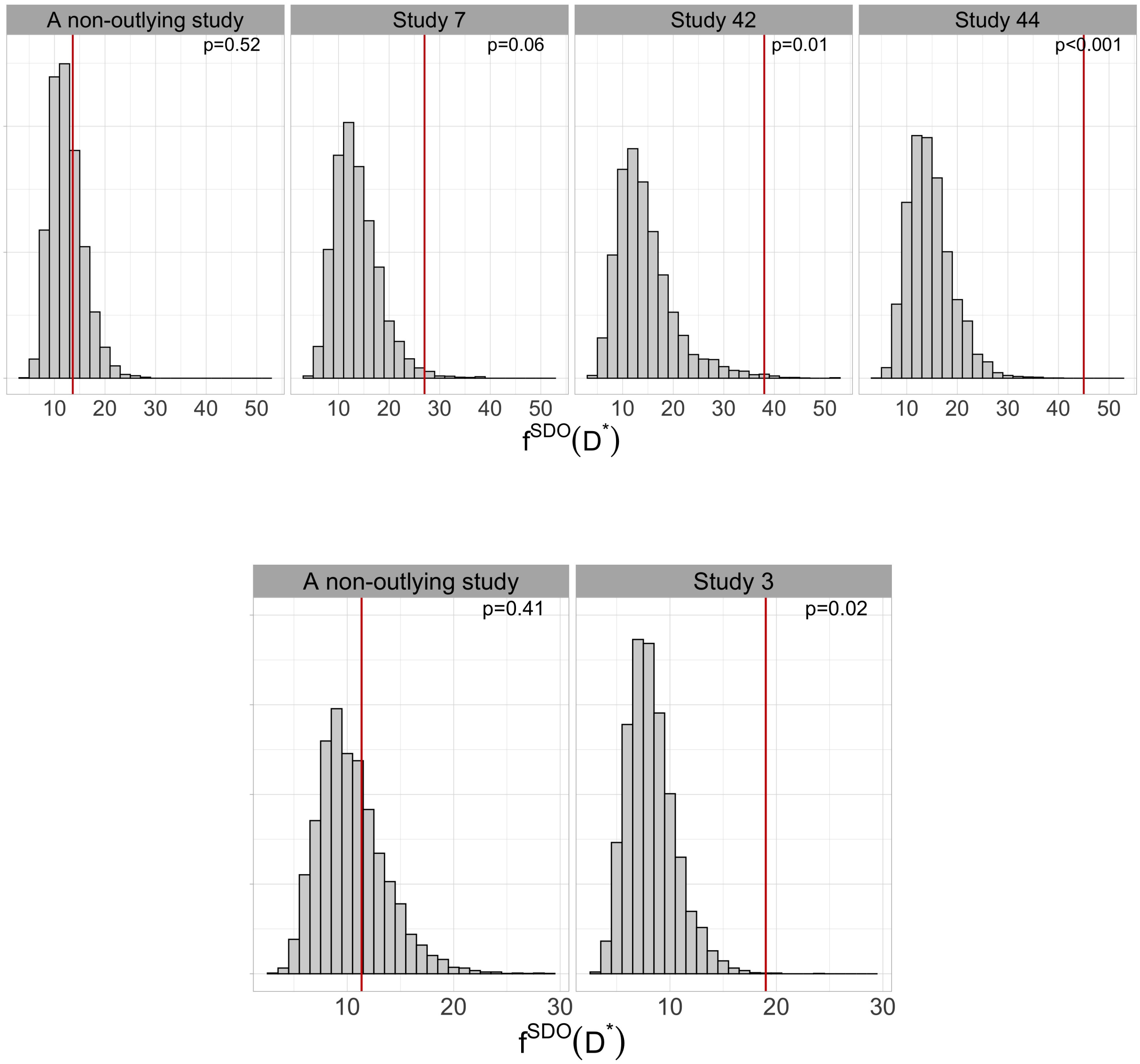

Figure 6: Histograms of draws from the posterior predictive distribution for the replicated vs.realised likelihood (vertical line) for the potential outliers identified under the SDO-based discrepancy measure f, alongside a randomly chosen non-outlying study used as comparison (with annotated posterior predictive p-values). The two upper plots correspond to lung cancer data, and the lower two plots to smoking cessation data.

	Chemo+Dual *vs* Dual	Chemo+Dual*vs* Immuno	Dual *vs* Immuno	Dual *vs* Monochemo	Immuno *vs* Targeted
Study 1	1.1881	0	1.073	1.5956	0
Study 2	2.04130	1.8437	2.7415	0	
Study 3	1.6037	0.1072	1.7111	2.5667	2.5667
Study 4	0.4828	0	0.436	0.6484	0
Study 5	3.7462	0	3.3834	5.031	0
Study 6	0.4387	0.0293	0.4681	0.7022	0.7022
Study 7	0.3918	0.6588	0.0355	0.0473	0.0473
Study 8	0.7692	0.0514	0.8208	1.2312	1.2312
Study 9	0 5.1246	3.3932	0	5.0456	
Study 10	0	9.1686	6.0709	0	9.0272
Study 11	1.3957	0.3313	0.4489	0.5985	0.5985
Study 12	1.16	0.0776	1.2377	1.8566	1.8566
Study 13	1.0392	0	0.9386	1.3956	0
Study 14	1.8226	0.1219	1.9447	2.9171	2.9171
Study 15	0.6708	1.1278	0.0608	0.081	0.081
Study 16	0.4425	0	0.3996	0.5942	0
Study 17	1.0522	0	0.9503	1.413	0
Study 18	0.3661	0.6155	0.0332	0.0442	0.0442
Study 19	2.3525	0.1573	2.5101	3.7651	3.7651
Study 20	0.5048	0.8487	0.0457	0.061	0.061
Study 21	6.9758	0.4665	7.4432	11.1647	11.1647
Study 22	1.2917	0.0864	1.3782	2.0673	2.0673
Study 23	0.3495	0.5876	0.0317	0.0422	0.0422
Study 24	1.4085	0.3344	0.453	0.604	0.604
Study 25	0.7256	0.0485	0.7743	1.1614	1.1614
Study 26	0.7432	0.0497	0.793	1.1895	1.1895
Study 27	1.6599	0.111	1.7711	2.6566	2.6566
Study 28	0	7.5961	5.0297	0	7.479
Study 29	2.4362	4.0962	0.2207	0.2943	0.2943
Study 30	1.5671	2.635	0.142	0.1893	0.1893
Study 31	0.9227	0.0617	0.9845	1.4767	1.4767
Study 32	0.5293	0.89	0.048	0.0639	0.0639
Study 33	0.9124	0.061	0.9736	1.4604	1.4604
Study 34	0.368	0.6187	0.0333	0.0445	0.0445
Study 35	0.9047	0.0605	0.9653	1.448	1.448
Study 36	0.3987	0.6704	0.0361	0.0482	0.0482
Study 37	0.9314	0.0623	0.9938	1.4908	1.4908
Study 38	3.7002	6.2214	0.3352	0.447	0.447
Study 39	2.2471	3.7783	0.2036	0.2715	0.2715
Study 40	0.8664	0.2057	0.2786	0.3715	0.3715
Study 41	5.2328	0	4.7261	7.0276	0
Study 42	2.3506	0	2.123	3.1568	0
Study 43	0.4943	0	0.4464	0.6638	0
Study 44	**0**	**13.3458**	**8.8368**	**0**	**13.1399**
Study 45	0 4.5442	3.0089	0	4.4741	
Study 46	1.2925	0	1.1674	1.7358	0
Study 47	2.8403	4.7756	0.2573	0.3431	0.3431
Study 48	3.2163	5.4078	0.2914	0.3885	0.3885
Study 49	5.9404	0	5.3652	7.9779	0
Study 50	4.1172 0	3.7186	5.5294	0	
Study 51	0.3254	0.5471	0.0295	0.0393	0.0393
Study 52	1.7079	2.8716	0.1547	0.2063	0.2063
Study 53	0.8102	0.0542	0.8645	1.2967	1.2967
Study 54	1.0484	0	0.9469	1.408	0
Study 55	1.4037	0	1.2678	1.8851	0
Study 56	0.8126	0	0.7339	1.0913	0
Study 57	1.2512	0.0837	1.3351	2.0026	2.0026
Study 58	1.4085	0.3344	0.453	0.604	0.604
Study 59	0.7981	0	0.7208	1.0718	0
Study 60	0.7246	1.2184	0.0656	0.0875	0.0875
Study 62	0.5845	0.9828	0.053	0.0706	0.0706
Study 63	2.8261	0.6709	0.909	1.2119	1.2119
Study 64	1.4457	2.4308	0.131	0.1746	0.1746
Study 65	3.1145	5.2366	0.2822	0.3762	0.3762
Study 66	0.5745	0.9659	0.052	0.0694	0.0694
Study 67	0	9.2824	6.1463	0	9.1393
Study 68	0.7068	0	0.6384	0.9493	0
Study 69	1.7788	0	1.6065	2.3888	0

Table 5: Per-study contribution matrix in the lung cancer data set.

References

Zhang, J., Fu, H. and Carlin, B. P. (2015). Detecting outlying trials in network meta-analysis. *Statistics in Medicine* **34**, 2695–2707.

Kass, R.E. and Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

Petropoulou, M., Salanti, G., Rücker, G., Schwarzer, G., Moustaki I., and Mavridis, D. (2021). A forward search algorithm for detection of extreme study effects in network meta-analysis. *Statistics in Medicine* 1, 1-25.

Gelfand, A. E. (1995). Model Determination Using Sampling-Based Methods, In: Gilks W, Richardson S and Spiegelhalter D (eds) Markov Chain Monte Carlo In Practice, London, Chapman Hall.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). Journal of the Royal Statistical Society B **71,** 319–392.

Noma H., Gosho M., Ishii R., Oba, K. and Furukawa, T. A. (2020). Outlier detection and influence diagnostics in network meta-analysis. *Research Synthesis Methods* **11**, 891–902.

Mavridis, D., Moustaki, I., Wall, M. and Salanti, G. (2017). Detecting outlying studies in meta-regression models using a forward search algorithm. *Research Synthesis Methods* **8**, 199–211.