Janiaud et al. BMC Medicine (2021) 19:157
https://doi.org/10.1186/s12916-021-02020-6

BMC Medicine

RESEARCH ARTICLE Open Access

Validity of observational evidence on putative risk and protective factors: appraisal of 3744 meta-analyses on 57 topics

Perrine Janiaud[1,2], Arnav Agarwal[3], Ioanna Tzoulaki[4,5], Evropi Theodoratou[6,7], Konstantinos K. Tsilidis[4,5], Evangelos Evangelou[4,5] and John P. A. Ioannidis[1,8,9,10,11*]

Abstract

Background: The validity of observational studies and their meta-analyses is contested. Here, we aimed to appraise thousands of meta-analyses of observational studies using a pre-specified set of quantitative criteria that assess the significance, amount, consistency, and bias of the evidence. We also aimed to compare results from meta-analyses of observational studies against meta-analyses of randomized controlled trials (RCTs) and Mendelian randomization (MR) studies.

Methods: We retrieved from PubMed (last update, November 19, 2020) umbrella reviews including meta-analyses of observational studies assessing putative risk or protective factors, regardless of the nature of the exposure and health outcome. We extracted information on 7 quantitative criteria that reflect the level of statistical support, the amount of data, the consistency across different studies, and hints pointing to potential bias. These criteria were level of statistical significance (pre-categorized according to 10^{-6}, 0.001, and 0.05 *p*-value thresholds), sample size, statistical significance for the largest study, 95% prediction intervals, between-study heterogeneity, and the results of tests for small study effects and for excess significance.

* Correspondence: jioannid@stanford.edu
[1]Meta-Research Innovation Center at Stanford (METRICS), Stanford, CA 94305, USA
[8]Department of Epidemiology and Population Health, Stanford University School of Medicine, Stanford, CA 94305, USA
Full list of author information is available at the end of the article

Results: 3744 associations (in 57 umbrella reviews) assessed by a median number of 7 (interquartile range 4 to 11) observational studies were eligible. Most associations were statistically significant at $P < 0.05$ (61.1%, 2289/3744). Only 2.6% of associations had $P < 10^{-6}$, ≥1000 cases (or ≥20,000 participants for continuous factors), $P < 0.05$ in the largest study, 95% prediction interval excluding the null, and no large between-study heterogeneity, small study effects, or excess significance. Across the 57 topics, large heterogeneity was observed in the proportion of associations fulfilling various quantitative criteria. The quantitative criteria were mostly independent from one another. Across 62 associations assessed in both RCTs and in observational studies, 37.1% had effect estimates in opposite directions and 43.5% had effect estimates differing beyond chance in the two designs. Across 94 comparisons assessed in both MR and observational studies, such discrepancies occurred in 30.8% and 54.7%, respectively.

Conclusions: Acknowledging that no gold-standard exists to judge whether an observational association is genuine, statistically significant results are common in observational studies, but they are rarely convincing or corroborated by randomized evidence.

Keywords: Umbrella review, Observation studies, Randomized clinical trials, Mendelian randomization

Background

The validity of observational studies of putative risk or protective factors is a subject of continuous debate. Critics focus on the weaknesses of the observational evidence and occasionally debates get further fueled by comparisons against other designs, in particular randomized trials. Usually debates address either single research questions or few associations [1, 2]. However, now we have the opportunity to assess systematically collected and synthesized evidence from thousands of observational associations. In the last decade, numerous umbrella reviews have summarized systematically the evidence from meta-analyses of observational epidemiological studies across entire fields of research [3, 4]. Umbrella reviews also typically assess the observational evidence by looking at the level of statistical support (statistical significance of results), the amount of data, the consistency across different studies, and hints pointing to potential bias. A series of seven standardized quantitative criteria (Table 1 and Additional file 1: Appendix Method 1) have been previously proposed and are commonly used [3–6].

Some of these umbrella reviews have also included systematic assessments of meta-analyses of randomized controlled trials (RCTs) and of Mendelian randomization (MR) studies (an alternative way to generate an equivalent to randomization under certain assumptions using genetic instruments) [7]. Juxtaposing observational and randomized evidence may allow to corroborate results and probe causality.

Here, we overview the evidence obtained from 3744 meta-analyses of observational studies included in umbrella reviews evaluating putative risk or protective non-genetic factors. We evaluate how these meta-analyses of observational studies perform on different quantitative criteria that address statistical significance, amount of evidence, consistency, and hints of bias. We also assess the concordance of observational epidemiological data

Table 1 The seven standardized criteria

Levels of evidence	Description
Convincing	• Associations with a statistical significance at $P < 10^{-6}$ • More than 1000 cases included (or more than 20,000 participants for continuous outcomes) • The largest component study reporting a significant result at $P < 0.05$ • A 95% prediction interval that excluded the null • Absence of large heterogeneity (I2<50%) • No evidence of small study effects ($P > 0.10$) • No evidence of excess significance ($P > 0.10$)
Highly suggestive	• Associations with a statistical significance at $P < 10^{-6}$ • More than 1000 cases included (or more than 20,000 participants for continuous outcomes) • The largest component study reporting a significant result at $P < 0.05$.
Suggestive	• Associations with a statistical significance at $P < 0.001$ • More than 1000 cases included (or more than 20,000 participants for continuous outcomes).
Weak	• Associations with a statistical significance at $P < 0.05$

Previous umbrella reviews have used various criteria to assess the evidence from meta-analysis of observational epidemiological studies. The combination of these criteria allows to tentatively classify evidence from meta-analyses of statistically significant risks and protective factors into four levels described below. A more detailed description of the criteria can be found in Additional file 1: Appendix Method 1

against corresponding meta-analyses of RCTs and MR studies.

Methods

Data sources and searches

We systematically searched PubMed, up to November 19, 2020, for studies labeled as umbrella reviews in their title: umbrella [Title] AND review [Title]. The protocol has been registered on the Open Science Framework [8].

Study selection

All umbrella reviews including meta-analyses of observational studies assessing putative risk or protective factors were eligible. We considered all putative factors (i.e., any attributes, characteristics, or exposure of an individual [9] that may either increase or decrease the occurrence of any type of health outcomes). Umbrella reviews not assessing any putative risk or protective factors in observational settings or not using any of the seven previously proposed standardized criteria (Table 1) to assess the evidence were excluded. One author (PJ) screened all resulting articles from the literature search for inclusion criteria and consulted with a second author (JPA) when in doubt. If two or more umbrella reviews had 50% of their associations (i.e., a putative risk or protective factor with a health outcome) assessed overlapping, we retained the one with the largest number of associations.

Data extraction

At the umbrella reviews level, we abstracted data regarding study design (observational studies alone or combined with other study types); number of factors evaluated by study design, when available; methodological quality tool used (e.g., AMSTAR [10]); and method used to evaluate the evidence (i.e., the seven standardized criteria [3–6] or other method).

We then extracted the following data for each meta-analysis included in each umbrella review examining the association of a putative risk or protective factor with a health outcome: exposure, outcome, study designs included (e.g., cohort or case-control studies), number of included studies, participants, metric used (e.g., odds ratio, risk ratio, hazard ratio, mean difference, standardized mean difference), and data necessary for the evaluation of the pre-specified seven standardized criteria (Table 1). Data extraction was repeated limited to data from prospective cohort studies. When cohorts were mentioned, without specification of whether these were prospective or retrospective, we kept these data and then excluded them in separate sensitivity analyses.

For umbrella reviews that also separately considered RCTs and MR studies besides the observational association studies, we extracted the effect size and corresponding 95% confidence interval, total number of participants, number of cases/events, and genetic instruments used for MR studies. We did not perform any new quality assessment and relied on the ones performed by umbrella review authors. Two authors (PJ and AA) independently extracted 20% of the included umbrella reviews, while the rest was split between them. Discrepancies were resolved through consensus.

Data synthesis and analysis

We started by reassessing the evidence for each association using the pre-specified list of criteria presented in Table 1 (more details in Additional file 1). In case of missing data, the criterion was considered as failed. The number and proportions of associations fulfilling each criterion and meeting the different levels of evidence (i.e., convincing, highly suggestive, suggestive, and weak) based on their combination (Table 1) were counted for each umbrella reviews (also labeled as topic). For each level of evidence, proportions were summarized across umbrella reviews using the restricted maximum likelihood random effects model meta-analysis and the arcsine transformation to normalize and stabilize the variance [11]. Similarly, proportions were summarized for each criterion but focusing solely on statistically significant associations (those with $P < 0.05$ for the random effects summary effect). The between-umbrella heterogeneity was estimated using I^2 [12].

The concordance between the 7 criteria was assessed by Cohen's kappa (κ), where a κ<0.6 represents weak agreement [13, 14]. First, we estimated the different κ across all umbrella reviews, including only statistically significant associations (those with $P < 0.05$ for the random effects summary effect). We then estimated the different κ and their corresponding confidence intervals within each umbrella review and combined them using random effects [15].

In previously published umbrellas, when all 7 criteria are met ($P < 10^{-6}$, ≥1000 cases (or ≥20,000 participants for continuous factors), $P < 0.05$ in the largest study, 95% prediction interval excluding the null [16, 17], and no large between-study heterogeneity, small study effects [18–20], or excess significance [21–23]), the evidence has been called "convincing" [3–6] since there is strong statistical support, large amount of evidence, consistency, and no overt signals in the bias tests. We should acknowledge, however, that there is no gold standard of what constitutes a genuine risk or protective factor. Some convincing associations may have some other problem in their evidence that invalidates them. Conversely, other associations that are not mapped as convincing may well be true. Allowing for this uncertainty, we tried to address which criteria were the most constraining to reach a convincing level of evidence, as each criterion was separately removed from being

required to have an association called convincing. This analysis was performed only on statistically significant associations for which information on all seven criteria was available. Numbers of additional associations reaching a convincing level of evidence were recorded. In addition to testing the different criteria, different statistical thresholds ($P < 0.001$ and <0.05) were also tested as alternatives to the original convincing level of statistical significance ($P < 10^{-6}$). We also recorded how evidence was impacted by restricting the assessment of associations to prospective cohorts.

For associations assessed both by meta-analyses of observational studies and by either meta-analyses of RCTs or MR studies, we compared the effect sizes and corresponding 95% confidence intervals. The estimates across different designs were paired according to outcome, exposure, comparison, and population. For RCTs, if there were more than one meta-analysis for the same topic, we retained the one with the largest number of studies included. For MR studies in case of multiple studies for one observational association, each study was compared with the corresponding meta-analysis of observational studies. We specifically examined if the direction and statistical significance of the associations were concordant with the direction and statistical significance of effects in meta-analyses of RCTs and MR studies. We considered the traditional $P < 0.05$ threshold of statistical significance and also the more recently adopted $P < 0.005$ [24].

Moreover, to investigate whether the difference between the meta-analyses estimates was beyond chance, Q tests were performed ($P < 0.10$) [25]. For ease of interpretation, we converted all weighted mean differences (WMDs) and standardized mean difference (SMDs) to odds ratio (OR) equivalents [26] and assumed that relative risks (RRs) and hazard ratios (HRs) were interchangeable with ORs (a reasonable assumption for mostly rare event rates and for a minority where event rates are substantial, the OR is substantially larger than the RR). We also checked how often OR estimates using the different designs differed by two-fold or more.

For factors with statistically significant results both in observational as well as RCTs or MR studies' evidence (and thus have the most consistent support), we recorded the pattern of the seven pre-specified criteria in the meta-analyses of observational epidemiological data.

Results

Eligible umbrella reviews and meta-analyses of observational associations

The literature search yielded 449 articles of which 180 umbrella reviews were potentially eligible. Of those, 123 umbrella reviews were excluded as they had limited or inadequately reported data available, and did not use the seven standardized criteria to assess the evidence of their included associations or reported associations overlapped by over 50% with another umbrella review (Fig. 1).

Fifty-seven umbrella reviews including 3744 associations assessed by meta-analyses of observational studies were included [5, 6, 27–81] (Fig. 2 and Table 2). The median number of estimates included in each meta-analysis was 7 (IQR 4 to 11) ranging from a minimum of 2 up to 309 estimates.

Assessment according to a set of 7 pre-specified quantitative criteria

Overall, 99 (2.6%) associations were graded as convincing, 253 (6.7%) as highly suggestive, 440 (11.8%) as suggestive, and 1497 (40.0%) as weak and 1455 (38.9%) were not statistically significant at $P < 0.05$ (Fig. 2). Meta-analyses of the proportions of convincing and highly suggestive associations across the 57 topics resulted in 1.3% (95% CI [1.0–2.2%]) summary proportion for convincing and 4.6% (95% CI [2.9–6.6%]) summary proportion for highly suggestive associations, and both had very high between-topic heterogeneity ($I^2 = 73.9\%$ and $I^2 = 85.7\%$, respectively) (Table 3 and Additional file 2: Figures 1 to 5). Convincing associations varied from 0 to 16.7% across topics and 29/57 umbrella reviews had no associations with convincing evidence [6, 29, 36, 38, 40, 41, 45, 47, 50, 52, 53, 56, 58, 62, 63, 65, 66, 68–72, 74–76, 78–81] (Table 2). Moreover, the number of non-statistically significant associations (those with P≥0.05) varied substantially between topics from 0% for the associations of depression with mortality outcomes, antipsychotics with life-threatening events, and health factors with loneliness [40, 55, 68] to 80.7% for risk factors of prostate cancer [41].

41.4% (1549/3744) of the associations had at least one missing criterion (Additional file 3: Figures 6 to 7). An additional 25 and 82 associations would have reached a convincing and highly suggestive level of evidence, respectively, if missing criteria were considered to be satisfied.

We performed meta-analyses for the proportion of associations that met each of the 7 pre-specified quantitative criteria across the 57 topics, limiting to the 2289 statistically significant associations. Only 29% (95% CI [24.9–33.3%]) of the associations had $P < 10^{-6}$. Conversely, 74.9% (95% CI [71.2–78.4%]) of the associations had the largest study with $P < 0.05$, and 75.3% (95% CI [72.2–78.3%]) and 77.7% (95% CI [72.6–82.5%]) of the associations with available data had no signals of small-study effects or excess significance, respectively. Between-topic heterogeneity for the presence of each criterion was typically high (Table 3 and Additional file 4: Figures 8 to 15).

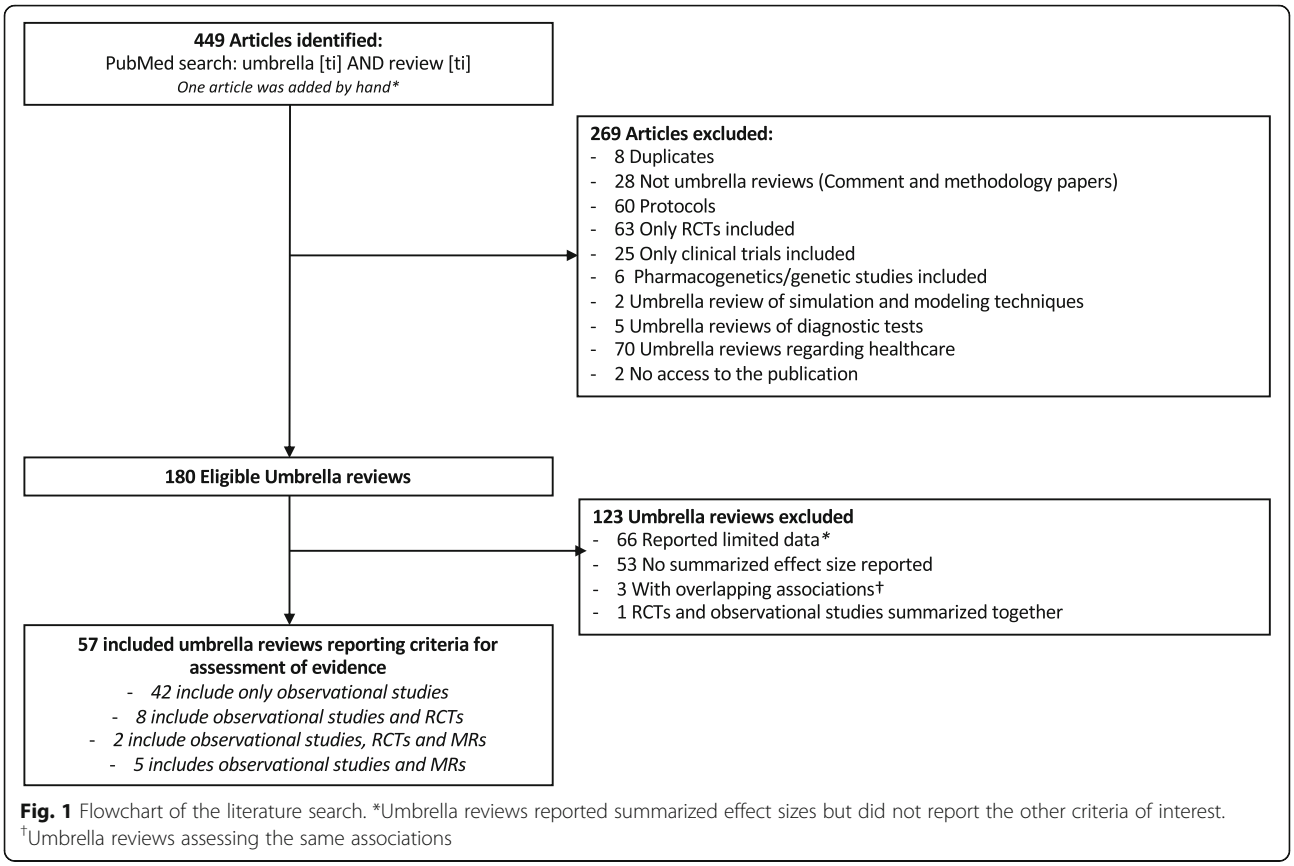

Fig. 1 Flowchart of the literature search. *Umbrella reviews reported summarized effect sizes but did not report the other criteria of interest. †Umbrella reviews assessing the same associations

Concordance between the 7 pre-specified quantitative criteria

There was a limited concordance between the different criteria, meaning that they provide mostly independent information (Fig. 3). Excluding the kappa coefficients for the concordance of different *P*-value thresholds, a weak to moderate concordance existed only between prediction intervals excluding the null and $P < 10^{-6}$ (κ=0.44) (Additional file 5: Table 1 and Additional file 6: Figures 16 to 43).

Impact of each pre-specified criterion on number of associations deemed to have convincing evidence

1457 statistically significant associations ($P < 0.05$) had information available on all 7 criteria. Replacing the *P*-value threshold of $<10^{-6}$ by <0.001 as a requirement for convincing evidence, convincing associations increased from 6.8 to 9.2% and increased even further to 9.7% when the threshold was set at <0.05 (Table 4). The most constraining criterion appeared to be the absence of large heterogeneity ($I^2 > 50\%$); removing it increased the number of convincing associations to 10.9%.

Analyses limited to prospective cohort studies

We were able to isolate 1141 associations which included only cohorts or where it was possible to separate cohort studies from other designs. Out of the 1141 associations, 849 were assessed by an unspecified mix of prospective and retrospective cohorts with no means to distinguish them from one another, 126 only by prospective cohort studies, 25 only by retrospective cohort studies, and 141 by a mix of study designs but where it was possible to separate the prospective cohort studies from the other designs. Across the 1141 associations, when limited to cohort studies, convincing associations decreased slightly from 5.7% ($n = 65$) to 4.2% ($n = 48$), and highly suggestive associations decreased from 13.5% ($n = 154$) to 11.7% ($n = 133$) (Fig. 2 and dataset available on the Open Science Framework [82]).

Comparison against RCTs and MR studies

Only 16 out of the 57 umbrella reviews also investigated evidence from RCTs, [36, 37, 44, 45, 47, 52, 56, 69, 77], MR studies [27, 33, 38, 62, 73], or both [6, 71] in addition to observational studies. Of those 16, 5 had no overlapping associations between the different study designs [37, 44, 47, 52, 77] and one only provided a narrative summary of MR studies [62]. For 121 of the 882 observational associations evaluated in the 16 included umbrella reviews, evidence from 62 meta-analyses of RCTs or 60 MR studies could be juxtaposed; of note, one association was assessed both by a meta-analyses of

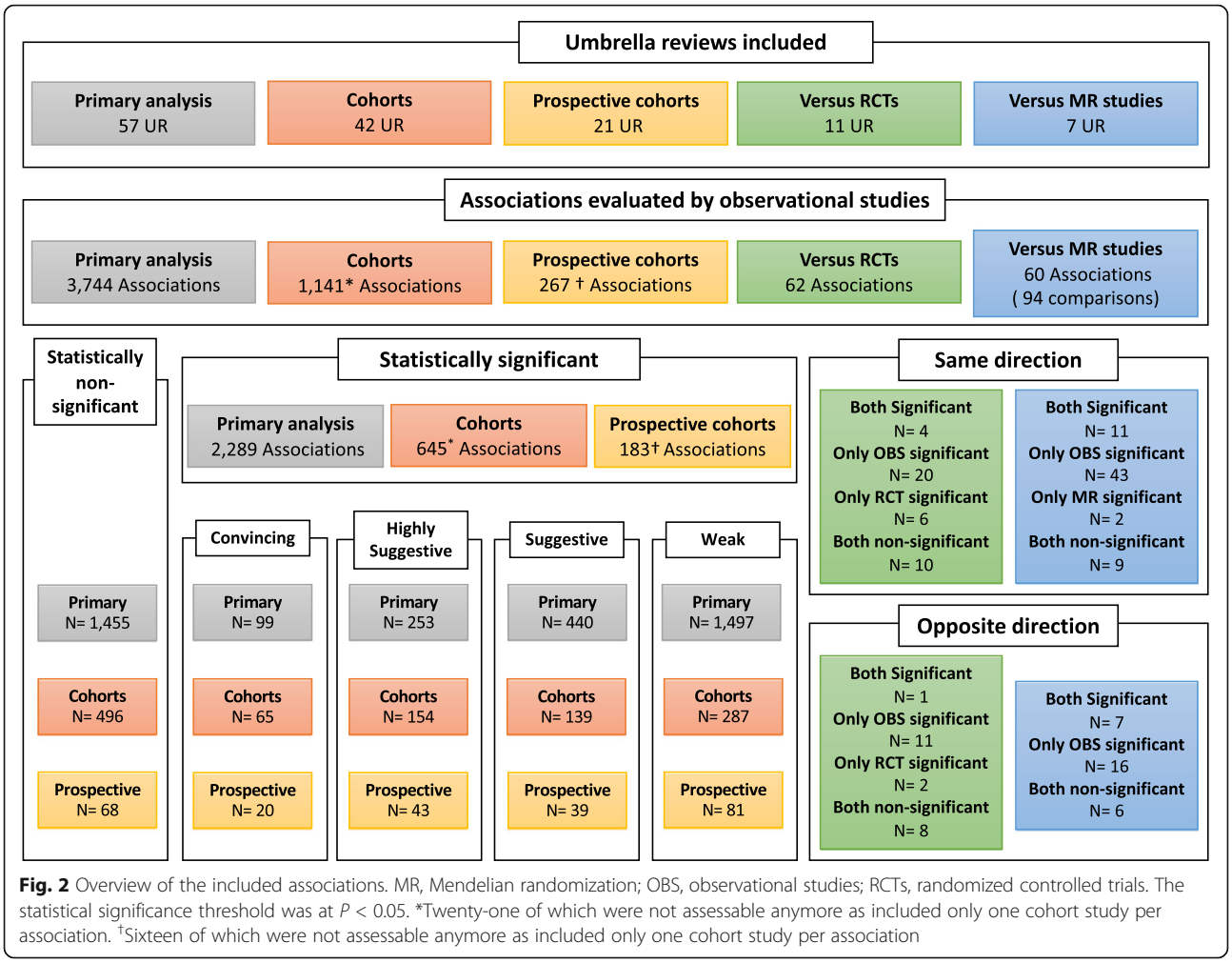

Fig. 2 Overview of the included associations. MR, Mendelian randomization; OBS, observational studies; RCTs, randomized controlled trials. The statistical significance threshold was at $P < 0.05$. *Twenty-one of which were not assessable anymore as included only one cohort study per association. †Sixteen of which were not assessable anymore as included only one cohort study per association

RCTs and by a MR study. Nine observational associations were assessed by more than one MR study using different genetic instruments, thus resulting in a total of 94 comparisons. Results are presented in Fig. 2 and Additional file 7: Table 2 and Additional file 8: Table 3.

When comparing meta-analyses of observational studies against meta-analyses of RCTs, half of the associations (31/62) were only statistically significant in meta-analyses of observational studies (at the P < 0.05 level), while eight were only statistically significant in meta-analyses of RCTs. Four estimates were statistically significant with point estimates in the same direction for both types of design. Conversely, for one association, the point estimates were statistically significant, but in different direction, statins significantly increased the risk of pancreatitis (OR= 1.41, 95% CI [1.15; 1.74], $P = 0.04$) when limiting the evidence to the meta-analysis of observational studies but the risk was decreased in the meta-analysis of RCTs (OR=0.77, 95% CI [0.61; 0.97]) (36). Overall, 37.1% (23/62) of the estimates showed point estimates in the opposite direction in observational and RCT meta-analyses (Additional file 7: Table 2). When the $P < 0.005$ level was used, only two associations were statistically significant in both meta-analyses of observational studies and RCTs. The differences between the meta-analyses estimates of observational studies and RCTs were beyond chance for 43.5% (27/62) associations ($P < 0.10$ for the χ^2 Q test) and 12.5% (8/64) differed in their effect sizes by two-fold or more in the two designs (Additional file 9: Figure 44 to 45).

Of 94 comparisons between meta-analyses of observational studies and MR studies, 62.8% (59/94) were solely statistically significant in observational studies (at the $P < 0.05$ level). Eleven comparisons showed a statistically significant evidence in both study designs with point estimates in the same direction. However, seven comparisons resulted in discordant results with statistically significant point estimates in opposite direction. Overall 30.8% (29/94) comparisons had point estimates in opposite directions between meta-analyses of observational studies and MR studies. Between MR studies, differences in the direction of the point estimates were also noted.

Table 2 Overview of the included umbrella reviews

Topic	First author	Year	Type of studies	N total studies [a]	Median [IQR] Min to max.[b]	N total associations[c]	N associations included	Convincing	Highly suggestive	Suggestive	Weak	Non-significant
One exposure with multiple outcomes												
Adiposity and cancer outcomes	Kyrgiou	2017	OBS	507	6 [4; 11.50] (2 to 44)	194	67[d]	8 (11.9%)	14 (20.9%)	15 (22.4%)	11 (16.4%)	19 (28.4%)
Antidepressant and adverse events	Dragioti	2019	OBS	1012	6 [4; 11.25] (2 to 44)	120	120	3 (2.5%)	8 (6.7%)	24 (20%)	38 (31.7%)	47 (39.2%)
Antipsychotics and life-threatening events	Papola	2019	OBS	68	9 [6.75; 12.75] (6 to 24)	6	6	1 (16.7%)	2 (33.3%)	3 (50%)	0 (0%)	0 (0%)
Low-dose aspirin and health outcomes	Veronese	2020	OBS-RCT	NR	3.50 [3; 7.75] (2 to 32)	156	42	0 (0%)	0 (0%)	0 (0%)	11 (26.2%)	31 (73.8%)
Aspirin and cancer outcomes	Song	2020	OBS	NR	11 [7; 18.25] (3 to 309)	27	18[e]	0 (0%)	0 (0%)	0 (0%)	12 (66.7%)	6 (33.3%)
Birth weight and later life events	Belbasis	2016	OBS	NR	10 [6.25; 16] (3 to 45)	78	78	3 (3.8%)	8 (10.3%)	10 (12.8%)	29 (37.2%)	28 (35.9%)
Chocolate and health outcomes	Veronesem	2018	OBS-RCT	NR	5 [4; 6] (4 to 6)	19	7	0 (0%)	0 (0%)	0 (0%)	4 (57.1%)	3 (42.9%)
Chronic kidney disease and mortality	Kim	2020	OBS-RCT	NR	9 [4; 14] (2 to 26)	105	49	0 (0%)	0 (0%)	22 (44.9%)	11 (22.4%)	16 (32.7%)
Coffee and cancer risk	Zhao	2020	OBS	448	15 [7; 21.25] (4 to 54)	36	36	0 (0%)	3 (8.3%)	7 (19.4%)	7 (19.4%)	19 (52.8%)
C-reactive protein and health outcomes	Markozannes	2020	OBS-MR	952	6 [4; 11] (3 to 53)	309	113	2 (1.8%)	12 (10.6%)	14 (12.4%)	67 (59.3%)	18 (15.9%)
Depression and mortality	Machado	2018	OBS	246	6 [4; 12] (3 to 111)	17	17	0 (0%)	4 (23.5%)	2 (11.8%)	11 (64.7%)	0 (0%)
Antidepressants during pregnancy and neonatal outcomes	Biffi	2020	OBS	NR	7 [4; 10] (3 to 28)	69	69	0 (0%)	5 (7.2%)	11 (15.9%)	18 (26.1%)	35 (50.7%)
Dietary Fiber and health outcomes	Veronese	2018	OBS	NR	12 [6; 19] (3 to 26)	21	21	2 (9.5%)	1 (4.8%)	9 (42.9%)	6 (28.6%)	3 (14.3%)
Fish and ω-3 Fatty Acids consumptions and cancer outcomes	Lee	2020	OBS	NR	5 [3; 10] (2 to 17)	57	52[f]	0 (0%)	0 (0%)	2 (3.8%)	10 (19.2%)	40 (76.9%)
Influenza vaccine in elderly and health outcomes	Demurtas	2020	OBS-RCT	NR	3.50 [2; 6.75] (2 to 27)	60	38	1 (2.6%)	3 (7.9%)	6 (15.8%)	15 (39.5%)	13 (34.2%)

Table 2 Overview of the included umbrella reviews *(Continued)*

Topic	First author	Year	Type of studies	N total studies [a]	Median [IQR] Min to max.[b]	N total associations[c]	N associations included	Convincing	Highly suggestive	Suggestive	Weak	Non-significant
Handgrip strength and health outcomes	Soysal	2020	OBS	NR	9 [8; 10.50] (7 to 34)	11	8[e]	0 (0%)	1 (12.5%)	1 (12.5%)	4 (50%)	2 (25%)
Human immunodeficiency virus infections and health outcomes	Grabovac	2019	OBS	NR	8 [4; 13.50] (2 to 43)	55	55	0 (0%)	0 (0%)	9 (16.4%)	30 (54.5%)	16 (29.1%)
Metformin and cancer outcomes	Yu	2019	OBS	327	7 [5; 15] (2 to 29)	33	33	1 (3%)	3 (9.1%)	5 (15.2%)	14 (42.4%)	10 (30.3%)
Magnesium and health outcomes	Veronese	2019	OBS-RCT	NR	6 [3; 9.50] (3 to 32)	55	19	0 (0%)	0 (0%)	2 (10.5%)	7 (36.8%)	10 (52.6%)
Obesity and gynecology/obstetric outcomes	Kalliala	2017	OBS-RCT	427	6 [3; 9] (2 to 40)	248	144	11 (7.6%)	28 (19.4%)	23 (16%)	42 (29.2%)	40 (27.8%)
Physical activity and cancer outcomes	Rezende	2017	OBS	297	6 [3.25; 10] (2 to 38)	46	46	1 (2.2%)	2 (4.3%)	5 (10.9%)	5 (10.9%)	33 (71.7%)
Physical activity and atrial fibrillation outcomes	Valenzuela	2020	OBS	NR	8 [7; 19] (6 to 20)	5	5	0 (0%)	0 (0%)	0 (0%)	3 (60%)	2 (40%)
Statins and multiple non-cardiovascular outcomes	Yazhou	2018	OBS-RCT	NR	6 [4; 9] (2 to 27)	278	115	0 (0%)	2 (1.7%)	21 (18.3%)	42 (36.5%)	50 (59.8%)
Serum uric acid and health outcomes	Li	2017	OBS-RCT-MR	NR	5 [3; 9] (2 to 31)	152	76	0 (0%)	7 (9.2%)	9 (11.8%)	41 (53.9%)	19 (25%)
Type 2 diabetes mellitus and cancer	Tsilidis	2015	OBS	474	14 [9; 21] (5 to 45)	27	27	2 (7.4%)	4 (14.8%)	4 (14.8%)	10 (37%)	7 (25.9%)
Tea consumption and cancer	Kim	2020	OBS	NR	10 [6.75; 16] (4 to 53)	150	68[g]	0 (0%)	1 (1.5%)	2 (2.9%)	18 (26.5%)	47 (69.1%)
Telomere length and health outcomes	Smith	2019	OBS	NR	5.50 [3; 8] (2 to 20)	50	50	0 (0%)	1 (2%)	0 (0%)	23 (46%)	26 (52%)
Vitamin D and health outcomes	Theodoratou	2014	OBS-RCT	NR	7 [5; 10] (2 to 37)	48	48	0 (0%)	6 (12.5%)	7 (14.6%)	16 (33.3%)	19 (39.6%)
Multiple exposures with one outcome												
Risk factor for attention deficit hyperactivity disorder	Kim	2020	OBS	NR	6 [4; 9] (2 to 30)	63	63	5 (7.9%)	3 (4.8%)	11 (17.5%)	26 (41.3%)	18 (28.6%)
Risk and protective factors for mental disorders with onset in childhood/adolescence	Marco	2020	OBS	192	6 [4.50; 9] (2 to 26)	23	23	0 (0%)	0 (0%)	1 (4.3%)	8 (34.8%)	14 (60.9%)
Environmental factors and serum biomarkers for atrial fibrillation	Belbasis	2020	OBS	NR	6 [4; 8] (3 to 31)	51	51	6 (11.8%)	11 (21.6%)	8 (15.7%)	10 (19.6%)	16 (31.4%)
Factors associated to loneliness	Solmi	2020	OBS	NR	13 [8; 18] (3 to 31)	5	5	0 (0%)	0 (0%)	1 (20%)	4 (80%)	0 (0%)

Table 2 Overview of the included umbrella reviews *(Continued)*

Topic	First author	Year	Type of studies	N total studies [a]	Median [IQR] Min to max.[b]	N total associations[c]	N associations included	Convincing	Highly suggestive	Suggestive	Weak	Non-significant
Risk factor for amyotrophic lateral sclerosis	Belbasis	2016	OBS	NR	8 [5.75; 9.25] (3 to 20)	16	16	0 (0%)	0 (0%)	3 (18.8%)	6 (37.5%)	7 (43.8%)
Risk and protective factors for anxiety and obsessive compulsive disorders	Fullana	2019	OBS	216	3 [2; 6] (2 to 112)	427	128[fg]	4 (3.1%)	2 (1.6%)	3 (2.3%)	60 (46.9%)	59 (46.1%)
Environmental risk factors and biomarkers for autism spectrum disorder	Kim	2019	OBS	NR	8 [3.50; 13] (2 to 24)	67	67	8 (11.9%)	7 (10.4%)	11 (16.4%)	26 (38.8%)	15 (22.4%)
Environmental risk factors for bipolar disorder	Bortolato	2017	OBS	54	8 [5; 10] (3 to 13)	7	7	1 (14.3%)	1 (14.3%)	2 (28.6%)	2 (28.6%)	1 (14.3%)
Risk factors for colorectal cancer metastasis and recurrence	Xu	2020	OBS	NR	6 [3.50; 9] (2 to 41)	47	47	0 (0%)	0 (0%)	4 (8.5%)	27 (57.4%)	16 (34%)
Non-genetic biomarkers and colorectal cancer risk	Zhang	2020	OBS-RCT-MR	NR	7 [3; 10] (2 to 28)	112	65	0 (0%)	0 (0%)	4 (6.2%)	25 (38.5%)	36 (55.4%)
Risk factors for chronic obstructive pulmonary disease	Bellou	2019	OBS-MR	NR	5 [4; 11] (3 to 22)	22	18	0 (0%)	0 (0%)	8 (44.4%)	5 (27.8%)	5 (27.8%)
Environmental risk factors for dementia	Bellou	2017	OBS	NR	7 [4.75; 13] (3 to 43)	76	76	7 (9.2%)	5 (6.6%)	10 (13.2%)	33 (43.4%)	21 (27.6%)
Risk factors for depression	Kohler	2018	OBS-MR	NR	7.50 [5; 11] (3 to 77)	140	134	0 (0%)	0 (0%)	41 (30.6%)	57 (42.5%)	36 (26.9%)
Risk factors for eating disorders	Solmi	2020	OBS	NR	6 [4; 9] (2 to 33)	49	49	0 (0%)	0 (0%)	6 (12.2%)	35 (71.4%)	8 (16.3%)
Risk factors for endometrial cancer	Raglan	2019	OBS	604	4 [3; 6] (2 to 28)	127	127	3 (2.4%)	13 (10.2%)	14 (11%)	26 (20.5%)	71 (55.9%)
Environmental risk factors for obesity	Solmi	2018	OBS-RCT	166	8 [6; 10.75] (2 to 22)	60	26	4 (15.4%)	2 (7.7%)	1 (3.8%)	15 (57.7%)	4 (15.4%)
Prognostic biomarkers for gastric cancer	Zhou	2019	OBS	>1000	7 [4; 11] (3 to 51)	119	119	3 (2.5%)	7 (5.9%)	3 (2.5%)	82 (68.9%)	24 (20.2%)
Risk factors for gestational diabetes	Giannakou	2019	OBS	NR	8 [5; 14] (3 to 40)	61	61	1 (1.6%)	13 (21.3%)	9 (14.8%)	28 (45.9%)	10 (16.4%)
Peripheral biomarkers and major mental disorders	Carvalho	2020	OBS	NR	7 [5; 13] (3 to 55)	358	318[g]	0 (0%)	0 (0%)	3 (0.9%)	175 (55%)	140 (44%)
Environmental risk factors for multiple sclerosis	Belbasis	2015	OBS	NR	8 [6; 12] (3 to 30)	44	44	2 (4.5%)	2 (4.5%)	2 (4.5%)	17 (38.6%)	21 (47.7%)
Prognostic biomarkers for pancreatic ductal adenocarcinoma	Wang	2020	OBS	>300	4 [3; 6] (2 to 43)	63	63	0 (0%)	2 (3.2%)	1 (1.6%)	41 (65.1%)	19 (30.2%)

Table 2 Overview of the included umbrella reviews *(Continued)*

Topic	First author	Year	Type of studies	N total studies [a]	Median [IQR] Min to max.[b]	N total associations[c]	N associations included	Convincing	Highly suggestive	Suggestive	Weak	Non-significant
Environmental risk factors and Parkinson's	Bellou	2016	OBS	755	7 [5; 10] (2 to 67)	75	75	2 (2.7%)	6 (8%)	9 (12%)	18 (24%)	40 (53.3%)
Risk and protective factors for prostate cancer	Markozannes	2016	OBS	1907	5 [3.75; 7] (2 to 45)	248	176[d]	0 (0%)	2 (1.1%)	7 (4%)	25 (14.2%)	142 (80.7%)
Non-genetic risk factors for pre-eclampsia	Giannakou	2017	OBS	NR	7 [4; 12.25] (3 to 34)	130	64[h]	1 (1.6%)	11 (17.2%)	5 (7.8%)	22 (34.4%)	25 (39.1%)
Risk and protective factors for psychosis	Ruada	2018	OBS	683	6 [3; 9] (2 to 55)	170	128[i]	1 (0.8%)	2 (1.6%)	11 (8.6%)	64 (50%)	50 (39.1%)
Environmental risk factors for rheumatic diseases	Belbasis	2018	OBS	NR	10.50 [7; 13] (3 to 51)	42	42	0 (0%)	0 (0%)	7 (16.7%)	26 (61.9%)	9 (21.4%)
Risk factors and peripheral biomarkers for schizophrenia spectrum disorders	Belbasis	2017	OBS-MR	NR	8 [5.25; 13] (3 to 42)	98	98	1 (1%)	4 (4.1%)	5 (5.1%)	52 (53.1%)	36 (36.7%)
Non-genetic risk factors for skin cancer	Belbasis	2016	OBS	NR	10 [7; 18] (3 to 41)	85	85	4 (4.7%)	9 (10.6%)	11 (12.9%)	34 (40%)	27 (31.8%)
Risk factors for type 2 diabetes mellitus	Bellou	2018	OBS-MR	NR	9.50 [6; 14] (3 to 88)	155	142	11 (7.7%)	34 (23.9%)	28 (19.7%)	43 (30.3%)	26 (18.3%)

IQR interquartile range, *MA* meta-analyses, *MR* Mendelian randomization, *NR* not reported, *OBS* observational studies, *OCD* obsessive and compulsive disorders, *RCT* randomized controlled trials

[a]Total number of primary studies included in the meta-analyses assessed by the umbrella reviews

[b]Median number IQR and minimum and maximum number of primary studies included in the associations assessed

[c]Total number of associations assessed in the included umbrella reviews

[d]The umbrella review presented data for continuous and binary outcomes but their principal analyses focused only on continuous outcomes which we included hence the lowest number of included associations in our work compared with the original umbrella review

[e]Some associations were excluded as they were assessed by a mix of RCTs and observational studies

[f]Associations assessed by only one study were removed

[g]Duplicated associations were excluded

[h]Excluded associations assessing genetic factors

[i]The authors of the umbrella review mention 170 associations but only report 145. Out of the 145, 17 meta-analyses were excluded because included only one study

For example, no significant associations were shown in MR studies between smoking and depression; however, the observational studies showed a significant increased risk of depression in smokers (OR= 1.68, 95% CI [1.55; 1.82]). All MR studies' point estimates were in the same direction (increased risk) except for one (OR=0.85, 95% CI [0.66; 1.1]) [38] (Additional file 8: Table 3). When using the $P < 0.005$ level for claiming statistical significance, only seven out of 18 associations remained statistically significant in MR studies. When comparing the meta-analysis' effects in observational studies versus the MR studies, there were significant heterogeneity (P < 0.10 for the χ^2 Q test) between the two designs for 54.7% (54/94) comparisons and 12 (12.8%) differed by two-fold or more in their effect sizes (Additional file 9: Figures 44 to 45).

Overall, only four associations assessed by observational studies and RCTs and another three comparisons assessed by observational and MR studies had consistently statistically significant results ($P < 0.05$) in the same direction. Of these seven associations, the seven pre-specified criteria had graded two of them as highly suggestive, two as suggestive and three as weak.

Discussion

We assessed the evidence obtained from observational studies for associations on 3744 putative risk and protective factors assessed by a median of 7 (IQR 4 to 11) estimates per meta-analysis from 57 umbrella reviews on diverse topics. Although the majority (61.1%) of the investigated associations were statistically significant at the traditional $P < 0.05$ level, only 2.6% and 6.7% were classified as having convincing or highly suggestive evidence, respectively, using a set of pre-specified criteria that have been used in the literature of umbrella reviews [3–6]. The proportions of associations meeting the various pre-specified criteria of statistical significance, amount of evidence, consistency, and lack of hints for bias and reaching different level of evidence varied across topics. Variability was highly prominent for the proportion of probed associations that had non-statistically significant (P≥0.05) results (0–80.7%).

The seven criteria that have been previously used to assess evidence from meta-analyses of observational associations have been developed ad hoc [3–6] aiming to capture sufficient statistical support, amount of evidence, consistency, and lack of signals that may herald bias [12, 20, 21]. It is unknown how well they can really identify convincing/strong evidence, let alone causality. A perfect gold standard is missing for causality in observational associations. Nevertheless, we could assess here the performance of these criteria against each other. They mostly showed low concordance among themselves and thus may offer relatively independent, complementary insights into the evidence of an observational association. Most associations did not offer any signal of small-study effects and excess significance. However, these results are to be interpreted with caution since both tests are not definite proof of presence or absence of bias; given the typically small or modest number of studies in each meta-analysis the power of these tests is very limited [83]. Conversely, substantial evidence of heterogeneity was common, with most meta-analyses of observational associations presenting I^2 estimates exceeding 50%. Heterogeneity was also the most constraining criterion. When removed from the list of criteria to reach a convincing level of evidence, the number of associations increased substantially. Heterogeneity in meta-analyses of observational studies may be due to bias but also genuine difference between studies [5]. It is often hard to detangle between the two.

It is important to acknowledge the limits of our proposed criteria and of the ways that they can be combined to reach an overall grading. *P*-value thresholds are set arbitrarily, the random effects meta-analysis may produce inconsistent results [84], the excess significance of bias has limited power if only a few studies are statistically significant [21, 22], and similarly both small-study effect and excess significance testing may be misleading when there is substantial heterogeneity [85]. Even if all 7 criteria are fulfilled, observational evidence could still remain at risk of unmeasured confounding, undetected bias, and reverse causality [6]. One illustration would be the downgrading of the evidence for associations for which we re-analyzed the data using only cohort studies.

Furthermore, we should acknowledge that different types of observational associations vary a lot in prior plausibility and thus the amount of statistical support that is required to make them convincing is likely to vary. Fields like pharmacoepidemiology might be very reluctant to adopt a *P*-value threshold of $P < 10^{-6}$ for signal detection of medication harms. In agnostic searches, conversely, even such *P*-value thresholds may not be low enough [86]. Field-specific setting of *P*-value thresholds has been proposed, e.g. through empirical calibration [87, 88], but such calibrations are still unspecified and lack consensus for the vast majority of fields in epidemiology.

Most decision-makers have required evidence of causality for interventions, but licensing based on observational evidence alone is becoming increasingly common [89, 90]. While discordant results between RCTs and observational studies were highlighted long ago [1, 91, 92], there is ongoing debate on whether overall there are big differences and even on whether these designs can be formally compared when the same factor/intervention is involved [93, 94]. Most of the evidence that has been systematically assessed to-date pertains to situations

Table 3 Meta-analyses of the proportions of associations for each criterion and level of evidence (random effects)

	n/N associations (crude proportion)	Proportions [95% CI] from meta-analysis of 57 topics[a]	I^2	Range of proportions across topics
Level of evidence:				
Convincing	99/3744 (2.6%)	1.3% [1.0%; 2.2%]	73.9%	0–16.7%
Highly suggestive	253/3744 (6.7%)	4.6% [2.9%; 6.6%]	85.7%	0–33.3%
Suggestive	440/3744 (11.8%)	11.0% [8.5%; 13.8%]	83.9%	0–50%
Weak	1497/3744 (40.0%)	39.1% [34.8%; 43.5%]	86.2%	0–71.4%
Non-significant	1455/3744 (38.9%)	34.7% [29.2%; 40.3%]	90.8%	0–80.7%
Criteria:				
Statistical significance				
$P < 10^{-6}$	762/2289 (33.3%)	29.0% [24.9%; 33.3%]	74.8%	0–66.7%
$P < 0.001$	1377/2289 (60.2%)	58.6% [54.1%; 63.0%]	73.3%	0–100%
Cases > 1000 (or > 20,000 participants for continuous outcomes)	1182/2107 (56.1%)	65.3% [56.9%; 73.2%]	94.9%	1.7–100%
Largest study with $P < 0.05$	1343/1781 (75.4%)	74.9% [71.2%; 78.4%]	63.6%	28.6–100%
95% prediction interval that excluded the null	642/2136 (30.1%)	30.3% [26.5%; 34.2%]	71.0%	9.0–100%
Absence of large heterogeneity (I^2<50%)	1050/2277 (46.1%)	46.6% [41.8%; 51.3%]	79.5%	0–88.2%
No evidence of small study effects ($P > 0.10$)	1628/2164 (75.2%)	75.3% [72.2%; 78.3%]	63.2%	40–100%
No evidence of excess significance ($P > 0.10$)	1599/2052 (77.9%)	77.7% [72.6%; 82.5%]	83.0%	33.3–100%

[a]Meta-analyses for the individual criteria excluded associations with missing data. Meta-analyses for the levels of evidence were conducted across all 3744 associations regardless of their statistical significance status. The meta-analyses for the individual criteria were conducted across the 2289 statistically significant associations. Out of 2289, statistically significant associations, 182 associations did not report on the number of cases,508 on whether the largest study had $P < 0.05$, 153 on the 95% prediction interval, 12 on the I^2 for heterogeneity, 125 on the small study effect test, and 237 on the excess of significance bias. Data on all 7 criteria were available for 1457 statistically significant meta-analyses

where therapeutic interventions are assessed [2]. On average, the two designs may give similar results [2], but single comparisons may deviate substantially in the effect size estimates and in some settings even average effects seem to differ markedly across designs [95]. The observational literature that we assessed was mostly compiled to assess risk factors rather than interventions per se. Most of these risk or protective factors would not be possible to operationalize into intervention equivalents. However, when both observational and randomized evidence were available, in our overview, point estimates in different direction were quite common, 37.1% for observational studies versus RCTs and 30.8% for observational versus MR studies. Discrepancies beyond chance in the effect size estimates occurred in 43.5% for observational studies versus RCTs and 54.7% for observational studies versus MR studies.

Our study has several limitations. First, the seven standardized criteria were pre-specified based on what had been done previously in umbrella reviews. However, no consensus exists for a gold standard against which any criteria may be affirmed to truly quantify strength of the evidence and risk of bias [36] in observational studies of risk factors. Other efforts to-date have focused mostly on interventional evidence from RCTs where some observational evidence may be included (e.g., GRADE [96]) or specifically for interventional observational studies (e.g. ROBIS [97]).

Second, even though we included tens of thousands of observational studies, our assessment covers only specific fields for which umbrella reviews had been performed and these may not necessarily be fully generalizable to all observational epidemiology. Furthermore, only 16 out of 57 umbrella reviews also investigated meta-analyses of RCTs and MR studies in addition to meta-analyses of observational studies. Thus, we might not be capturing all meta-analyses of RCTs and all MR studies that reflect our included associations. MR studies are fairly recent and may even be more difficult to capture as they are often included in large genome-wide associations without being clearly identified. Moreover, both false positives and false-negative claims of causality may be made with MR studies, e.g., in the presence of weak genetic instruments.

Third, we used existing umbrella reviews which themselves focus on already existing meta-analyses. We did not appraise ourselves the quality of the included meta-analyses as this was already performed by the umbrella reviews' authors but flawed meta-analyses are not uncommon [98] and results should be taken with caution. The original studies may also be affected by selection bias, missing data, inadequate follow-up, and poor study

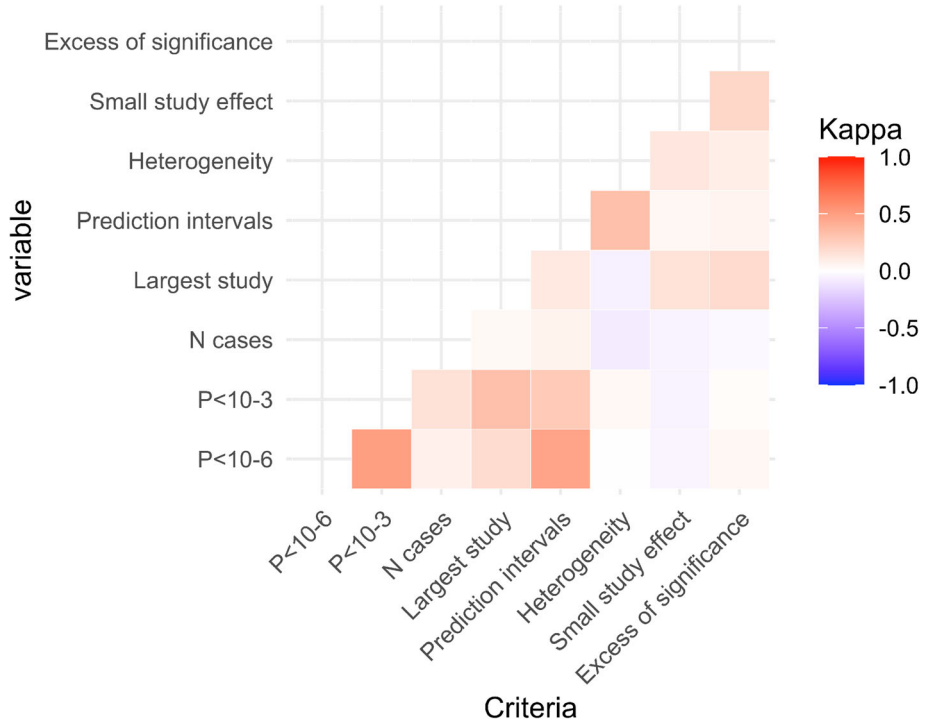

Fig. 3 Kappa heatmap for the seven criteria across all umbrella reviews. Only statistically significant associations (with $P < 0.05$ for the random effects summary effect) were included in the Cohen's kappa analysis. A $\kappa<0.6$ (lighter red) represents a weak, $0.6 \leq \kappa < 0.8$ a moderate (red), and $\kappa \geq 0.8$ (dark red) a strong strength of agreement. Conversely, a $\kappa>-0.6$ represents a weak (light blue), $-0.8<\kappa \leq -0.6$ a moderate (blue), and $\kappa \leq -0.8$ (dark blue) a strong disagreement. The kappa estimated within each umbrella reviews and combined using random effects meta-analyses are presented eFigure 5 and eFigure 6

conduct. For example, the serum uric acid [6] and statins [36] umbrella reviews also assessed the original studies in depth and found errors (e.g., incorrect data combining different level of exposure, use of duplicated data, and inclusion of different populations) that led to downgraded associations. Such errors require in-depth re-evaluation of the primary studies and their data. If anything, the proportion of associations with convincing or highly suggestive evidence might decrease even further, if one were to downgrade evidence because of the poor quality of meta-analyses and of primary studies. Finally, some of the included meta-analyses in umbrellas of different topics may have had some overlap, but we kept them so as to have each topic represented in its totality. We estimate that approximately 5% of the meta-analyses may be duplicates across two different topics,

Table 4 Changes in number of associations that are graded as having convincing evidence when one criterion is dropped or replaced by a more lenient version

Credibility assessment	N associations *(total=1457)*[a]	Proportion
Convincing	**99**	**6.8%**
Replace $P < 10^{-6}$ by <0.001	134	9.2%
Replace $P < 10^{-6}$ by < 0.05	142	9.7%
Without the minimum number of cases criterion	149	10.2%
Without the largest study at $p < 0.05$ criterion	103	7.1%
Without the 95% prediction interval criterion	106	7.3%
Without the heterogeneity I^2<50% criterion	159	10.9%
Without the small study effects criterion	122	8.4%
Without the excess significance criterion	111	7.6%

[a]These are the associations that are statistically significant ($P < 0.05$) and also have information on all criteria

but the exact number depends on how exactly duplication/overlap is defined. Regardless, the proportion is low to affect the results materially.

Conclusion

Allowing for these caveats, overall, our bird's eye view evaluation across 3744 meta-analyses of observational evidence on risk factors suggests that strong, large-scale, consistent, and uncontested observational evidence is probably very uncommon, even though statistically significant results are very common. It is also uncommon to find consistent corroborating evidence from RCTs or MR studies. Associations from meta-analyses of observational studies can offer interesting leads but require great caution, especially when high validity is required for decision-making.

Abbreviations

AMSTAR: A MeaSurement Tool to Assess systematic Reviews; HR: Hazard ratio; MR: Mendelian randomization; OR: Odds ratio; P: *P*-value; RCT: Randomized controlled trials; RR: Relative risks

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12916-021-02020-6.

Additional file 1. Appendix Method 1: Criteria for evaluating evidence strength.

Additional file 2: Figures 1 to 5. Forest plots of the proportions of associations for each level of evidence.

Additional file 3: Figures 6 to 7. Missing data.

Additional file 4: Figure 8 to 15. Forest plots of the proportions of associations fulfilling each criteria.

Additional file 5: Table 1. Kappa matrix across credibility criteria (below) and meta-analyzed across umbrella reviews (above).

Additional file 6: Figures 16 to 43. Forest plots of the meta-analyses of kappa estimated for each individual association.

Additional file 7: Table 2. Comparisons of observational studies and RCTs.

Additional file 8: Table 3. Comparisons of observational studies and MR studies.

Additional file 9: Figures 44 to 45. Forest plots of the ROR comparing meta-analyses of observational studies with meta-analyses of RCTs and with MR studies.

Acknowledgements

Not applicable

Authors' contributions

PJ and JPAI conceived the study. All authors (PJ, AA, IT, ET, KKT, EE, and JPAI) approved the protocol. PJ performed database searches and study selection. PJ and AA carried out the data extraction. IT, ET, and EE provided raw data. PJ carried out the analyses. PJ and JPAI drafted the manuscript. All authors (PJ, AA, IT, ET, KKT, EE and JPAI) critically revised the manuscript and approved the final version before submission. JPAI supervised the study.

Funding

METRICS is supported by a grant from the Laura and John Arnold Foundation. The work of JPAI is supported by an unrestricted gift from Sue and Bob O'Donnell. ET is supported by a CRUK Career Development Fellowship (C31250/A22804).

Availability of data and materials

The data used in this analysis are available to share.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declared that they have no competing interests.

Author details

[1]Meta-Research Innovation Center at Stanford (METRICS), Stanford, CA 94305, USA. [2]Department of Clinical Research, University Hospital Basel, University of Basel, CH-4056 Basel, Switzerland. [3]Department of Medicine, University of Toronto, 1 King's College Circle #3172, Toronto, ON M5S 1A8, Canada. [4]Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, University Campus, 45110 Ioannina, Greece. [5]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London W2 1PG, UK. [6]Centre for Global Health, The University of Edinburgh, Edinburgh EH8 9AG, UK. [7]Cancer Research UK Edinburgh Centre, Institute of Genetics and Cancer, Western General Hospital, The University of Edinburgh, Edinburgh EH4 2XU, UK. [8]Department of Epidemiology and Population Health, Stanford University School of Medicine, Stanford, CA 94305, USA. [9]Stanford Prevention Research Center, Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA. [10]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA 94305, USA. [11]Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, CA 94305, USA.

Received: 17 February 2021 Accepted: 28 May 2021
Published online: 06 July 2021

References

1. Ioannidis JP, Haidich AB, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. JAMA. 2001;286(7):821–30. https://doi.org/10.1001/jama.286.7.821.
2. Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Cochrane Database Syst Rev. 2014;4:MR000034.
3. Solmi M, Correll CU, Carvalho AF, Ioannidis JPA. The role of meta-analyses and umbrella reviews in assessing the harms of psychotropic medications: beyond qualitative synthesis. Epidemiol Psychiatr Sci. 2018;16:1–6.
4. Papatheodorou S. Umbrella reviews: what they are and why we need them. Eur J Epidemiol. 2019;34(6):543–6. https://doi.org/10.1007/s10654-019-00505-6.
5. Bellou V, Belbasis L, Tzoulaki I, Evangelou E, Ioannidis JPA. Environmental risk factors and Parkinson's disease: an umbrella review of meta-analyses. Parkinsonism Relat Disord. 2016;23:1–9. https://doi.org/10.1016/j.parkreldis.2015.12.008.
6. Li X, Meng X, Timofeeva M, Tzoulaki I, Tsilidis KK, Ioannidis JP, et al. Serum uric acid levels and multiple health outcomes: umbrella review of evidence from observational studies, randomised controlled trials, and Mendelian randomisation studies. BMJ. 2017;357:j2376.
7. Lawlor DA, Harbord RM, Sterne JAC, Timpson N, Davey SG. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Stat Med. 2008;27(8):1133–63. https://doi.org/10.1002/sim.3034.
8. Janiaud P, Agarwal A. Umbrella review of umbrella reviews. 2018 [cited 2021 May 26]; Available from: https://osf.io/g2hd7.
9. WHO | Risk factors [Internet]. WHO. [cited 2018 Jan 11]. Available from: http://www.who.int/topics/risk_factors/en/.
10. Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol. 2007;7(1):10. https://doi.org/10.1186/1471-2288-7-10.

11. Rücker G, Schwarzer G, Carpenter J, Olkin I. Why add anything to nothing? The arcsine difference as a measure of treatment effect in meta-analysis with zero cells. Stat Med. 2009;28(5):721–38. https://doi.org/10.1002/sim.3511.
12. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21(11):1539–58. https://doi.org/10.1002/sim.1186.
13. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37–46. https://doi.org/10.1177/001316446002000104.
14. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med. 2012; 22(3):276–82.
15. Baldwin JR, Reuben A, Newbury JB, Danese A. Agreement between prospective and retrospective measures of childhood maltreatment: a systematic review and meta-analysis. JAMA Psychiatry. 2019;76(6):584–93. https://doi.org/10.1001/jamapsychiatry.2019.0097.
16. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. J R Stat Soc Ser A Stat Soc. 2009;172(1):137–59. https://doi.org/10.1111/j.1467-985X.2008.00552.x.
17. Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. BMJ. 2011;342(feb10 2):d549. https://doi.org/10.1136/bmj.d549.
18. Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ. 2011;343(jul22 1):d4002. https://doi.org/10.1136/bmj.d4002.
19. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA. 1995;273(5):408–12. https://doi.org/10.1001/jama.1995.03520290060030.
20. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997;315(7109):629–34. https://doi.org/10.1136/bmj.315.7109.629.
21. Ioannidis JPA. Clarifications on the application and interpretation of the test for excess significance and its extensions. J Math Psychol. 2013;57(5):184–7. https://doi.org/10.1016/j.jmp.2013.03.002.
22. Ioannidis JPA. Excess significance bias in the literature on brain volume abnormalities. Arch Gen Psychiatry. 2011;68(8):773–80. https://doi.org/10.1001/archgenpsychiatry.2011.28.
23. Ioannidis JPA, Trikalinos TA. An exploratory test for an excess of significant findings. Clin Trials Lond Engl. 2007;4(3):245–53. https://doi.org/10.1177/1740774507079441.
24. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E-J, Berk R, et al. Redefine statistical significance. Nat Hum Behav. 2018;2(1):6–10. https://doi.org/10.1038/s41562-017-0189-z.
25. Hoaglin DC. Misunderstandings about Q and 'Cochran's Q test' in meta-analysis. Stat Med. 2016;35(4):485–95. https://doi.org/10.1002/sim.6632.
26. Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. Stat Med. 2000;19(22):3127–31. https://doi.org/10.1002/1097-0258(20001130)19:22<3127::AID-SIM784>3.0.CO;2-M.
27. Belbasis L, Köhler CA, Stefanis N, Stubbs B, van Os J, Vieta E, et al. Risk factors and peripheral biomarkers for schizophrenia spectrum disorders: an umbrella review of meta-analyses. Acta Psychiatr Scand. 2018;137(2):88–97. https://doi.org/10.1111/acps.12847.
28. Belbasis L, Bellou V, Evangelou E, Ioannidis JPA, Tzoulaki I. Environmental risk factors and multiple sclerosis: an umbrella review of systematic reviews and meta-analyses. Lancet Neurol. 2015;14(3):263–73. https://doi.org/10.1016/S1474-4422(14)70267-4.
29. Belbasis L, Bellou V, Evangelou E. Environmental risk factors and amyotrophic lateral sclerosis: an umbrella review and critical assessment of current evidence from systematic reviews and meta-analyses of observational studies. Neuroepidemiology. 2016;46(2):96–105. https://doi.org/10.1159/000443146.
30. Belbasis L, Savvidou MD, Kanu C, Evangelou E, Tzoulaki I. Birth weight in relation to health and disease in later life: an umbrella review of systematic reviews and meta-analyses. BMC Med. 2016;14(1):147. https://doi.org/10.1186/s12916-016-0692-5.
31. Belbasis L, Stefanaki I, Stratigos AJ, Evangelou E. Non-genetic risk factors for cutaneous melanoma and keratinocyte skin cancers: an umbrella review of meta-analyses. J Dermatol Sci. 2016;84(3):330–9. https://doi.org/10.1016/j.jdermsci.2016.09.003.
32. Bellou V, Belbasis L, Tzoulaki I, Middleton LT, Ioannidis JPA, Evangelou E. Systematic evaluation of the associations between environmental risk factors and dementia: an umbrella review of systematic reviews and meta-analyses. Alzheimers Dement J Alzheimers Assoc. 2017;13(4):406–18. https://doi.org/10.1016/j.jalz.2016.07.152.
33. Bellou V, Belbasis L, Tzoulaki I, Evangelou E. Risk factors for type 2 diabetes mellitus: an exposure-wide umbrella review of meta-analyses. PLoS One. 2018;13(3):e0194127. https://doi.org/10.1371/journal.pone.0194127.
34. Bortolato B, Köhler CA, Evangelou E, León-Caballero J, Solmi M, Stubbs B, et al. Systematic assessment of environmental risk factors for bipolar disorder: an umbrella review of systematic reviews and meta-analyses. Bipolar Disord. 2017;19(2):84–96. https://doi.org/10.1111/bdi.12490.
35. Giannakou K, Evangelou E, Papatheodorou SI. Genetic and non-genetic risk factors for pre-eclampsia: an umbrella review of systematic reviews and metaanalyses of observational studies. Ultrasound Obstet Gynecol Off J Int Soc Ultrasound Obstet Gynecol. 2018;51(6):720–30.
36. He Y, Li X, Gasevic D, Brunt E, McLachlan F, Millenson M, et al. Statins and multiple noncardiovascular outcomes: umbrella review of meta-analyses of observational studies and randomized controlled trials. Ann Intern Med. 2018;169(8):543–53. https://doi.org/10.7326/M18-0808.
37. Kalliala I, Markozannes G, Gunter MJ, Paraskevaidis E, Gabra H, Mitra A, et al. Obesity and gynaecological and obstetric conditions: umbrella review of the literature. BMJ. 2017;359:j4511.
38. Köhler CA, Evangelou E, Stubbs B, Solmi M, Veronese N, Belbasis L, et al. Mapping risk factors for depression across the lifespan: an umbrella review of evidence from meta-analyses and Mendelian randomization studies. J Psychiatr Res. 2018;103:189–207. https://doi.org/10.1016/j.jpsychires.2018.05.020.
39. Kyrgiou M, Kalliala I, Markozannes G, Gunter MJ, Paraskevaidis E, Gabra H, et al. Adiposity and cancer at major anatomical sites: umbrella review of the literature. BMJ. 2017;356:j477.
40. Machado MO, Veronese N, Sanches M, Stubbs B, Koyanagi A, Thompson T, et al. The association of depression and all-cause and cause-specific mortality: an umbrella review of systematic reviews and meta-analyses. BMC Med. 2018;16(1):112. https://doi.org/10.1186/s12916-018-1101-z.
41. Markozannes G, Tzoulaki I, Karli D, Evangelou E, Ntzani E, Gunter MJ, et al. Diet, body size, physical activity and risk of prostate cancer: an umbrella review of the evidence. Eur J Cancer Oxf Engl 1990. 2016;69:61–9.
42. Radua J, Ramella-Cravaro V, Ioannidis JPA, Reichenberg A, Phiphopthatsanee N, Amir T, et al. What causes psychosis? An umbrella review of risk and protective factors. World Psychiatry Off J World Psychiatr Assoc WPA. 2018; 17(1):49–66.
43. de Rezende LFM, de Sá TH, Markozannes G, Rey-López JP, Lee I-M, Tsilidis KK, et al. Physical activity and cancer: an umbrella review of the literature including 22 major anatomical sites and 770 000 cancer cases. Br J Sports Med. 2018;52(13):826–33.
44. Solmi M, Köhler CA, Stubbs B, Koyanagi A, Bortolato B, Monaco F, et al. Environmental risk factors and nonpharmacological and nonsurgical interventions for obesity: an umbrella review of meta-analyses of cohort studies and randomized controlled trials. Eur J Clin Invest. 2018;20:e12982.
45. Theodoratou E, Tzoulaki I, Zgaga L, Ioannidis JPA. Vitamin D and multiple health outcomes: umbrella review of systematic reviews and meta-analyses of observational studies and randomised trials. BMJ. 2014;348(apr01 2): g2035. https://doi.org/10.1136/bmj.g2035.
46. Tsilidis KK, Kasimis JC, Lopez DS, Ntzani EE, Ioannidis JPA. Type 2 diabetes and cancer: umbrella review of meta-analyses of observational studies. BMJ. 2015;350(jan02 1):g7607. https://doi.org/10.1136/bmj.g7607.
47. Veronese N, Demurtas J, Celotto S, Caruso MG, Maggi S, Bolzetta F, et al. Is chocolate consumption associated with health outcomes? An umbrella review of systematic reviews and meta-analyses. Clin Nutr. 2019;38(3):1101–8.
48. Veronese N, Solmi M, Caruso MG, Giannelli G, Osella AR, Evangelou E, et al. Dietary fiber and health outcomes: an umbrella review of systematic reviews and meta-analyses. Am J Clin Nutr. 2018;107(3):436–44. https://doi.org/10.1093/ajcn/nqx082.
49. Zhou C, Zhong X, Song Y, Shi J, Wu Z, Guo Z, et al. Prognostic biomarkers for gastric cancer: an umbrella review of the evidence. Front Oncol. 2019;9: 1321. https://doi.org/10.3389/fonc.2019.01321.
50. Zhao L-G, Li Z-Y, Feng G-S, Ji X-W, Tan Y-T, Li H-L, et al. Coffee drinking and cancer risk: an umbrella review of meta-analyses of observational studies. BMC Cancer. 2020;20(1):101. https://doi.org/10.1186/s12885-020-6561-9.
51. Yu H, Zhong X, Gao P, Shi J, Wu Z, Guo Z, et al. The potential effect of metformin on cancer: an umbrella review. Front Endocrinol. 2019;10:617. https://doi.org/10.3389/fendo.2019.00617.

52. Veronese N, Demurtas J, Pesolillo G, Celotto S, Barnini T, Calusi G, et al. Magnesium and health outcomes: an umbrella review of systematic reviews and meta-analyses of observational and intervention studies. Eur J Nutr. 2020;59(1):263–72. https://doi.org/10.1007/s00394-019-01905-w.
53. Smith L, Luchini C, Demurtas J, Soysal P, Stubbs B, Hamer M, et al. Telomere length and health outcomes: an umbrella review of systematic reviews and meta-analyses of observational studies. Ageing Res Rev. 2019;51:1–10. https://doi.org/10.1016/j.arr.2019.02.003.
54. Raglan O, Kalliala I, Markozannes G, Cividini S, Gunter MJ, Nautiyal J, et al. Risk factors for endometrial cancer: an umbrella review of the literature. Int J Cancer. 2019;145(7):1719–30.
55. Papola D, Ostuzzi G, Gastaldon C, Morgano GP, Dragioti E, Carvalho AF, et al. Antipsychotic use and risk of life-threatening medical events: umbrella review of observational studies. Acta Psychiatr Scand. 2019;140(3):227–43. https://doi.org/10.1111/acps.13066.
56. Kim JY, Steingroever J, Lee KH, Oh J, Choi MJ, Lee J, et al. Clinical interventions and all-cause mortality of patients with chronic kidney disease: an umbrella systematic review of meta-analyses. J Clin Med. 2020; 9(2):394.
57. Kim JY, Son MJ, Son CY, Radua J, Eisenhut M, Gressier F, et al. Environmental risk factors and biomarkers for autism spectrum disorder: an umbrella review of the evidence. Lancet Psychiatry. 2019;6(7):590–600. https://doi.org/10.1016/S2215-0366(19)30181-6.
58. Grabovac I, Veronese N, Stefanac S, Haider S, Jackson SE, Koyanagi A, et al. Human immunodeficiency virus infection and diverse physical health outcomes: an umbrella review of meta-analyses of observational studies. Clin Infect Dis Off Publ Infect Dis Soc Am. 2020;70(9):1809–15. https://doi.org/10.1093/cid/ciz539.
59. Giannakou K, Evangelou E, Yiallouros P, Christophi CA, Middleton N, Papatheodorou E, et al. Risk factors for gestational diabetes: an umbrella review of meta-analyses of observational studies. PLoS One. 2019;14(4): e0215372. https://doi.org/10.1371/journal.pone.0215372.
60. Fullana MA, Tortella-Feliu M, de la Cruz LF, Chamorro J, Pérez-Vigil A, Ioannidis JPA, et al. Risk and protective factors for anxiety and obsessive-compulsive disorders: an umbrella review of systematic reviews and meta-analyses. Psychol Med. 2020;50(8):1300–15. https://doi.org/10.1017/S0033291719001247.
61. Dragioti E, Solmi M, Favaro A, Fusar-Poli P, Dazzan P, Thompson T, et al. Association of antidepressant use with adverse health outcomes: a systematic umbrella review. JAMA Psychiatry. 2019;76(12):1241–55.
62. Bellou V, Belbasis L, Konstantinidis AK, Evangelou E. Elucidating the risk factors for chronic obstructive pulmonary disease: an umbrella review of meta-analyses. Int J Tuberc Lung Dis Off J Int Union Tuberc Lung Dis. 2019; 23(1):58–66.
63. Belbasis L, Dosis V, Evangelou E. Elucidating the environmental risk factors for rheumatic diseases: an umbrella review of meta-analyses. Int J Rheum Dis. 2018;21(8):1514–24. https://doi.org/10.1111/1756-185X.13356.
64. Kim JH, Kim JY, Lee J, Jeong GH, Lee E, Lee S, et al. Environmental risk factors, protective factors, and peripheral biomarkers for ADHD: an umbrella review. Lancet Psychiatry. 2020;7(11):955–70. https://doi.org/10.1016/S2215-0366(20)30312-6.
65. Solmi M, Dragioti E, Arango C, Radua J, Ostinelli E, Kilic O, et al. Risk and protective factors for mental disorders with onset in childhood/adolescence: an umbrella review of published meta-analyses of observational longitudinal studies. Neurosci Biobehav Rev. 2021;120:565–73.
66. Valenzuela PL, Santos-Lozano A, Morales JS, López-Ortiz S, Pinto-Fraga J, Castillo-García A, et al. Physical activity, sports and risk of atrial fibrillation: umbrella review of meta-analyses. Eur J Prev Cardiol. 2020;16: 2047487320923183.
67. Belbasis L, Mavrogiannis MC, Emfietzoglou M, Evangelou E. Environmental factors, serum biomarkers and risk of atrial fibrillation: an exposure-wide umbrella review of meta-analyses. Eur J Epidemiol. 2020;35(3):223–39. https://doi.org/10.1007/s10654-020-00618-3.
68. Solmi M, Veronese N, Galvano D, Favaro A, Ostinelli EG, Noventa V, et al. Factors associated with loneliness: an umbrella review of observational studies. J Affect Disord. 2020;271:131–8. https://doi.org/10.1016/j.jad.2020.03.075.
69. Veronese N, Demurtas J, Thompson T, Solmi M, Pesolillo G, Celotto S, et al. Effect of low-dose aspirin on health outcomes: an umbrella review of systematic reviews and meta-analyses. Br J Clin Pharmacol. 2020;86(8):1465–75. https://doi.org/10.1111/bcp.14310.
70. Song Y, Zhong X, Gao P, Zhou C, Shi J, Wu Z, et al. Aspirin and its potential preventive role in cancer: an umbrella review. Front Endocrinol. 2020;11:3. https://doi.org/10.3389/fendo.2020.00003.
71. Zhang X, Gill D, He Y, Yang T, Li X, Monori G, et al. Non-genetic biomarkers and colorectal cancer risk: umbrella review and evidence triangulation. Cancer Med. 2020;9(13):4823–35. https://doi.org/10.1002/cam4.3051.
72. Xu W, He Y, Wang Y, Li X, Young J, Ioannidis JPA, et al. Risk factors and risk prediction models for colorectal cancer metastasis and recurrence: an umbrella review of systematic reviews and meta-analyses of observational studies. BMC Med. 2020;18(1):172. https://doi.org/10.1186/s12916-020-01618-6.
73. Markozannes G, Koutsioumpa C, Cividini S, Monori G, Tsilidis KK, Kretsavos N, et al. Global assessment of C-reactive protein and health-related outcomes: an umbrella review of evidence from observational studies and Mendelian randomization studies. Eur J Epidemiol. 2021;36(1):11–36.
74. Biffi A, Cantarutti A, Rea F, Locatelli A, Zanini R, Corrao G. Use of antidepressants during pregnancy and neonatal outcomes: an umbrella review of meta-analyses of observational studies. J Psychiatr Res. 2020;124: 99–108. https://doi.org/10.1016/j.jpsychires.2020.02.023.
75. Solmi M, Radua J, Stubbs B, Ricca V, Moretti D, Busatta D, et al. Risk factors for eating disorders: an umbrella review of published meta-analyses. Braz J Psychiatry. 2021;43(3):314–23.
76. Lee KH, Seong HJ, Kim G, Jeong GH, Kim JY, Park H, et al. Consumption of fish and ω-3 fatty acids and cancer risk: an umbrella review of meta-analyses of observational studies. Adv Nutr Bethesda Md. 2020;11(5):1134–49. https://doi.org/10.1093/advances/nmaa055.
77. Demurtas J, Celotto S, Beaudart C, Sanchez-Rodriguez D, Balci C, Soysal P, et al. The efficacy and safety of influenza vaccination in older people: an umbrella review of evidence from meta-analyses of both observational and randomized controlled studies. Ageing Res Rev. 2020;62:101118. https://doi.org/10.1016/j.arr.2020.101118.
78. Soysal P, Hurst C, Demurtas J, Firth J, Howden R, Yang L, et al. Handgrip strength and health outcomes: Umbrella review of systematic reviews with metaanalyses of observational studies. J Sport Health Sci. 2021;10(3):290–5.
79. Carvalho AF, Solmi M, Sanches M, Machado MO, Stubbs B, Ajnakina O, et al. Evidence-based umbrella review of 162 peripheral biomarkers for major mental disorders. Transl Psychiatry. 2020;10(1):152. https://doi.org/10.1038/s41398-020-0835-5.
80. Wang Y, Zhong X, Zhou L, Lu J, Jiang B, Liu C, et al. Prognostic biomarkers for pancreatic ductal adenocarcinoma: an umbrella review. Front Oncol. 2020;10:1466. https://doi.org/10.3389/fonc.2020.01466.
81. Kim TL, Jeong GH, Yang JW, Lee KH, Kronbichler A, van der Vliet HJ, et al. Tea consumption and risk of cancer: an umbrella review and meta-analysis of observational studies. Adv Nutr Bethesda Md. 2020;11(6):1437–52. https://doi.org/10.1093/advances/nmaa077.
82. Janiaud P, Agarwal A. Umbrella review of umbrella reviews. 2018 [cited 2021 May 26]; Available from: https://osf.io/xj5cf/
83. Ioannidis JPA, Trikalinos TA. The appropriateness of asymmetry tests for publication bias in meta-analyses: a large survey. CMAJ. 2007;176(8):1091–6. https://doi.org/10.1503/cmaj.060410.
84. Cornell JE, Mulrow CD, Localio R, Stack CB, Meibohm AR, Guallar E, et al. Random-effects meta-analysis of inconsistent effects: a time for change. Ann Intern Med. 2014;160(4):267–70. https://doi.org/10.7326/M13-2886.
85. Terrin N, Schmid CH, Lau J, Olkin I. Adjusting for publication bias in the presence of heterogeneity. Stat Med. 2003;22(13):2113–26. https://doi.org/10.1002/sim.1461.
86. Patel CJ, Ji J, Sundquist J, Ioannidis JPA, Sundquist K. Systematic assessment of pharmaceutical prescriptions in association with cancer risk: a method to conduct a population-wide medication-wide longitudinal study. Sci Rep. 2016;6(1):31308. https://doi.org/10.1038/srep31308.
87. Schuemie MJ, Ryan PB, Hripcsak G, Madigan D, Suchard MA. Improving reproducibility by using high-throughput observational studies with empirical calibration. Philos Transact A Math Phys Eng Sci. 2018;13: 376(2128).
88. Schuemie MJ, Cepeda MS, Suchard MA, Yang J, Tian Y, Schuler A, et al. How confident are we about observational findings in healthcare: a benchmark study. Harv Data Sci Rev. 2020; [cited 2021 Feb 2]; Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7755157/.
89. Djulbegovic B, Glasziou P, Klocksieben FA, Reljic T, VanDenBergh M, Mhaskar R, et al. Larger effect sizes in nonrandomized studies are associated with higher rates of EMA licensing approval. J Clin Epidemiol. 2018;98:24–32. https://doi.org/10.1016/j.jclinepi.2018.01.011.

90. Razavi M, Glasziou P, Klocksieben FA, Ioannidis JPA, Chalmers I, Djulbegovic B. US Food and Drug Administration approvals of drugs and devices based on nonrandomized clinical trials: a systematic review and meta-analysis. JAMA Netw Open. 2019;2(9):e1911111. https://doi.org/10.1001/jamanetworkopen.2019.11111.
91. Sacks H, Chalmers TC, Smith H. Randomized versus historical controls for clinical trials. Am J Med. 1982;72(2):233–40. https://doi.org/10.1016/0002-9343(82)90815-4.
92. Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy. I: Medical. Stat Med. 1989;8(4):441–54. https://doi.org/10.1002/sim.4780080408.
93. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. N Engl J Med. 2000;342(25):1878–86. https://doi.org/10.1056/NEJM200006223422506.
94. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med. 2000;342(25):1887–92. https://doi.org/10.1056/NEJM200006223422507.
95. Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JPA. Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials: meta-epidemiological survey. BMJ. 2016;352:i493.
96. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011;64(4):383–94. https://doi.org/10.1016/j.jclinepi.2010.04.026.
97. Whiting P, Savović J, Higgins JPT, Caldwell DM, Reeves BC, Shea B, et al. ROBIS: a new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol. 2016;69:225–34. https://doi.org/10.1016/j.jclinepi.2015.06.005.
98. Ioannidis JPA. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. Milbank Q. 2016;94(3):485–514. https://doi.org/10.1111/1468-0009.12210.

Publisher's Note