nature mental health

Supplementary information

https://doi.org/10.1038/s44220-024-00328-9

A meta-analysis of loneliness and risk of dementia using longitudinal data from >600,000 individuals

In the format provided by the authors and unedited

Supplementary Information

A Meta-analysis of Loneliness and Risk of Dementia using Longitudinal Data from >600,000 Individuals

Table of Contents:

List of Abbreviations

AD = Alzheimer's disease
BMI = body mass index
CES-D = Center for Epidemiological Studies-Depression
CIND = cognitive impairment, no dementia
CI = non-specific cognitive impairment
IQCODE = Informant Questionnaire on Cognitive Decline in the Elderly
ISCED-97 = International Standard Classification of Education, 1997
MCI = mild cognitive impairment
MMSE = Mini-Mental State Examination
SF-36 = 36-Item Short Form Health Survey
TICSm = modified Telephone Interview for Cognitive Status
UCLA = University of California, Los Angeles
VaD = vascular dementia

Data Extraction File and R scripts for the meta-analysis are accessible at Open Science Framework: https://osf.io/tfphs/

Luchetti, M. *et al.* Loneliness and Risk for Dementia. (2024). doi:10.17605/OSF.IO/TFPHS

Table S1: Preregistration deviations

Deviations					
#	Details		Original Wording	Deviation Description	Reader Impact
1	Type	Variables	**Main Outcome(s)** "Studies that evaluate risk for incident (all cause) dementia or incident Alzheimer's disease. Incidence of dementia may be detected through clinical diagnosis, neuropathological evaluation, or a cut-off for cognitive performance."	There was an oversight in the operationalization and description of the main outcome. We conducted separate meta-analyses for all-cause and cause-specific dementias. We also did not specify other possible sources of information, such as doctor diagnosis, reported by study participants and/or knowledgeable proxies.	This deviation has a relatively small impact on readers' interpretations of the findings. It has been implemented to distinguish loneliness-associated risk for cause-specific dementia, when this information was available. The results were not already known, so the risk of bias is low.
	Reason	Other (Please Explain)			
	Timing	During data collection			
2	Type	Analysis	**Intervention(s), exposure(s)** "We will consider studies that assess both; i.e., loneliness in general, and its social and emotional dimensions." **Analysis of subgroups or subsets** "… we will group measures based on the most common scale assessed to use loneliness (for example, versions of the UCLA scale vs. other scales), number of items (1-item vs. multi-item scale), type of loneliness measured (emotional, social or unidimensional construct), and formulation of items (direct or indirect questions to assess loneliness)."	It was not possible to evaluate specific types of loneliness because the selected studies (except Shibata *et al*. 2021) did not distinguish between social and emotional loneliness. Most studies used a direct, single item to assess loneliness, often categorized as presence/absence of loneliness. We therefore explored whether loneliness entered as a categorical/binary predictor versus as a continuous/ordinal scale moderated the association with dementia/cognitive impairment. We did not report meta-regressions for number of items and formulation of the items as these moderators overlapped with the type of loneliness scale used and data source.	This deviation is due to a lack of data available in the published literature. It points to the necessity for additional work to address social and emotional components of loneliness.
	Reason	Plan not possible			
	Timing	After data access			
3	Type	Analysis	**Analysis of subgroups or subsets** "We will test whether results vary by the country/nationality of the samples (North America, Central-South America, Europe, Asia)." "We will test study characteristics and measures as moderators: quality score, data source (new data vs. published study), age of the sample, proportion of females, proportion of minoritized individuals"	Studies from North and Central America were grouped together as we did not identify any study from South America. Most studies did not include information on race or ethnicity, and therefore we were not able to assess the proportion of individuals from race/ethnic minorities as a moderator.	This deviation is due to a lack of data available in the published literature.
	Reason	Plan not possible			
	Timing	After data access			

4	Type	Analysis	**Analysis of subgroups or subsets** "We will test whether results differed between studies that used a clinical diagnosis to assess dementia/cognitive impairment compared to studies that relied on cognitive performance"	During the review process, we extracted additional information from the selected articles to better specify the type and source of information used to classify dementia/cognitive impairment across studies. The meta-regression compared studies with a diagnosis based on clinical symptoms, neuropsychological evaluation, medical and/or death records with studies that exclusively used validated performance-based cut-offs and self or proxy reports of a doctor diagnosis of dementia.	This deviation has little impact on the results. The reader is provided with additional information on the type of cognitive assessment and dementia diagnosis used across studies.
	Reason	Peer review			
	Timing	After results known			

Unregistered Steps

#	Details		Original Wording	Unregistered Step Description	Reader Impact
1	Type	Analysis	"We will test whether results vary by dementia types (Alzheimer's disease vs. all-cause dementia) and cognitive impairment type (CIND vs. MCI)"	There was an oversight in the operationalization and description of the main outcome. We conducted separate analyses for all-cause and cause-specific dementias. Most studies also assessed CIND/CI. Thus, we ran the meta-analysis in a subset of studies that assessed CI, but did not ascertain absence of dementia, or did not distinguish between mild and severe cognitive impairment.	This unregistered step has little impact on the current results and paper. The analysis helps the reader by providing additional information on the influence of the type of cognitive assessment when examining the association between loneliness and CIND/CI.
	Timing	After data access			
2	Type	Analysis	The target population/participants included "middle-aged and older adults" without specifying an age range.	Upon the request of a reviewer, we performed a supplementary analysis within each individual cohort study restricting the analytical sample to adults aged 50 and older. We also re-ran the meta-analysis including samples aged 50 and older.	This unregistered step has little impact on the current results and paper. To help the reader evaluate the effect of age on the association, we report results from samples that included relatively younger participants, as well as results from samples restricted to adults aged 50 or older.
	Timing	After results known			
3	Type	Analysis	"Analysis will account for age, sex/gender, education, and race/ethnicity; where possible, follow-up analyses will include depression and/or measures of social isolation."	Upon the request of a reviewer, we included additional covariates in the sensitivity analysis. Within each individual cohort study, we controlled for measures of depression and isolation, as well as	This unregistered step has little impact on the current results and paper. To help the reader evaluate the strength of the association between loneliness and
	Timing	After results known			

			"To reduce variability across studies, we will extract estimates of statistical models that account for age, sex/gender, education, and race/ethnicity; results of models with additional covariates, particularly models including depressive symptoms and/or measures of social isolation, will be extracted for use in sensitivity analysis."	established modifiable clinical factors for dementia (i.e., diabetes, hypertension, and obesity). We also re-ran the meta-analysis including results from models with the additional covariates.	dementia/cognitive impairment after the inclusion of depression, isolation, and modifiable clinical factors, we provide the results of the additional analysis performed with each cohort.

Note: We followed recommendations from Willroth and Atherton (2024)[1].

Note S1: Description of the individual cohort studies

A general description of the selected cohorts is provided below. Supplementary Tables S2-S4 describe measures and coding schemes for loneliness, cognitive status, and control variables across studies. The same analyses were conducted across all samples with harmonized variables to facilitate comparisons and meta-analysis. For the classification of cognitive status, we used the definition validated in each individual sample. In each study, we identified dementia and/or CIND/CI.

The Health and Retirement Study (HRS)

The HRS is a longitudinal study of Americans aged 50 years and older and their spouses, regardless of age. Sampling for HRS was based on a multi-stage area probability design that included geographical stratification and clustering. HRS also oversampled for Black and Hispanic households to ensure representation of minorities. The 3-item UCLA-R scale[2] of loneliness was introduced as part of the Left-Behind Questionnaire and first administered to a random half of the sample in 2006, and the other half first completed the measure in 2008; these two sub-samples were combined to form our baseline assessment 2006/08. Cognitive status was assessed with the modified Telephone Interview for Cognitive Status (TICSm)[3] every two years from 2006 to 2022 (max. 9 possible assessments). TICSm scores ≤6 (out of 27) were indicative of dementia; scores between 7 and 11 were indicative of CIND[3]. Our analysis included data from 12,217 participants without dementia at baseline (n = 10,079 were included in the analysis for CIND; we excluded those participants with CIND at the baseline assessment and those who developed dementia before CIND); participants had to have baseline information on age (≥ 40 years), sex, education (in years), race (white vs. black/other minorities) and ethnicity (Hispanic vs. non-Hispanic origins), and two or more cognitive assessments over the study period. The follow-up ranged from 2 to 16 years.

HRS data are publicly available and accessible after registration at:
https://hrs.isr.umich.edu/data-products

Sonnega A, Faul J, Ofstedal MBeth, Langa KM, Phillips JWR, Weir DR. Cohort Profile: The Health and Retirement Study (HRS). *Int J Epidemiol* **43**, 576-585 (2014).

The English Longitudinal Study of Ageing (ELSA)

The ELSA is a longitudinal study of aging of individuals over the age of 50 in England and their spouses, regardless of age. ELSA was developed to be a sister study to HRS. The initial sample was recruited in 2002/03 from consenting households that had participated in the Health Survey for England (HSE) in 1998, 1999, and 2001. The 3-item UCLA scale was first administered at the 2004-05 (wave 2) assessment, which consisted of our baseline. Prior work[4,5] has used a classification system for dementia cases based on a combination of a self-report doctor diagnosis of Alzheimer's disease or dementia and/or a score ≥ 3.5 on a short-form of the Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE)[6], as reported by a knowledgeable proxy. Dementia status was assessed every two years from 2004 to 2023 (max. 9 possible assessments). ELSA also administered the TICSm from the 2014 assessment on. Using the last four waves of assessment (up to 2021-23), we assessed the risk of developing CIND, excluding ELSA participants who reported a diagnosis of dementia or an IQCODE above the 3.5 cut-off at the baseline assessment. Our analysis included 6,771 participants without dementia at baseline (n = 6,349 were included in the analysis for CIND); participants had to have baseline information on age (≥ 40 years), sex, education (7-point ordinal scale), and race (white vs. non-white), and two or more cognitive assessments over the study period. The follow-up ranged from 1 to 19 years.

ELSA data are publicly available and accessible after registration at: https://beta.ukdataservice.ac.uk/datacatalogue/series/series?id=200011

Banks, J. *et al.* English Longitudinal Study of Ageing: Waves 0-10, 1998-2023. Data collection. UK Data Service. SN: 5050. (2024). http://doi.org/10.5255/UKDA-SN-5050-27

The Survey of Health, Ageing and Retirement in Europe (SHARE)

The SHARE is a cross-national study of health and aging in 28 European countries and Israel (29 countries total). SHARE is a sister study to HRS and ELSA. The target population is adults aged 50, who live in and speak the language of their respective SHARE country. As with HRS and ELSA, current partners living in the household were also enrolled regardless of age. SHARE included 12 countries at the first wave in 2004-06 (Austria, Belgium, Denmark, France, Germany, Greece, Israel, Italy, the Netherlands, Spain, Sweden, and Switzerland). In this first wave, loneliness was measured as part of the drop-off questionnaire; participants were asked "How often do you feel lonely?" from *almost all the time* to *almost none of the time* (4-point scale). From the 2011-12 assessment (wave 4), the 3-item of the UCLA scale was included in the questionnaire. At this latter assessment, the questionnaire was administered across 15 countries (Austria, Belgium, Czech Republic, Denmark, France, Germany, Hungary, Italy, the Netherlands, Poland, Portugal, Slovenia, Spain, Sweden, and Switzerland). At each wave, up to the 2021-22 assessment (wave 9), participants completed a memory task (immediate and delayed recall of 10 words) and a verbal fluency (category) task. Following previous publications[7,8], we identified participants with CI based on cognitive performance: Individuals with a memory and verbal fluency score of 1.5 SD equal or below the age-graded mean of the sample on both tasks were classified as cognitively impaired. Note that at the 2017-18 assessment (wave 7), only part of the sample had cognitive data, due to the SHARE method design at that particular wave. When estimating the risk of CI, we conducted the analyses in two separate samples, Sample 1 (n = 15,028), which included participants with loneliness assessed in 2004, and Sample 2 (n = 27,673), which included participants who completed the 3-item UCLA in 2011. For the analysis with Sample 2, we excluded participants who were part of Sample 1. That is, the two samples did not overlap. Starting from 2011 (wave 4), SHARE also recorded doctor diagnosis of Alzheimer's disease, dementia, or senility, reported directly by the participant and/or their proxy. Following prior publications[9–11], we used this information to estimate risk of dementia across waves 4 to 9 (n = 35,531). The Dementia Sample in part overlapped with Sample 1 and 2. For each sample, participants had to have baseline information on age (≥ 40 years), sex, and education (7-point ordinal scale), and two or more cognitive assessments (or reported diagnosis) over the study period. The follow-up ranged from 1 to 18 years for Sample 1 and from 1 to 11 for Sample 2 and the Dementia Sample.

SHARE data are publicly available and accessible after registration at: https://share-eric.eu/data/data-access

Börsch-Supan, A. *et al.* Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). *Int J Epidemiol* **42**, 992–1001 (2013).

SHARE-ERIC. Survey of Health, Ageing and Retirement in Europe (SHARE). Release version: 9.0.0. Datasets. Wave 1 http://dx.doi.org/10.6103/SHARE.w1.900; Wave 2 http://dx.doi.org/10.6103/SHARE.w2.900; Wave 4 http://dx.doi.org/10.6103/SHARE.w4.900; Wave 5, http://dx.doi.org/10.6103/SHARE.w5.900; Wave 6, http://dx.doi.org/10.6103/SHARE.w6.900; Wave 7 http://dx.doi.org/10.6103/SHARE.w7.900; Wave 8,

http://dx.doi.org/10.6103/SHARE.w8.900; Wave 9, http://dx.doi.org/10.6103/SHARE.w9.900 (2024).

The Irish Longitudinal Study on Ageing (TILDA)

The TILDA is a longitudinal study of adults aged 50 and older, and their spouses, regardless of age. The initial sample was recruited between 2009 and 2011 using a multistage probability sampling of all households in Ireland. Follow-up assessments occurred in 2012-13, 2014-15, 2016, and 2018. The 3-item of the UCLA scale was administered at the 2009-11 baseline assessment. Cognitive status was assessed at each wave with the Mini-Mental State Examination (MMSE)[12]; scores between 24 and 30 were classified as normal cognition, scores between 18 and 23 were classified as CIND, and scores < 18 were classified as dementia[13,14]. Our analysis included 4,833 participants without CIND at baseline; participants had to have baseline information on age (range 50-80+), sex, and education (7-point ordinal scale), and two or more cognitive assessments over the study period. The follow-up ranged from 2 to 8 years. Please note that TILDA does not release interview dates for participants; we thus used the year of assessment to calculate follow-up length, which is the same approach taken by Gateway of Global Aging in their harmonized files. Further, analysis with dementia as an outcome (MMSE < 18) was not performed; there were only 0.4% (n = 19) incident cases of dementia over the follow-up.

TILDA anonymized datasets were accessed at the Inter-university Consortium for Political and Social Research: https://www.icpsr.umich.edu/web/ICPSR/series/726

Kenny, R. A. The Irish Longitudinal Study on Ageing (TILDA), Wave 1, 2009-2011 (ICPSR 34315), Wave 2, 2012-2013 (ICPSR 37105), Wave 3, 2014-2015 (ICPSR 37106), Wave 4, 2016 (ICPSR 38670), Wave 5, 2018 (ICPSR 38674). Inter-university Consortium for Political and Social Research [distributor], 2023-03-07.

The Mexican Health and Aging Study (MHAS)

The MHAS is a longitudinal study of adults aged 50 and older, and their spouse, regardless of age, living in private households in Mexico. The initial sample was recruited in 2001 from households that had participated in the 2000 National Employment Survey. Follow-up assessments occurred in 2003, 2012, 2015, 2018 and 2021. Loneliness was measured at the 2001 baseline with a single item (i.e., “In the past week… You felt lonely”, response: yes/no), which was part of a 9-item version of the CES-D, validated for the Mexican population[15,16]. At each wave, cognitive status and functionality in activities of daily living were assessed both using direct interviews and proxy interviews. Based on this information, the MHAS convened a group of experts, who followed and adapted the criteria for preclinical and clinical phases of all-cause dementia recommended by the National Institute of Aging and the Alzheimer’s Association, to classify participants into the three following categories[17]. (1) Dementia: participants with impairment in two or more cognitive domains and one or more functional limitations, or having a proxy IQCODE score ≥ 3.4. (2) CIND: participants with impairment in two or more cognitive domains and no functional limitations. (3) Normal cognition: participants with normal cognitive function and no functional limitations or impairment in instrumental activity of daily living due to physical limitations, and proxy respondents with an IQCODE score < 3.4. This classification is available for participants in 2001, 2003, 2012, and 2015 waves, and has been used in previous studies with MHAS[18]. Our analysis included 11,672 participants without dementia at baseline (n = 7,349 were included in the analysis for CIND); participants had to have baseline information on age (≥ 40 years), sex, and education (in years), and two or more cognitive assessments over the study period. The follow-up ranged from 2 to 15 years.

MHAS data are publicly available and accessible after registration at:
https://www.mhasweb.org/DataProducts/Home.aspx

Wong, R., Michaels-Obregon, A., & Palloni, A. Cohort profile: The Mexican Health and Aging Study (MHAS). *Int J Epidemiol* **46**, e2 (2017).

The Korean Longitudinal Study of Aging (KLoSA)

The KLoSA is a longitudinal study of adults aged 45 years and older, living in private households in Korea. The initial sample was recruited in 2006 using a multistage, stratified probability sampling of households in all geographic areas except for Jeju Island. Spouses living in the household were also enrolled if meeting age inclusion criteria. Loneliness was measured at the 2006 baseline with a single item (i.e., "In the past week… You felt lonely", from *rarely or none of the time* to *all the time* (4-point scale), which was part of a 10-item version of the CES-D[19]. Cognitive status was assessed at each wave with the MMSE; scores between 24 and 30 were classified as normal cognition, scores between 18 and 23 were classified as CIND, and scores < 18 were classified as dementia[14]. Our analysis included 7,991 participants without dementia at baseline (n = 6,266 were included in the analysis for CIND); participants had to have baseline information on age, sex, and education (4-point ordinal scale), and two or more cognitive assessments over the study period. The follow-up ranged from 2 to 14 years.

KLoSA data are publicly available and accessible after registration at:
https://survey.keis.or.kr/eng/myinfo/login.jsp

Shin, C. Korean Longitudinal Study of Aging. in *Encyclopedia of Gerontology and Population Aging* (D. Gu & M. E. Dupre Eds.). (Springer, 2019).

The China Health and Retirement Longitudinal Study (CHARLS)

The CHARLS is a nationally longitudinal study of adults aged 45 years and older, and their spouse, regardless of age, living in private households in China. The sample also has urban/rural representation. The initial sample was recruited in 2011/12 using a stratified, multistage probability sampling of households in all county-level units except for Tibet. Follow-up assessments occurred in 2013, 2015, 2018 and 2020. Loneliness was measured at the 2011/12 baseline with a single item (i.e., "In the past week… You felt lonely", from *rarely or none of the time* to *most or all the time* (4-point scale), which was part of a 10-item version of the CES-D[20]. Following prior publications[21], a composite cognitive score was derived at each wave based on performance on a memory recall, serial subtraction, time orientation and a figure drawing task; participants with a composite score equal or below 1.5 SD of the age-graded mean of the sample were classified as impaired. Our analysis included 8,742 participants without cognitive impairment at baseline; participants had to have baseline information on age (≥ 40 years), sex, and education (4-point ordinal scale), and two or more cognitive assessments over the study period. The follow-up ranged from 2 to 9 years. Please note that CHARLS does not release interview dates for participants; we thus used the year of assessment and age (derived from date of birth) to calculate follow-up length, which is similar to the approach taken by Gateway of Global Aging in their harmonized files.

CHARLS data are publicly available and accessible after registration at:
https://charls.charlsdata.com/pages/data/111/en.html

Zhao, Y., Hu, Y., Smith, J.P., Strauss, J., & Yang, G. Cohort profile: The China Health and Retirement Longitudinal Study (CHARLS). *Int J Epidemiol* **43**, 61-68. (2014).

The Household, Income and Labour Dynamics in Australia (HILDA)

The HILDA is a longitudinal survey of Australian households that includes household members as young as 16. The survey has been conducted annually since 2001 and includes 21 waves of data collection. Loneliness was assessed in 2001/02 with a single item, "I often feel very lonely", from *strongly disagree* to *strongly agree* (7-point scale). Measures to classify cognitive status were available for the 2012/13 and 2016/17 assessments. Following previous publications[22], CI was defined as a score equal or below 1.5 SD of the age-graded mean of the sample on the symbol digit modality test and backward digit span. Neither cognitive status nor dementia screening measures were available at the 2001/02 baseline. We categorized participants who needed a family member or friend to assist in the baseline interview and who had either "poor" or "very poor" understanding of questions as potential dementia cases and excluded them from the analyses (n = 4). Our analysis included 2,963 participants who had baseline information on age (≥ 40 years), sex, ethnicity (Aboriginal or Torres Strait Islander origin) and education (7-point ordinal scale), and a cognitive assessment at the 2012/13 and/or 2016/17 wave. The follow-up ranged from 11 to 15 years.

HILDA data are available after registration at:
https://melbourneinstitute.unimelb.edu.au/hilda/for-data-users

Summerfield, M. *et al.* HILDA User Manual – Release 18. (Melbourne Institute: Applied Economic and Social Research, 2019).

Table S2: Loneliness measures across individual cohort studies

Cohort	Description	Baseline Wave	Harmonization
HRS	3-item UCLA-R Loneliness scale (3-point response scale). Items were scored in the direction of higher loneliness and the mean taken across items.	W2006/08	Scores were standardized within each sample (M = 0, SD = 1), except for MHAS (lonely=1, not lonely=0).
ELSA		W2 (2004-05)	
SHARE (Sample 2, Dementia Sample)		W4 (2011-12)	
TILDA		W1 (2009-11)	
SHARE (Sample 1)	Single item, “How often do you feel lonely?” from *almost all the time* to *almost none of the time* (4-point ordinal scale). Responses were reverse-scored in the direction of higher loneliness.	W1 (2004-06)	
MHAS	Single item from CES-D, “In the past week… You felt lonely”, response *yes/no*.	W1 (2001)	
KLoSA	Single item from CES-D, “In the past week… You felt lonely”, from *rarely* to *almost all the time* (4-point ordinal scale).	W1 (2006)	
CHARLS		W1 (2011)	
HILDA	Single item, "I often feel very lonely", from *strongly disagree* to *strongly agree* (7-point ordinal scale).	W1 (2001)	

Table S3: Classification of cognitive status across individual cohort studies

Cohort	Cognitive Status	Study Waves
HRS	**Measure(s):** TICSm[3]. **Classification (Dementia, CIND):** The total score is based on participants' performance on a 10-word immediate and delayed memory recall, backward counting, and serial subtraction task. Scores ≤6 (out of 27) were indicative of dementia; scores between 7 and 11 were indicative of CIND.	W2006, W2008, W2010, W2012, W2014, W2016, W2018, W2020, W2022
ELSA	**Measure(s):** Short-form IQCODE[6] and self-reported doctor diagnosis; TICSm (from W7). **Classification (Dementia):** Dementia was defined as a self-reported doctor diagnosis of Alzheimer's disease or dementia or an IQCODE score ≥ 3.5, as done in prior publications[4]. **Classification (CIND):** CIND was defined as TICSm scores between 7 and 11; CIND analysis excluded ELSA participants who reported a diagnosis of dementia or an IQCODE score above the cut-off at the baseline assessment.	W2 (2004-05), W3 (2006-07), W4 (2008-09), W5 (2010-11), W6 (2012-13), W7 (2014-15), W8 (2016-17), W9 (2018-19), W10 (2021-23)
SHARE	**Measures:** Proxy and/or self-reported doctor diagnosis of Alzheimer's disease, dementia or senility (from W4); performance on a 10-word immediate and delayed memory recall and verbal fluency task. **Classification (Dementia):** Presence/absence, as done in prior publications[9]. **Classification (CI):** Participants with a score equal or below 1.5 SD of the age-graded mean of the sample, on both the memory and verbal fluency task, were classified as impaired[7].	W1 (2004-06), W2 (2006-07), W4 (2011-12), W5 (2013), W6 (2015-16), W7 (2017-18), W8 (2019-20), W9 (2021-22)
TILDA	**Measures:** MMSE[14]. **Classification (Dementia, CIND):** Scores between 24 and 30 were classified as normal cognition, scores between 18 and 23 were classified as CIND, and scores < 18 were classified as dementia[13,14].	W1 (2009-11), W2 (2012-13), W3 (2014-15), W4 (2016),W5 (2018)
KLoSA		W1 ('06), W2 ('08), W3 ('10), W4 ('12),W5 ('14), W6 ('16), W7 ('18), W8 ('20).
MHAS	**Measures:** *The study team provided a data-driven classification[17] for dementia, CIND and normal function across waves*. **Classification (Dementia, CIND):** Dementia was defined as an impairment in two or more cognitive domains and one or more instrumental activities of daily living (IADLs) limitations, or proxy respondents with IQCODE scores ≥ 3.4; CIND was defined as an impairment in two or more cognitive domains and no IADL functional limitations; unimpaired cases were self-respondents with normal cognitive function and no IADL limitations or impairment in IADLs due to physical limitations, and proxy respondents with IQCODE score < 3.4.	W1 (2001), W2 (2003), W3 (2012), W4 (2015)
CHARLS	**Measures:** A composite cognitive score was derived from the performance on 10-word immediate and delayed memory recall, serial subtraction, time orientation and a figure drawing task, as done in prior publications. **Classification (CI):** Participants with a total composite score equal or below 1.5 SD of the age-graded mean of the sample were classified as impaired[21].	W1 (2011), W2 (2013), W3 (2015), W4 (2018), W5 (2020)
HILDA	**Measures:** Performance on a symbol digit modality test and backward digit span. **Classification (CI):** Participants with a score equal or below 1.5 SD of the age-graded mean of the sample on both the symbol digit modality test and backward digit span were classified as impaired[22].	W12 (2012-13), w16 (2016-17)

Note. Coding scheme across studies and waves: 1= presence of dementia/CIND (i.e., cognitive impairment, no dementia) or CI (i.e., non-specific cognitive impairment); 0 = absence of dementia/CIND or CI.

Table S4: Selected control variables across individual cohort studies

Variable	Cohort	Description	Scale/Harmonization
Age	All cohorts	Years of age.	Continuous. Values were standardized within each sample (M = 0, SD = 1).
	*ELSA	Ages are collapsed after 90.	
	*TILDA	Ages are collapsed after 80. Participants younger than 50 are not included in the analysis because age is not released.	
	*CHARLS	Date of birth was provided in lunar calendar for most participants. We used age and solar calendar dates provided by the Gateway of Global Aging harmonized files.	
Sex	All cohorts	Biological sex at birth.	Categorical. Coded 1=female individual, 0=male individual.
Race	HRS, ELSA	White, black, other (HRS); white, non-white (ELSA).	Categorical. Coded 1=non-white, 0=white.
Ethnicity	HRS, HILDA	Hispanic/Latinx (HRS), Aboriginal or Torres Strait Islander (HILDA).	Categorical. Coded 1= ethnic minority, 0=no minority.
Education	HRS, MHAS	Years of education.	Continuous or ordinal. Values were standardized within each sample (M = 0, SD = 1).
	ELSA	Level of education, from 0-*no qualification* to 6-*degree.*	
	SHARE	Level of education (ISCED-97 classification), from 0-*pre-primary level of education* to 6-*second stage of tertiary education.*	
	TILDA	Level of education, from 1-*none/some primary* to 7-*post graduate/higher degree.*	
	KLoSA, CHARLS	Level of education, from 1-*elementary school/lower* to 4-*college/university.*	
	HILDA	Level of education, from 1-<*11 years* to 7-*postgraduate, masters or doctorate.*	
Depressive Symptoms	HRS, ELSA, MHAS	Sum of 7 symptoms (1-yes, 0-no) from the CES-D; an eighth item on loneliness was excluded from the computation (score 0-7). *MHAS included an additional item (tired) but we used the same 7 items administered in HRS*.	Continuous. Values were standardized within each sample (M = 0, SD = 1).
	SHARE	Sum of 12 symptoms (1-yes, 0-no) from the EURO-D (score 0–12)[23].	
	TILDA, KLoSA, CHARLS	CES-D 20-item (TILDA) or 10-item version (KLoSA, CHARLS); response on each item was from 0-*rarely* to 3-*almost all the time*. Items were reversed scored when needed and the mean was taken across items, excluding an item on loneliness (score 0-3).	
	HILDA	Mental Health Inventory, 5-item scale from the SF-36 general health survey[24]; response on each item was from 1-*strongly disagree* to 7-*strongly agree*. Items were reversed scored when needed and the mean was taken across items (score 1-7).	
No spouse or partner	All cohorts	Marital status.	Categorical. Coded 1= single/never married, separated, divorced, widowed, 0= married or cohabitation with a partner.
Living alone	All cohorts	Number of people living in the household either provided by the study or derived from household information.	Categorical. Coded 1 = one-person household, 0 = two or more people in the household.

No children	All cohorts	Number of living children.	Categorical. Coded 1=no children, 0=one or more children.
Diabetes	All cohorts	Self and/or proxy reported doctor diagnosis of diabetes or high blood sugar.	Categorical. Coded 1=diabetes, 0=no diabetes
Hypertension	All cohorts	Self and/or proxy reported doctor diagnosis of hypertension or high blood pressure.	Categorical. Coded 1=hypertension, 0=no hypertension
Obesity	All cohorts	BMI derived from self-reported or measured weight and height.	Categorical. Coded 1=obesity (BMI ≥ 30), 0=no obesity

Note: All covariates were derived from the baseline assessment in each study. The only exception was HILDA; for this study, information on modifiable clinical factors was not available at baseline and was derived from the first available assessment at wave 3 (diabetes and hypertension) and wave 6 (obesity).

Table S5: MOOSE checklist for meta-analyses of observational studies

Item No	Recommendation	Reported on Page No
Reporting of background should include		
1	Problem definition	3
2	Hypothesis statement	-
3	Description of study outcome(s)	3 and Methods
4	Type of exposure or intervention used	3 and Methods
5	Type of study designs used	3, 4, and Methods
6	Study population	3, 4, and Methods
Reporting of search strategy should include		
7	Qualifications of searchers (e.g., librarians and investigators)	Title Page
8	Search strategy, including time period included in the synthesis and key words	4, Methods, Extended Data Table 2
9	Effort to include all available studies, including contact with authors	Methods, Supplementary Table S6, Data Extraction File
10	Databases and registries searched	4, Methods, Extended Data Table 2
11	Search software used, name and version, including special features used (e.g., explosion)	4, Methods, Extended Data Table 2
12	Use of hand searching (e.g., reference lists of obtained articles)	4, Methods, Extended Data Table 2
13	List of citations located and those excluded, including justification	Supplementary Table S6
14	Method of addressing articles published in languages other than English	Methods, Extended Data Table 2
15	Method of handling abstracts and unpublished studies	Methods, Extended Data Table 2
16	Description of any contact with authors	Supplementary Table S6, Data Extraction File
Reporting of methods should include		
17	Description of relevance or appropriateness of studies assembled for assessing the hypothesis to be tested	4, Table 1, Extended Data Table 3 and 4, Supplementary Note S1 and Tables S2-S4
18	Rationale for the selection and coding of data (e.g., sound clinical principles or convenience)	Methods, Supplementary Table S6
19	Documentation of how data were classified and coded (e.g., multiple raters, blinding and interrater reliability)	Methods, Data Extraction File
20	Assessment of confounding (e.g., comparability of cases and controls in studies where appropriate)	4, Methods
21	Assessment of study quality, including blinding of quality assessors, stratification or regression on possible predictors of study results	Methods, Extended Data Table 6, Data Extraction File
22	Assessment of heterogeneity	Methods
23	Description of statistical methods (e.g., complete description of fixed or random effects models, justification of whether the chosen models account for predictors of study results, dose-response models, or cumulative meta-analysis) in sufficient detail to be replicated	4, Methods, Data Extraction File, R script for the meta-analysis
24	Provision of appropriate tables and graphics	Table 1-2, Figure 1-4, Extended Data Tables 1-6 and Fig. 1-2
Reporting of results should include		
25	Graphic summarizing individual study estimates and overall estimate	Figure 2-4, Extended Data Fig.1-2.
26	Table giving descriptive information for each study included	Table 1, Extended Data Table 3 and 4
27	Results of sensitivity testing (e.g., subgroup analysis)	5-6, Extended Data Table 1, Extended Data Fig. 1,

28	Indication of statistical uncertainty of findings	5-6, Extended Data Fig. 2, Supplementary Information Note S2
Reporting of discussion should include		
29	Quantitative assessment of bias (e.g., publication bias)	7
30	Justification for exclusion (e.g., exclusion of non-English language citations)	Supplementary Table S6
31	Assessment of quality of included studies	Extended Data Table 6
Reporting of conclusions should include		
32	Consideration of alternative explanations for observed results	6-7
33	Generalization of the conclusions (ie, appropriate for the data presented and within the domain of the literature review)	8
34	Guidelines for future research	8
35	Disclosure of funding source	Acknowledgments

Table S6: Final decision and reasons for exclusion of the screened full-text articles

No.	1st Author	Year	Decision
1	Dabiri	2024	**Selected.** Secondary Outcome = Cognitive Impairment. Note: The authors analyzed data from the Rush Memory and Aging Project (as Wlison *et al.* 2007). The authors examined loneliness-associated risk for MCI in addition to all-cause dementias. Because most dementia cases were AD cases, we used the estimates for AD from the original article of Wilson *et al*., but included the estimates for MCI from the article of Dabiri *et al*. DOI: 10.1016/j.bbi.2023.11.034
2	Mahalingam	2023	**Selected.** Primary Outcome = Dementia (all-cause). Secondary Outcome = Cognitive Impairment. Note: The authors provide a coordinated analysis of multiple datasets. We selected results from three cohort studies that assessed loneliness in association with dementia and cognitive imapirment. Results from ELSA were not considered because we re-analyzed ELSA data over a longer follow-up. DOI: 10.1002/alz.13072
3	Sutin	2023	**Selected.** Primary Outcome = Dementia (all-cause). Cause-specific dementias (AD, VaD, and frontotemporal dementia). DOI: 10.1017/S1041610222001028
4	Freak-Poli	2022	**Selected.** Primary Outcome = Dementia (all-cause). DOI: 10.3233/JAD-210330
5	Joyce	2022	**Selected.** Primary Outcome = Dementia (all-cases). Note: We considered the analysis with adjudicated dementia cases, not suspected cases. DOI: 10.1002/gps.5644
6	Salinas	2022	**Selected.** Primary Outcome = Dementia (all-cause). Cause-specific dementias (AD, VaD). DOI: 10.1212/WNL.0000000000200039
7	Goldberg	2021	**Selected.** Primary Outcome = Dementia (all-cause). DOI: 10.1017/S1041610221000776
8	Shibata	2021	**Selected.** Primary Outcome = Dementia (all-cause). Note: Assessed loneliness and its components (emotional and social loneliness). DOI: 10.1093/geronb/gbaa196
9	Sundström	2020	**Selected.** Primary Outcome = Dementia (all-cause). Cause-specific dementias (AD, VaD). DOI: 10.1093/geronb/gbz139
10	Zhou	2019	**Selected.** Secondary Outcome = Cognitive Impairment. Note: Sex-stratified analysis. DOI: doi:10.3390/ijerph16162877
11	Zhou	2018	**Selected.** Primary Outcome = Dementia (all-cause). DOI: 10.1080/13607863.2016.1277976
12	Holwerda	2014	**Selected.** Primary Outcome = Dementia (all-cause). DOI: 10.1136/jnnp-2012-302755
13	Chen	2011	**Selected.** Primary Outcome = Dementia (all-cause). DOI: 10.1371/journal.pone.0024817
14	Lobo	2008	**Selected.** Primary Outcome = Dementia (all-cause). Cause-specific dementias (AD). Secondary Outcome = Cognitive Impairment. Note: Unreported information for analysis with dementia; the authors were contacted but were unable to provide the data at the time of the request. This information was retrieved from a previous meta-analysis (Lara *et al*., 2019). DOI: 10.1007/BF03033815
15	Wilson	2007	**Selected.** Primary Outcome = Cause-specific dementias (AD). DOI: 10.1001/archpsyc.64.2.234
16	He	2000	**Selected.** Primary Outcome = Cause-specific dementias (AD). Note: Unreported information for analysis with dementia (all-cause). The authors were contacted but were unable to provide the data at the time of the request. URL: https://link.gale.com/apps/doc/A169717216/AONE?u=fcla_main&sid=googleScholar&xid=58ca62fb
17	Bickel	1994	**Selected.** Primary Outcome = Dementia (all-cause). Note: Unreported information for analysis with dementia subtypes. DOI: 10.1017/S0033291700026945
18	Huang	2023	**Selected for a sensitivity analysis.** Secondary Outcome = Cognitive Impairment. Analysis are repeated entering Huang *et al*. (instead Zhou *et al*. 2019). The article assesses cognitive impairment in a sample that overlaps, in part, with Zhou *et al*. Huang *et al*. report non-stratified analysis with a longer follow-up, but include exclusively adults aged over 80. Further, loneliness is assessed as categorical predictor (not lonely [ref.], sometime lonely, and often lonely). DOI: 10.1080/13607863.2022.2116396
19	Wei	2022	**Selected for a sensitivity analysis.** Secondary Outcome = Cognitive Impairment. Analysis are repeated entering Wei *et al*. (instead Zhou *et al*. 2019). The article assesses cognitive impairment using the same sample of Zhou *et al*. Wei *et al*. report results for the all sample, while Zhou *et al*. report sex-stratified analysis. However, Wei *et al*. control for a large number of covariates, which reduce comparability with other studies in the meta-analysis. DOI: 10.1186/s12877-021-02742-5
20	Rolandi	2020	**Selected for a sensitivity analysis.** Primary Outcome = Dementia (all-cause). This study is not included in the main analysis because of the statistical approach used to estimate risk which reduce comparability with other studies in the meta-analysis. DOI: 10.1186/s13195-020-00661-y

21	Wang	2020	**Selected for a sensitivity analysis.** Secondary Outcome = Cognitive Impairment. This study is not included in the main analysis because of the statistical approach used to estimate risk which reduce comparability with other studies in the meta-analysis. Further, loneliness is assessed as categorical predictor (not lonely [ref.], slightly lonely, and lonely). DOI: 10.1080/13607863.2019.1655704
22	Rawtaer	2017	**Selected for a sensitivity analysis.** Secondary Outcome = Cognitive Impairment. The article applies population norms to identify cognitive changes in MMSE, but their operationalization of impairment differs from other studies, which reduce comparability with the studies in the meta-analysis. Note: An additional outcome is MCI and dementia combined. Authors were inquired to provide separate estimate risks for MCI and dementia, but were unable at the time of the request. DOI: 10.3233/JAD-160862
23	Ma	2018	**Not included because of unreported information.** Primary Outcome = Dementia (all-cause). Authors were contacted but were unable to provide the data at the time of the request. DOI: 10.1007/s12603-018-1054-0
24	Wilson	2015	**Not included because of unreported information.** Secondary Outcome = Cognitive Impairment. Authors were contacted but were unable to provide the data at the time of the request. DOI: 10.1037/neu0000154
25	Fang	2023	**Not satisfying inclusion criteria.** Cross-sectional design. DOI: 10.1177/07334648231192053
26	Tao	2022	**Not satisfying inclusion criteria.** Not estimating risk of dementia or cognitive impairment. Further, the sample overlaps with Salinas *et al*. (2023). DOI: 10.1016/j.eclinm.2022.101643
27	De Lange	2021	**Not satisfying inclusion criteria.** Not estimating risk of dementia or cognitive impairment. Further, the sample overlaps with Sutin *et al*. (2023). DOI: 10.1016/j.bbr.2021.113510
28	Jang	2021	**Not satisfying inclusion criteria.** Cross-sectional design. DOI: 10.1186/s12877-021-02066-4
29	Burr	2020	**Not satisfying inclusion criteria.** Not estimating risk of dementia or cognitive impairment. DOI: 10.1093/geront/gnaa055
30	Griffin	2020	**Not satisfying inclusion criteria.** Not estimating risk of dementia or cognitive impairment. Further, the sample overlaps with Sutin *et al*. (2020). DOI: 10.1177/0898264318800587
31	Zhong	2017	**Not satisfying inclusion criteria.** Not estimating risk of dementia or cognitive impairment. Further, the sample overlaps with Zhou *et al*. (2018, 2019), Wei *et al*. (2022) and Huang *et al*. (2023). DOI: 10.1016/j.jagp.2015.12.009
32	Batty	2024	**Not included.** Reasons: Examine former contact sports participants. The association between loneliness and dementia is examined comparing three contact sports and the control group. medRxiv pre-print DOI: 10.1101/2024.01.15.24301327
33	Li	2023	**Not included.** Reasons: (a) Definition of the predictor, group-based changes in loneliness. (b) Potential overlap with ELSA, only uses the last three waves of data. DOI: 10.1016/j.jagp.2022.12.002
34	Elvainio	2022	**Not included.** Reasons: (a) The sample overlaps with Sutin *et al*. (2023), which used a longer follow-up of UK Biobank. (b) The measure of loneliness is derived in a way that might underestimate loneliness presence vs. absence. DOI:10.1136/bmjopen-2021-053936
35	Gardener	2022	**Not included.** Reasons: (a) Definition of the predictor, social connectivity. Loneliness is evaluated in combination with social isolation. (b) The outcome is MCI and dementia combined. DOI 10.3233/JAD-210519
36	Shen	2022	**Not included.** Reasons: (a) The sample overlaps with Sutin *et al*. (2023), which used a longer follow-up of UK Biobank. (b) The measure of loneliness is derived in a way that might underestimate loneliness presence vs. absence. DOI:10.1212/WNL.0000000000200583
37	Zhang	2022	**Not included.** Reasons: (a) The predictor is a composite index of psychological well-being, which include loneliness. No results reported for loneliness. (b) The sample overlap with Zhou *et al*. (2018, 2019), Wei *et al*. (2022) and Huang *et al*. (2023). DOI: 10.1186/s12877-021-02681-1
38	Akhter-Khan	2021	**Not included.** Reasons: (a) Definition of the predictor, loneliness is categorized as transient/persistent, which reduces comparability with other studies in the meta-analysis. (b) The sample overlaps with Salinas *et al*. (2022). DOI: 10.1002/alz.12327
39	Yang	2021	**Not included.** Reasons: (a) AD8 is used as outcome to assess risk of dementia. (b) Does not report risk estimates. DOI: 10.1080/07317115.2020.1799891
40	Kim	2021	**Not included.** Reasons: (a) Definition of the predictor, focuses on changes in loneliness. (b) Twin sample. DOI: 10.3389/fgene.2021.661474
41	Solfrizzi	2019	**Not included.** Reasons: Definition of the predictor, social frailty. Does not assess loneliness. DOI: 10.1016/j.jalz.2019.04.013
42	Poey	2017	**Not included.** Reasons: (a) Potential overlap with HRS. (b) Inclusion of a large number of covariates. (c) Unclear design or uncertain length of the follow-up. DOI:10.1093/geront/gnw154

43	Tilvis	2004	**Not included.** Reasons: Operationalization of cognitive impairment. Not based on population norms, nor on a validated cut-off. DOI: 10.1093/gerona/59.3.M268
44	Yu	2023	**Not selected.** We re-analyzed HRS data using a longer follow-up. The analysis is not comparable to other studies in the meta-analysis. DOI: 10.1093/gerona/glac213
45	Giné-Garriga	2021	**Not selected.** We re-analyzed SHARE data with a longer follow-up. DOI: 10.1016/j.maturitas.2020.11.010
46	Luchetti	2020	**Not selected.** We re-analyzed SHARE data with a longer follow-up. DOI: 10.1002/gps.5304
47	Rafnsson	2020	**Not selected.** We re-analyzed ELSA data with a longer follow-up. DOI:10.1093/geronb/gbx087
48	Sutin	2020	**Not selected.** We re-analyzed HRS data with a longer follow-up. DOI:10.1093/geronb/gby112
49	Weiss	2020	**Not selected.** We re-analyzed HRS data with a longer follow-up. The analysis is not comparable to other studies in the meta-analysis. DOI: 10.1371/journal.pone.0239994

Note: Green = selected for the main analysis; blue = selected for supplementary analysis; yellow = not included because of unreported information; orange (and light orange) = not included with reasons; light gray = data re-analyzed using a longer follow-up.

Note S2: Supplemental analyses to evaluate publication bias

PET-PEESE results

In supplemental analyses, we used the precision-effect test and precision-effect estimate with standard errors (PET-PEESE) to estimate the effect size that would be expected in a hypothetical study with a standard error of zero. PET adjusts for small-study effects (i.e., the tendency of studies with smaller samples to produce larger effect sizes) and assesses if there is a true effect beyond publication bias. If PET is significant (i.e., an indication of true effect), the additional PEESE provides a better estimate as it corrects for a non-linear association between the reported effect size and standard errors.

For all-cause dementia, PET indicated a true effect of loneliness beyond publication bias (b_0 = 0.190, $p < .05$). PEESE was thus considered (b_0 = 0.212, $p < .05$), indicating that when controlling for small-study effects, there was a significant association between loneliness and dementia. The bias-corrected effect size for PET-PEESE was slightly reduced (HR = 1.236, 95% CI [1.172,1.304]) compared to the main analysis. A similar pattern also emerged for CIND/CI, with a bias-corrected PET-PEESE effect size slightly reduced (HR = 1.125, 95% CI [1.097, 1.154]), as compared to the main meta-analytic model.

P-curve analyses

To help distinguish whether significant findings are likely or unlikely the result of selective reporting, p-curve analyses were conducted using the online app (version 4.06) provided at http://www.p-curve.com/app4/. The observed p-curve includes the statistically significant results, as the app excludes non-significant results. The p-curve analysis provides information about the evidential value: If the half p-curve test is right-skewed with $p < .05$, or both the half and full test are right-skewed with $p < .1$, then p-curve analysis indicates the presence of evidential value. The full report of the p-curve analyses for dementia and CIND/CI is attached at the end of this supplementary file, after citations. For both outcomes, the analysis indicated the presence of evidential value, providing reassurance that the significant results were unlikely due to selective reporting.

Citations

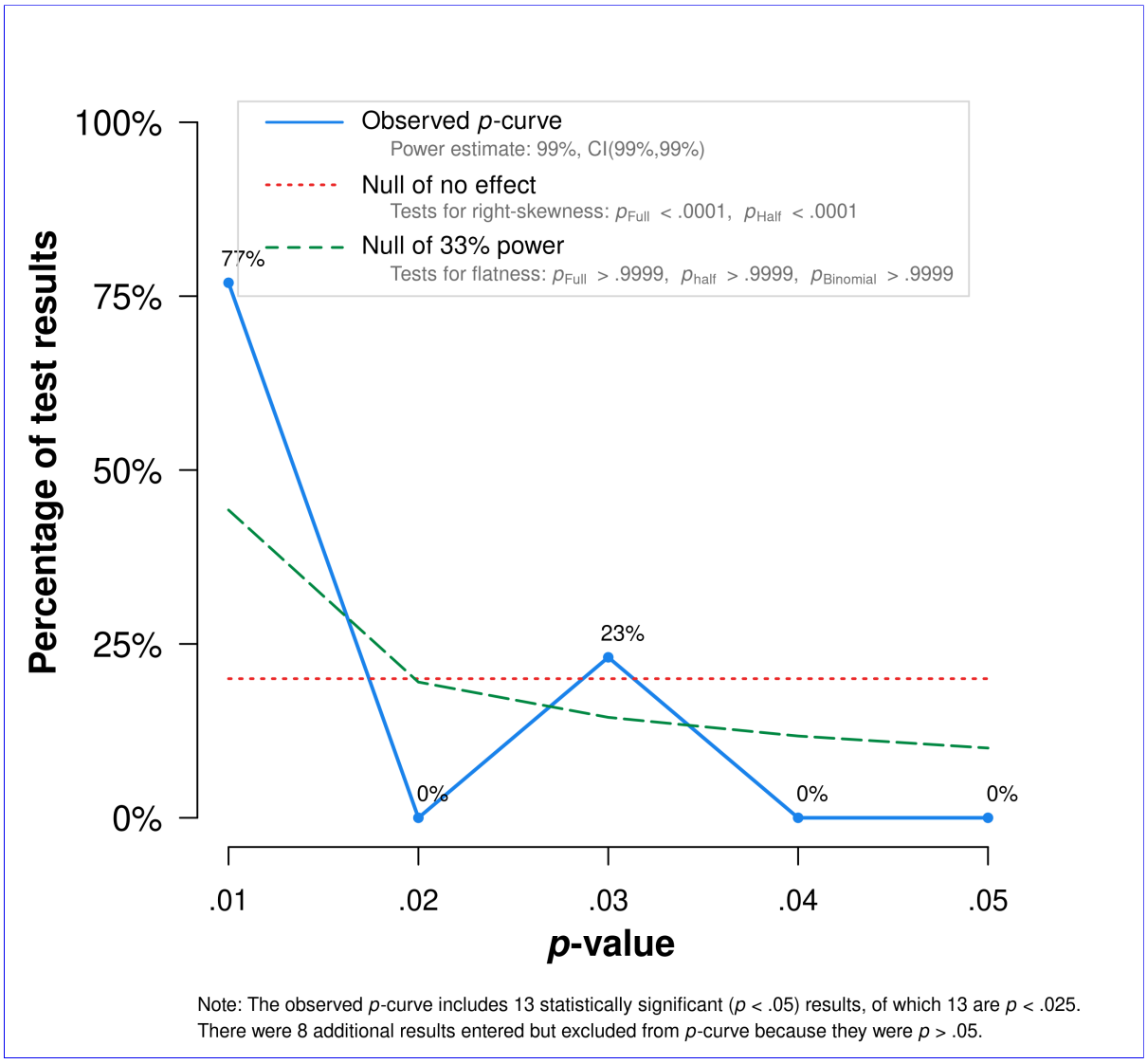
1. Willroth, E. C. & Atherton, O. E. Best laid plans: A guide to reporting preregistration deviations. *Adv Methods Pract Psychol. Sci.* **7** (2024).
2. Hughes, M. E., Waite, L. J., Hawkley, L. C. & Cacioppo, J. T. A short scale for measuring loneliness in large surveys: Results from Two population-based studies. *Res Aging* **26**, 655–672 (2004).
3. Crimmins, E. M., Kim, J. K., Langa, K. M. & Weir, D. R. Assessment of cognition using surveys and neuropsychological assessment: The Health and Retirement Study and the Aging, Demographics, and Memory Study. *J Gerontol B Psychol Sci Soc Sci.* **66B**, i162–i171 (2011).
4. Rafnsson, S. B., Orrell, M., d'Orsi, E., Hogervorst, E. & Steptoe, A. Loneliness, social integration, and incident dementia over 6 years: Prospective findings from the English Longitudinal Study of Ageing. *J Gerontol B Psychol Sci Soc Sci.* **75**, 114–124 (2020).
5. Almeida-Meza, P., Steptoe, A. & Cadar, D. Markers of cognitive reserve and dementia incidence in the English Longitudinal Study of Ageing. *Br J Psychiatry* **218**, 243–251 (2021).
6. Quinn, T. J. *et al.* Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE) for the diagnosis of dementia within community dwelling populations. *Cochrane Database Syst Rev.* CD010079 (2014).
7. Luchetti, M. *et al.* Loneliness is associated with risk of cognitive impairment in the Survey of Health, Ageing and Retirement in Europe. *Int J Geriatr Psychiatry.* **35**, 794–801 (2020).
8. Lugo-Palacios, D. G. & Gannon, B. Health care utilisation amongst older adults with sensory and cognitive impairments in Europe. *Health Econ Rev.* **7**, 44 (2017).
9. Ferreira, R. G., Brandão, M. P. & Cardoso, M. F. An update of the profile of older adults with dementia in Europe: Findings from SHARE. *Aging Ment Health.* **24**, 374–381 (2020).
10. Giné-Garriga, M. *et al.* Is loneliness a predictor of the modern geriatric giants? Analysis from the survey of health, ageing, and retirement in Europe. *Maturitas.* **144**, 93–101 (2021).
11. Gontié, R. *et al.* Relationship between physical activity and incidence of dementia in people aged 50 and over in Europe. *Aging Ment Health.* **27**, 1429–1435 (2023).
12. Folstein, M. F., Folstein, S. E. & McHugh, P. R. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res.* **12**, 189–198 (1975).
13. Bassuk, S. S., Wypij, D. & Berkmann, L. F. Cognitive Impairment and mortality in the community-dwelling elderly. *Am J Epidemiol.* **151**, 676–688 (2000).
14. Creavin, S. T. *et al.* Mini-Mental State Examination (MMSE) for the detection of Alzheimer's dementia and other dementias in asymptomatic and previously clinically unevaluated people aged over 65 years in community and primary care populations. *Cochrane Database Syst Rev.* CD011145 (2016).
15. Aguilar-Navarro, S. G., Fuentes-Cantú, A., Ávila-Funes, J. A. & García-Mayo, E. J. Validez y confiabilidad del cuestionario del ENASEM para la depresión en adultos mayores. *Salud pública Méx.* **49**, 256–262 (2007).
16. Radloff, L. S. The CES-D Scale: A self-report depression scale for research in the general population. *Appl Psychol Meas.* **1**, 385–401 (1977).
17. Michaels-Obregón, A., Mejía-Arango, S. & Wong, R. The Mexican Health and Aging Study: Cognitive function measures scoring and classification across waves 2001- 2015, Version 2. [PDF document]. (2020).
18. Mejia-Arango, S. *et al.* Effect of demographic and health dynamics on cognitive status in mexico between 2001 and 2015: Evidence from the Mexican Health and Aging Study. *Geriatrics.* **6**, 63 (2021).
19. Kim, H., Park, S.-M., Jang, S.-N. & Kwon, S. Depressive symptoms, chronic medical illness, and health care utilization: Findings from the Korean Longitudinal Study of Ageing (KLoSA). *Int Psychogeriatr.* **23**, 1285–1293 (2011).

20. Zhou, L., Ma, X. & Wang, W. Relationship between cognitive performance and depressive symptoms in chinese older adults: The China Health and Retirement Longitudinal Study (CHARLS). *J Affect Dis.* **281**, 454–458 (2021).
21. Zhang, K. & Zhang, W. Adverse childhood experiences and mild cognitive impairment in later life: Exploring Rural/urban and gender differences using CHARLS. *J Appl Gerontol.* **41**, 1454–1464 (2022).
22. Aschwanden, D., Sutin, A. R., Luchetti, M., Stephan, Y. & Terracciano, A. Personality and dementia risk in England and Australia. *GeroPsych.* **33**, 197–208 (2020).
23. Castro-Costa, E. *et al.* Ascertaining late-life depressive symptoms in Europe: An evaluation of the survey version of the EURO-D scale in 10 nations. The SHARE Project. *Int J Methods Psychiatr Res.* **17**, 12–29 (2008).
24. Ware, J. E. & Gandek, B. Overview of the SF-36 Health Survey and the International Quality of Life Assessment (IQOLA) Project. *J Clin Epidemiol.* **51**, 903–912 (1998).

Outcome = DEMENTIA

P-CURVE RESULTS - App 4.06

App's Last Update: 2017 11 30



Note: The observed *p*-curve includes 13 statistically significant ($p < .05$) results, of which 13 are $p < .025$.
There were 8 additional results entered but excluded from *p*-curve because they were $p > .05$.

The image above is in high resolution (400 dpi), you can save it and use in peer-reviewed publications.
Below we report the table previous versions of the app embedded in the image above; it includes more details than those reported within the new figure's legend.

	Binomial Test *(Share of results p<.025)*	**Continuous Test** *(Aggregate with Stouffer Method)*	
		Full p-curve (p's<.05)	**Half p-curve (p's<.025)**
1) Studies contain evidential value. *(Right skew)*	*p*=.0001	*Z*=-11.72, *p*<.0001	*Z*=-9.55, *p*<.0(
2) Studies' evidential value, if any, is inadequate. *(Flatter than 33% power)*	*p*>.9999	*Z*=8.96, *p*>.9999	*Z*=10.51, *p*>.9
Power of tests included in *p*-curve *(correcting for selective reporting)*	**Statistical Power** Estimate: 99% 90% Confidence interval: (99% , 99%)		

Interpretation:

P-Curve analysis combines the half and full p-curve to make inferences about evidential value. In particu half p-curve test is right-skewed with $p<.05$ or both the half and full test are right-skewed with $p<.1$, the analysis indicates the presence of evidential value. This combination test, introduced in Simonsohn, Sim Nelson (2015 .pdf) 'Better P-Curves' paper, is much more robust to ambitious *p*-hacking than the simple fu test is.

Here both conditions are met, indicating evidential value.

Similarly, *p*-curve analysis indicates that evidential value is inadequate or absent if the 33% power test is the full *p*-curve or both the half *p*-curve and binomial 33% power test are $p<.1$. Here neither condition is curve does not indicate evidential value is inadequate nor absent.

As with all *p*-values, these cutoffs are just benchmarks; the lower the *p*-values are, the less consistent th with the respective null hypotheses. A $p=.049$ is essentially the same as a $p=.051$, while a $p=.0001$ is m compelling than either.

To appreciate the advantage of these combination tests in relation to the previously used full *p*-curve Figure 2 and pages 1149-1151 in the 'Better P-Curves' paper (.pdf) and check out its Supplement 2 (.pdf)

Brief Explanations of Main Results:

1) **Binomial tests** compare the observed proportion of significant results that are p<.025 (in this case: 10 expected proportions when there is no effect (50%), and when studies have 1/3 power (71%). This latte varies (by a few %s) as a function of the degrees of freedom of the tests submitted to *p*-curve.

2) **Continuous tests** are obtained by computing *pp*-values for each test (probability of at least as extreme conditional on p<.05), and converting them to Z scores(N(0,1)). The sum of these Z scores (13 in this case by the square-root of the number of tests included (again: 13 in this case) is the reported Z score in th (and corresponding p-value). This approach is known as Stouffer's Method. (Prior to App 3.0 we relied c method instead. See "Better P-Curves" paper.)

Note that the binomial and continuous tests are by definition one-sided (e.g., *more* right skewed than flat negative Z values to indicate deviation in the direction of the alternative hypothesis of interest; for e negative Z value for the Right-Skew test is evidence against the flat null, and thus in favor of Right-Skew.

3) **Statistical power** is obtained by comparing the expected *p*-curve for each possible value of power be and 99% to the observed *p*-curve, and selecting the level of power that leads to the expected *p*-curve closely resembles the observed *p*-curve. (We quantify the similarity with the overall Z arising from aggre resulting pp-values via the Stouffer method, pp-values which depend on the assumed level of power). Th possible is Z=0.

Dropping Highest/Lowest *p*-values

(Cumulative meta-analysis)

In order to assess the extent to which *p*-curve's overall results hinge on a few studies, the figure belo

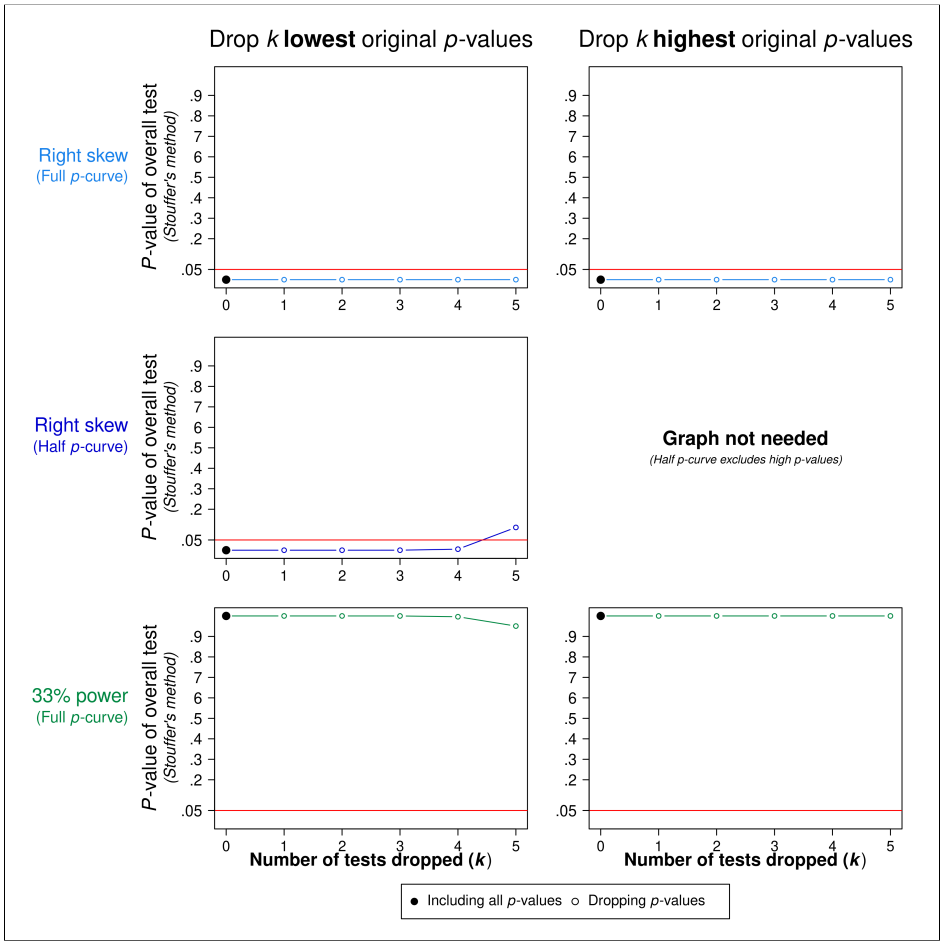
them excluding a progressively larger number of the most extreme *p*-values originally included in *p*-curve.

The first column of charts, reports results that first exclude the smallest *p*-value in *p*-curve, then th smallest, and so on.
For example, if *p*-curve contained the following four *p*-values: *p*=.001, *p*=.004, *p*=.01 and *p*=.045, the 1 would report results with all four *p*-values, the next marker when one excludes *p*=.001, then excluding bo and *p*=.004, and so on.

In the second column one proceeds in opposite order. First excluding *p*=.045, then *p*=.045 and *p*=.01, and s

The graph plots what happens until there is only half the *p*-values left, but in most situations one is only i in what happens as the single or handful of most extreme *p*-values are excluded.

We should place more confidence in sets of studies whose overall evidential value survives the exclusi most extreme few results.



Calculations for each test entered into *p*-curve:

Test entered by user	*p*-value	*pp*-values				Z Scores			
		Full *p*-curve		Half *p*-curve		Full *p*-curve		Half *p*-curve	
		Righ Skew	Power of 33%	Righ Skew	Power of 33%	Righ Skew	Power of 33%	Righ Skew	Power of 33%
Z=0.493539417	.62163	NA	NA	NA	NA	NA	NA	NA	NA
Z=0.495834111	.62001	NA	NA	NA	NA	NA	NA	NA	NA
Z=1.56211086	.11826	NA	NA	NA	NA	NA	NA	NA	NA
Z=15.40562495	<.00001	<.00001	>.9999	<.00001	>.9999	-7.76	8.13	-7.67	8.13
Z=2.807615362	.00499	.09982	.69866	.19964	.85935	-1.28	0.52	-0.84	1.08

Test entered by user	*p*-value	*pp*-values				Z Scores			
		Full *p*-curve		Half *p*-curve		Full *p*-curve		Half *p*-curve	
		Righ Skew	Power of 33%	Righ Skew	Power of 33%	Righ Skew	Power of 33%	Righ Skew	Power of 33%
Z=2.309755955	.02090	.41803	.34778	.83607	.69558	-0.21	-0.39	0.98	0.51
Z=0.758821025	.44796	NA	NA	NA	NA	NA	NA	NA	NA
Z=2.26333692	.02361	.47230	.30605	.94460	.67611	-0.07	-0.51	1.59	0.46
Z=0.237111731	.81257	NA	NA	NA	NA	NA	NA	NA	NA
Z=2.242388352	.02494	.49873	.28675	.99745	.66709	-0.00	-0.56	2.80	0.43
Z=2.936191013	.00332	.06645	.76112	.13291	.88851	-1.50	0.71	-1.11	1.22
Z=3.835029345	.00013	.00251	.96837	.00502	.98524	-2.81	1.86	-2.57	2.18
Z=5.160483472	<.00001	<.00001	.99958	.00001	.99980	-4.42	3.34	-4.27	3.54
Z=1.240607625	.21475	NA	NA	NA	NA	NA	NA	NA	NA
Z=1.830743528	.06714	NA	NA	NA	NA	NA	NA	NA	NA
Z=1.415984692	.15678	NA	NA	NA	NA	NA	NA	NA	NA
Z=6.047239366	<.00001	<.00001	>.9999	<.00001	>.9999	-5.42	4.28	-5.30	4.45
Z=3.87093315	.00011	.00217	.97126	.00434	.98658	-2.85	1.90	-2.62	2.21
Z=8.571577134	<.00001	<.00001	>.9999	<.00001	>.9999	-7.76	6.89	-7.67	7.00
Z=3.241730389	.00119	.02376	.86997	.04752	.93931	-1.98	1.13	-1.67	1.55
Z=6.745671127	<.00001	<.00001	>.9999	<.00001	>.9999	-6.19	5.01	-6.08	5.15
SUM of Z-Scores in column, dividing by sqrt(N of tests) Z Scores reported under p-curve figure above----->						**-11.72**	**8.96**	**-9.55**	**10.51**

Explaining these calculations with an example:

Take the first significant result entered: **Z=15.40562495**. It is associated with a two-sided *p*-value of **0**. *pp*-v the probability of at least as extreme a significant p-value. For right skew we compute these under the effect; because *p*-values would be distributed uniform between 0 and .05, we simply divide by .05 (multi and get the *pp*-value for right skew, that is 0*20=**0**. One minus that gives us the *pp*-value for left skew (r above).

For the *pp*-value under the null that the test is powered to 33% things are a bit more complicated. This ex will not be quite enough, but: we find the non-centrality parameter for the corresponding distribution an of freedom that gives 33% power. We then evaluate in that non-central distribution the observed tes Z=15.40562495, and now divide by 33% rather than 5% because now 1/3 of tests are expected to be *p*< than only 5% of them.

More importantly, the interpretation of the *pp*-value for 33% power is as follows. If the underlying effect big enough to give the sample of the study obtaining Z=15.40562495 33% power, then with probability **1** get a *p*-value of 0 or higher.

For the half *p*-curve we proceed similarly. First, for right skew we divide by .025 (multiply by 40). When a >.025 it is not included in half p-curve, we see "NA" in the table above. For 33% power, in turn, we use non-centrality parameter but this time we divide by the share of p-values expected to be p<.025 when 33%.

The last four columns report the Z-Scores associated with those *pp*-values. So for the full *p*-curve right value we had **pp=0**. Evaluating the standard normal distribution in that percentile gives us the reported **Z=**

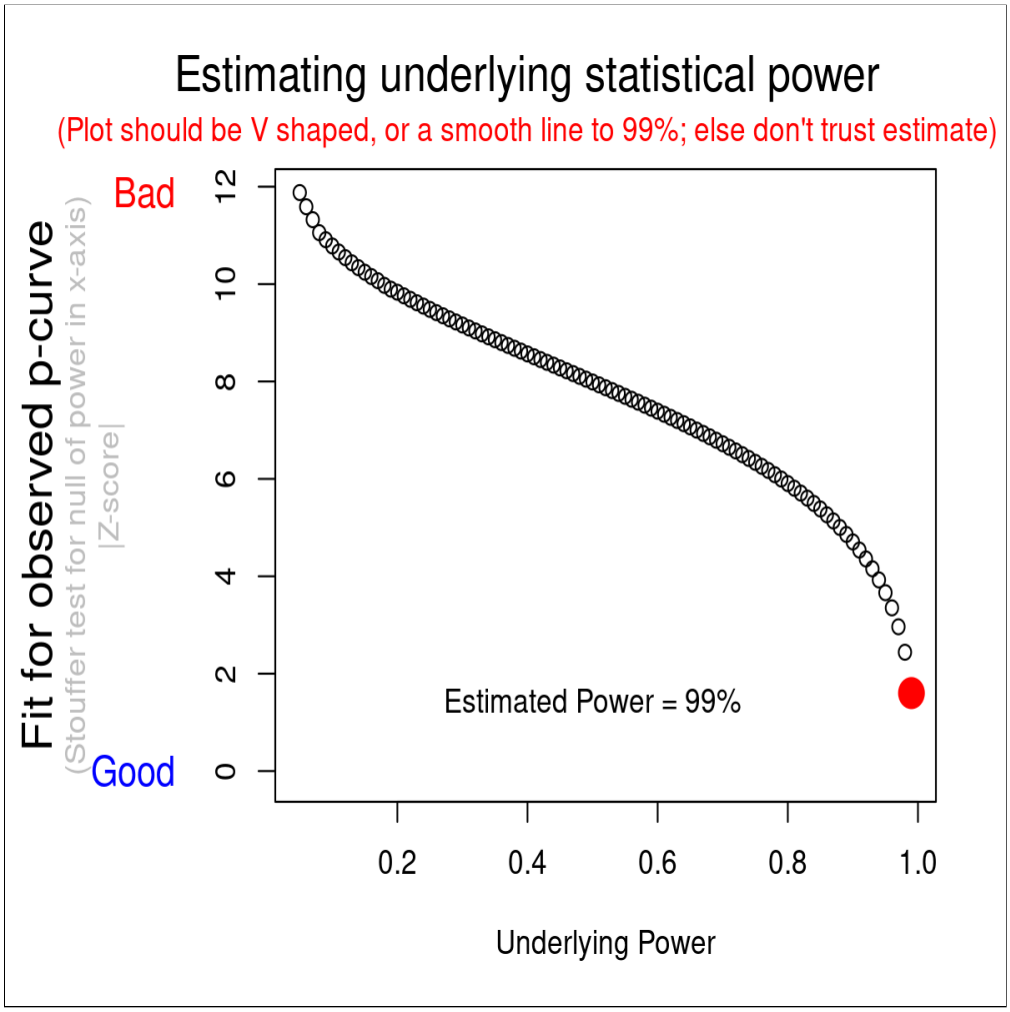
Note: when a *p*-value or p*p*-value is smaller than 2.22e-16 the app uses that value instead. The calcul hence only approximate for extremely significant results (e.g., t(98)=7.12).

Diagnostic plot for power estimation

This figure plots how consistent the observed *p*-curve is with each possible value of power between 5% ar

To create the figure we compute *pp*-values for the null that all studies are powered with a given level of p combine those pp-values using Stouffer's method. The best fitting level of power will lead to an overa Z=0, p=.5.

This approach is different from the one used with App 3.0 where instead the Kolmogorov-Smirnov test w the resulting distribution of pp-values and the uniform. The results with both methods are very similar. advantage of the KS test approach is that it reports absolute fit between expected and observed *p*-curve. advantage of the Stouffer method is that it is the approach used to compute the confidence interval an more parsimonious.
The table with results at the top of this page reports **99%** as the estimate of power. This means that if all the set were truly powered to 99%, half the time we would see a flatter p-curve than the one we see, an time we would see a more right-skewed one. So 99% is our best guess.



Confidence interval for power

To build the confidence interval for power we proceed as we do to obtain the estimate of power, but ra finding the underlying statistical power that leads to an overall Stouffer test combining the resulting *pp* *p*=.5, we find the level of power that gives *p*=.05 and *p*=.95.

For example, above we saw that the lower end of the confidence interval for power was **99%**. This means assume that's the level of power we would observe a p-curve this right-skewed, or more right-skewed, a by the Stouffer combination of the resulting pp-values, only 5% of the time. The other end of the c

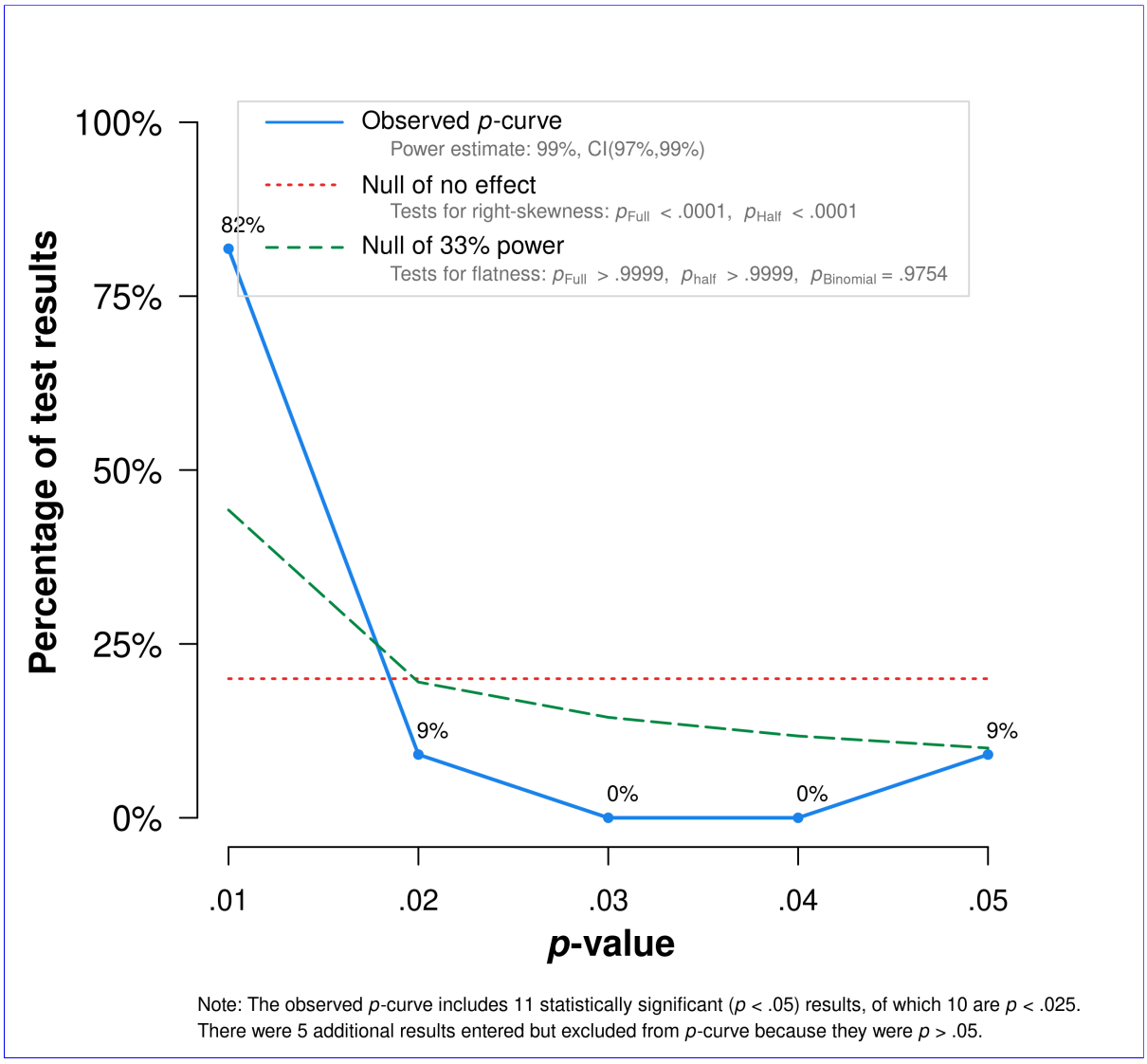
interval (**99%**), in turn, means that if power were that high, we would see as flat a p-curve, or flatter, 9 time. Note that this is a 90% confidence interval (for a 95% one, we would look for levels of power l overall p-values of 2.5% and 97.5% respectively). We use 90% to make it consistent with the one-sided te the 33% power null. If p-curve is significantly flatter than expected with 33% power, then the (90%) c interval for power will not include 33% power.

Thank you for using the *p*-curve app.

Outcome = CIND/CI

P-CURVE RESULTS - App 4.06

App's Last Update: 2017 11 30



Note: The observed *p*-curve includes 11 statistically significant ($p < .05$) results, of which 10 are $p < .025$.
There were 5 additional results entered but excluded from *p*-curve because they were $p > .05$.

The image above is in high resolution (400 dpi), you can save it and use in peer-reviewed publications.
Below we report the table previous versions of the app embedded in the image above; it includes more details than those reported within the new figure's legend.

	Binomial Test *(Share of results p<.025)*	**Continuous Test** *(Aggregate with Stouffer Method)*	
		Full p-curve (p's<.05)	**Half p-curve (p's<.025)**
1) Studies contain evidential value. *(Right skew)*	*p*=.0059	*Z*=-11.2, *p*<.0001	*Z*=-11.36, *p*<.0(
2) Studies' evidential value, if any, is inadequate. *(Flatter than 33% power)*	*p*=.9754	*Z*=8.03, *p*>.9999	*Z*=9.8, *p*>.999
Power of tests included in *p*-curve *(correcting for selective reporting)*		**Statistical Power** Estimate: 99% 90% Confidence interval: (97% , 99%)	

Interpretation:

P-Curve analysis combines the half and full p-curve to make inferences about evidential value. In particu half p-curve test is right-skewed with $p<.05$ or both the half and full test are right-skewed with $p<.1$, the analysis indicates the presence of evidential value. This combination test, introduced in Simonsohn, Sim Nelson (2015 .pdf) 'Better P-Curves' paper, is much more robust to ambitious *p*-hacking than the simple fu test is.

Here both conditions are met, indicating evidential value.

Similarly, *p*-curve analysis indicates that evidential value is inadequate or absent if the 33% power test is the full *p*-curve or both the half *p*-curve and binomial 33% power test are $p<.1$. Here neither condition is curve does not indicate evidential value is inadequate nor absent.

As with all *p*-values, these cutoffs are just benchmarks; the lower the *p*-values are, the less consistent th with the respective null hypotheses. A $p=.049$ is essentially the same as a $p=.051$, while a $p=.0001$ is m compelling than either.

To appreciate the advantage of these combination tests in relation to the previously used full *p*-curve Figure 2 and pages 1149-1151 in the 'Better P-Curves' paper (.pdf) and check out its Supplement 2 (.pdf)

Brief Explanations of Main Results:

1) **Binomial tests** compare the observed proportion of significant results that are p<.025 (in this case: 91 expected proportions when there is no effect (50%), and when studies have 1/3 power (71%). This latte varies (by a few %s) as a function of the degrees of freedom of the tests submitted to *p*-curve.

2) **Continuous tests** are obtained by computing *pp*-values for each test (probability of at least as extreme conditional on p<.05), and converting them to Z scores(N(0,1)). The sum of these Z scores (11 in this case by the square-root of the number of tests included (again: 11 in this case) is the reported Z score in th (and corresponding p-value). This approach is known as Stouffer's Method. (Prior to App 3.0 we relied c method instead. See "Better P-Curves" paper.)

Note that the binomial and continuous tests are by definition one-sided (e.g., *more* right skewed than flat negative Z values to indicate deviation in the direction of the alternative hypothesis of interest; for e negative Z value for the Right-Skew test is evidence against the flat null, and thus in favor of Right-Skew.

3) **Statistical power** is obtained by comparing the expected *p*-curve for each possible value of power be and 99% to the observed *p*-curve, and selecting the level of power that leads to the expected *p*-curve closely resembles the observed *p*-curve. (We quantify the similarity with the overall Z arising from aggre resulting pp-values via the Stouffer method, pp-values which depend on the assumed level of power). Th possible is Z=0.

Dropping Highest/Lowest *p*-values

(Cumulative meta-analysis)

In order to assess the extent to which *p*-curve's overall results hinge on a few studies, the figure belo

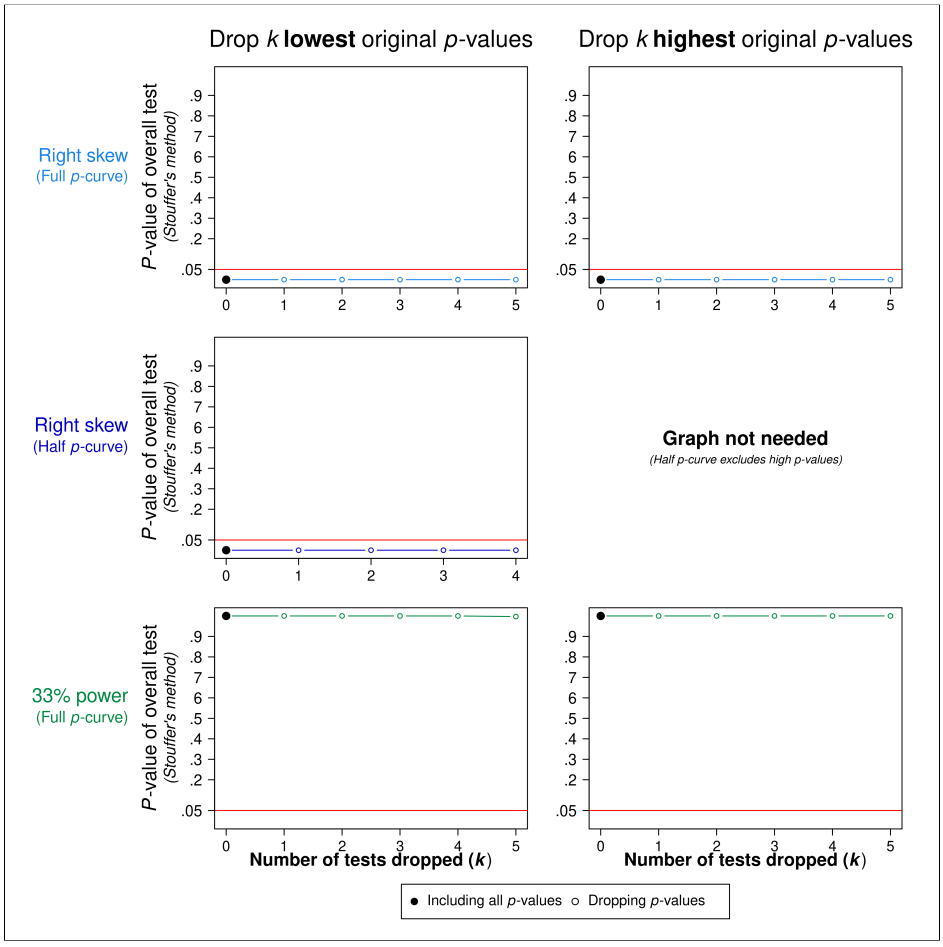
them excluding a progressively larger number of the most extreme *p*-values originally included in *p*-curve.

The first column of charts, reports results that first exclude the smallest *p*-value in *p*-curve, then th smallest, and so on.
For example, if *p*-curve contained the following four *p*-values: *p*=.001, *p*=.004, *p*=.01 and *p*=.045, the 1 would report results with all four *p*-values, the next marker when one excludes *p*=.001, then excluding bo and *p*=.004, and so on.

In the second column one proceeds in opposite order. First excluding *p*=.045, then *p*=.045 and *p*=.01, and s

The graph plots what happens until there is only half the *p*-values left, but in most situations one is only i in what happens as the single or handful of most extreme *p*-values are excluded.

We should place more confidence in sets of studies whose overall evidential value survives the exclusi most extreme few results.



Calculations for each test entered into *p*-curve:

Test entered by user	*p*-value	*pp*-values				Z Scores			
		Full *p*-curve		Half *p*-curve		Full *p*-curve		Half *p*-curve	
		Righ Skew	Power of 33%	Righ Skew	Power of 33%	Righ Skew	Power of 33%	Righ Skew	Power of 33%
Z=4.984632098	<.00001	.00001	.99918	.00002	.99962	-4.22	3.15	-4.06	3.36
Z=1.607246247	.10800	NA	NA	NA	NA	NA	NA	NA	NA
Z=1.530138198	.12598	NA	NA	NA	NA	NA	NA	NA	NA
Z=0.136881143	.89112	NA	NA	NA	NA	NA	NA	NA	NA
Z=0.1012981	.91931	NA	NA	NA	NA	NA	NA	NA	NA

Test entered by user	*p*-value	*pp*-values				Z Scores			
		Full *p*-curve		Half *p*-curve		Full *p*-curve		Half *p*-curve	
		Righ Skew	Power of 33%	Righ Skew	Power of 33%	Righ Skew	Power of 33%	Righ Skew	Power of 33%
Z=2.527884498	.01148	.22950	.52346	.45901	.77758	-0.74	0.06	-0.10	0.76
Z=3.164002022	.00156	.03112	.84706	.06225	.92862	-1.86	1.02	-1.54	1.47
Z=4.631887914	<.00001	.00007	.99713	.00014	.99866	-3.80	2.76	-3.62	3.00
Z=5.119681463	<.00001	.00001	.99951	.00001	.99977	-4.37	3.29	-4.22	3.50
Z=4.875286994	<.00001	.00002	.99877	.00004	.99943	-4.09	3.03	-3.92	3.25
Z=6.374690412	<.00001	<.00001	>.9999	<.00001	>.9999	-5.78	4.62	-5.67	4.78
Z=3.887501499	.00010	.00203	.97251	.00405	.98717	-2.87	1.92	-2.65	2.23
Z=1.354785269	.17549	NA	NA	NA	NA	NA	NA	NA	NA
Z=7.767049646	<.00001	<.00001	>.9999	<.00001	>.9999	-7.29	6.06	-7.19	6.19
Z=4.106605472	.00004	.00080	.98510	.00161	.99304	-3.15	2.17	-2.95	2.46
Z=2.027637033	.04260	.85195	.07284	NA	NA	1.04	-1.45	NA	NA
SUM of Z-Scores in column, dividing by sqrt(N of tests) Z Scores reported under p-curve figure above----->						**-11.2**	**8.03**	**-11.36**	**9.8**

Explaining these calculations with an example:

Take the first significant result entered: **Z=4.984632098**. It is associated with a two-sided *p*-value of **0**. *pp*-v the probability of at least as extreme a significant p-value. For right skew we compute these under the effect; because *p*-values would be distributed uniform between 0 and .05, we simply divide by .05 (multi and get the *pp*-value for right skew, that is 0*20=**1.0E-5**. One minus that gives us the *pp*-value for left shown above).

For the *pp*-value under the null that the test is powered to 33% things are a bit more complicated. This ex will not be quite enough, but: we find the non-centrality parameter for the corresponding distribution an of freedom that gives 33% power. We then evaluate in that non-central distribution the observed tes Z=4.984632098, and now divide by 33% rather than 5% because now 1/3 of tests are expected to be *p*< than only 5% of them.

More importantly, the interpretation of the *pp*-value for 33% power is as follows. If the underlying effect big enough to give the sample of the study obtaining Z=4.984632098 33% power, then with probability **0.** would get a *p*-value of 0 or higher.

For the half *p*-curve we proceed similarly. First, for right skew we divide by .025 (multiply by 40). When a >.025 it is not included in half p-curve, we see "NA" in the table above. For 33% power, in turn, we use non-centrality parameter but this time we divide by the share of p-values expected to be p<.025 when 33%.

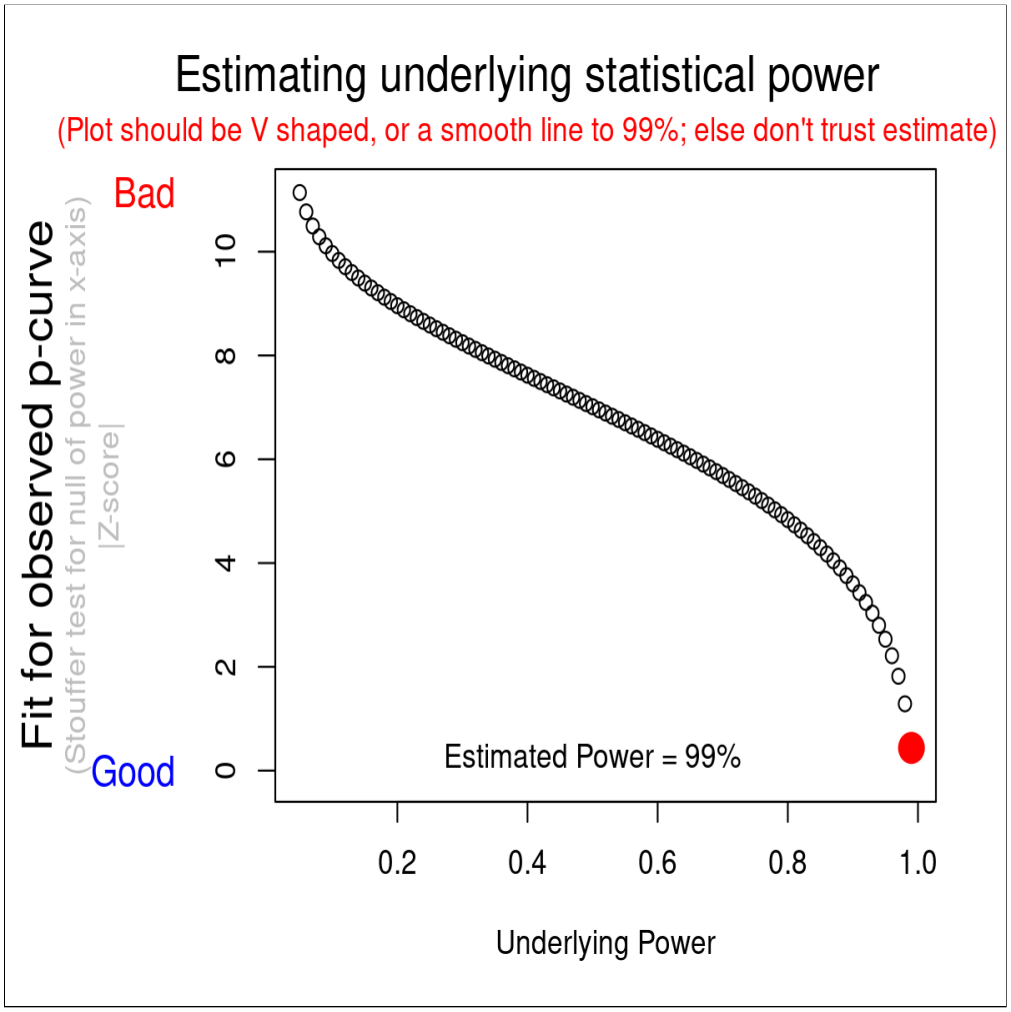
The last four columns report the Z-Scores associated with those *pp*-values. So for the full *p*-curve right value we had **pp=1.0E-5**. Evaluating the standard normal distribution in that percentile gives us the **Z=-4.22**.

Note: when a *p*-value or p*p*-value is smaller than 2.22e-16 the app uses that value instead. The calcul hence only approximate for extremely significant results (e.g., t(98)=7.12).

Diagnostic plot for power estimation

This figure plots how consistent the observed *p*-curve is with each possible value of power between 5% ar To create the figure we compute *pp*-values for the null that all studies are powered with a given level of p combine those pp-values using Stouffer's method. The best fitting level of power will lead to an overa Z=0, p=.5.

This approach is different from the one used with App 3.0 where instead the Kolmogorov-Smirnov test w
the resulting distribution of pp-values and the uniform. The results with both methods are very similar.
advantage of the KS test approach is that it reports absolute fit between expected and observed *p*-curve.
advantage of the Stouffer method is that it is the approach used to compute the confidence interval anc
more parsimonious.
The table with results at the top of this page reports **99%** as the estimate of power. This means that if all
the set were truly powered to 99%, half the time we would see a flatter p-curve than the one we see, an
time we would see a more right-skewed one. So 99% is our best guess.



Confidence interval for power

To build the confidence interval for power we proceed as we do to obtain the estimate of power, but ra
finding the underlying statistical power that leads to an overall Stouffer test combining the resulting *pp*
p=.5, we find the level of power that gives *p*=.05 and *p*=.95.

For example, above we saw that the lower end of the confidence interval for power was **97%**. This means
assume that's the level of power we would observe a p-curve this right-skewed, or more right-skewed, a
by the Stouffer combination of the resulting pp-values, only 5% of the time. The other end of the c
interval (**99%**), in turn, means that if power were that high, we would see as flat a p-curve, or flatter, 9
time. Note that this is a 90% confidence interval (for a 95% one, we would look for levels of power l
overall p-values of 2.5% and 97.5% respectively). We use 90% to make it consistent with the one-sided te

the 33% power null. If p-curve is significantly flatter than expected with 33% power, then the (90%) c interval for power will not include 33% power.

Thank you for using the *p*-curve app.