Evidence-based medicine vs precision medicine: fighting with small and big data

John P.A. Ioannidis, MD, DSc

Berlin, March 2019

C.F. Rehnborg Professor in Disease Prevention
Professor of Medicine, of Health Research and Policy, of Biomedical Data Science, and of Statistics, Stanford University
Co-Director, Meta-Research Innovation Center at Stanford (METRICS)
Director, PhD program in Epidemiology and Clinical Research
PI, Wellness Living Laboratory

Evidence-based medicine

- David Sackett definition, 1996 = "the conscientious, explicit and judicious use of **current best evidence** in making decisions about the **care of <u>individual</u> patients**. ... [It] means integrating **<u>individual clinical expertise</u>** with the **best available external clinical evidence** from **systematic research**."

- Evidence-based medicine is individualized, precision-medicine from its very beginning

Precision medicine (health)

- is a medical model that proposes the customization of healthcare, with medical decisions, practices, or products being tailored to the individual patient (Wikipedia definition)

- Individual = 1/Population

- By definition, precision medicine is aiming to have the most tiny and the most negligible impact possible at a population level.

Precision medicine (or health) = the study of the most Insignificant

Και πολλά μέλλει να μάθεις αν το Ασήμαντο εμβαθύνεις

You'll come to learn a great deal if you study the Insignificant in depth

Odysseus Elytis, Nobel prize for literature 1979
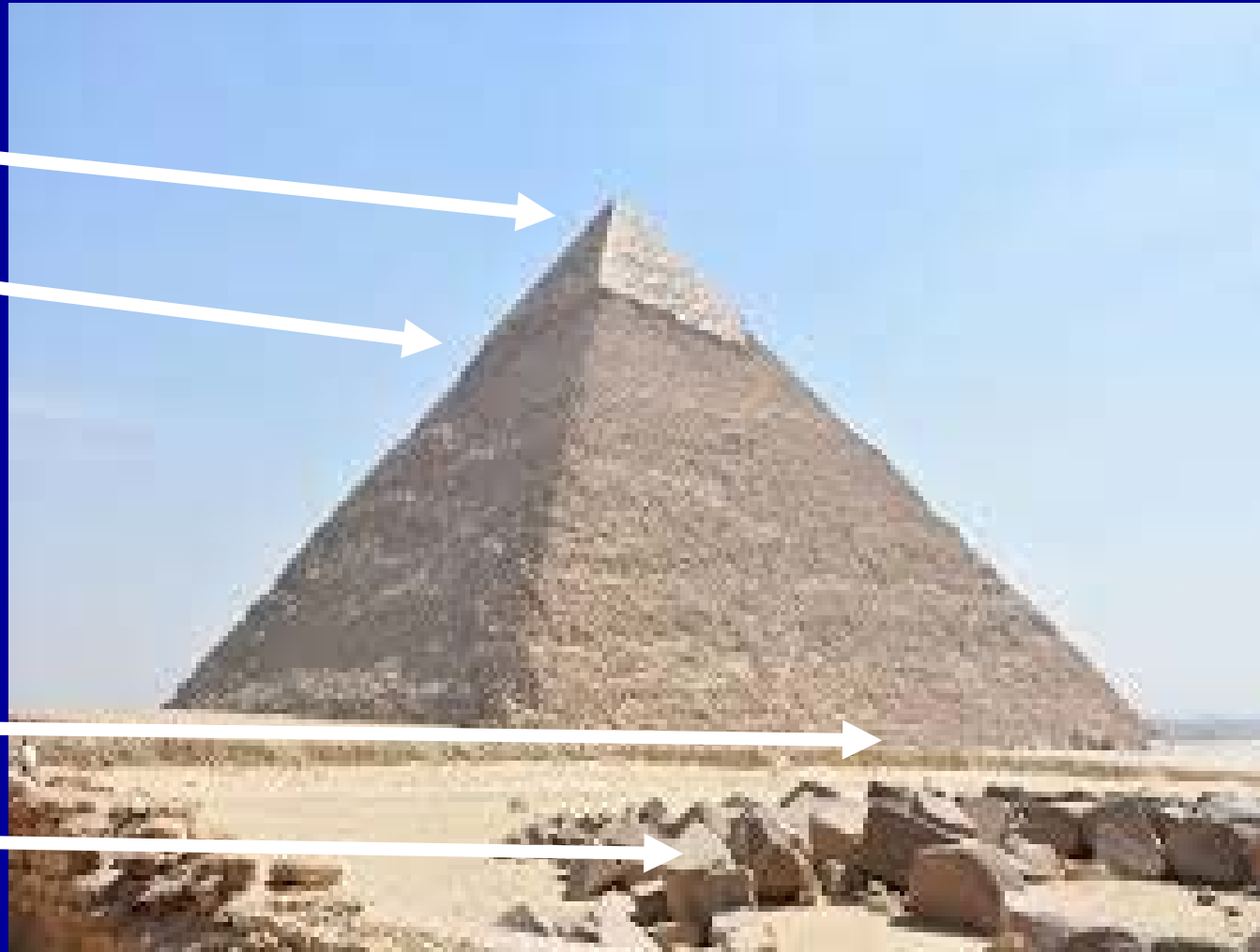
Hierarchies of evidence

MA, SR

RCT

…

…

…

Experts

Tweets

Multiple types of evidence

Clinical evidence

Observational evidence

Mechanistic evidence

Other evidence

Lots of sand

Evidence is less than optimal

Destroyed pyramid in Abu Rawash

How good is the quality of the clinical evidence?

- All 1394 systematic reviews published on the Cochrane Database of Systematic Reviews from January 2013 to June, 2014.

- GRADE (Grades of Recommendation, Assessment, Development, and Evaluation) summary of findings performed in 608 (43.6%).

- Quality of the evidence for the first listed primary outcome: 13.5% high, 30.8% moderate, 31.7% low, 24% very low level.

- Even when all outcomes listed were considered, only 19.1% had at least one outcome with high quality of evidence.

- Of the reviews with high quality of evidence, only 25 had both significant results and a favorable interpretation of the intervention.

Bulldozed pyramid by property developers in Peru (money, money, money!)

Re-analysis: can we trust the data?

Restoring Study 329: efficacy and harms of paroxetine and imipramine in treatment of major depression in adolescence

Joanna Le Noury,[1] John M Nardo,[2] David Healy,[1] Jon Jureidini,[3] Melissa Raven,[3] Catalin Tufanaru,[4] Elia Abi-Jaoude[5]

ABSTRACT

OBJECTIVES
To reanalyse SmithKline Beecham's Study 329 (published by Keller and colleagues in 2001), the primary objective of which was to compare the efficacy and safety of paroxetine and imipramine with placebo in the treatment of adolescents with unipolar major depression. The reanalysis under the restoring invisible and abandoned trials (RIAT) initiative was done to see whether access to and reanalysis of a full dataset from a randomised controlled trial would have clinically relevant implications for evidence based medicine.

DESIGN
Double blind randomised placebo controlled trial.

SETTING
12 North American academic psychiatry centres, from 20 April 1994 to 15 February 1998.

PARTICIPANTS
275 adolescents with major depression of at least eight weeks in duration. Exclusion criteria included a range of comorbid psychiatric and medical disorders and suicidality.

INTERVENTIONS
Participants were randomised to eight weeks double blind treatment with paroxetine (20-40 mg), imipramine (200-300 mg), or placebo.

MAIN OUTCOME MEASURES
The prespecified primary efficacy variables were change from baseline to the end of the eight week acute treatment phase in total Hamilton depression scale (HAM-D) score and the proportion of responders (HAM-D score ≤8 or ≥50% reduction in baseline HAM-D) at acute endpoint. Prespecified secondary outcomes were changes from baseline to endpoint in depression items in K-SADS-L, clinical global impression, autonomous functioning checklist, self-perception profile, and sickness impact scale; predictors of response; and number of patients who relapse during the maintenance phase. Adverse experiences were to be compared primarily by using descriptive statistics. No coding dictionary was prespecified.
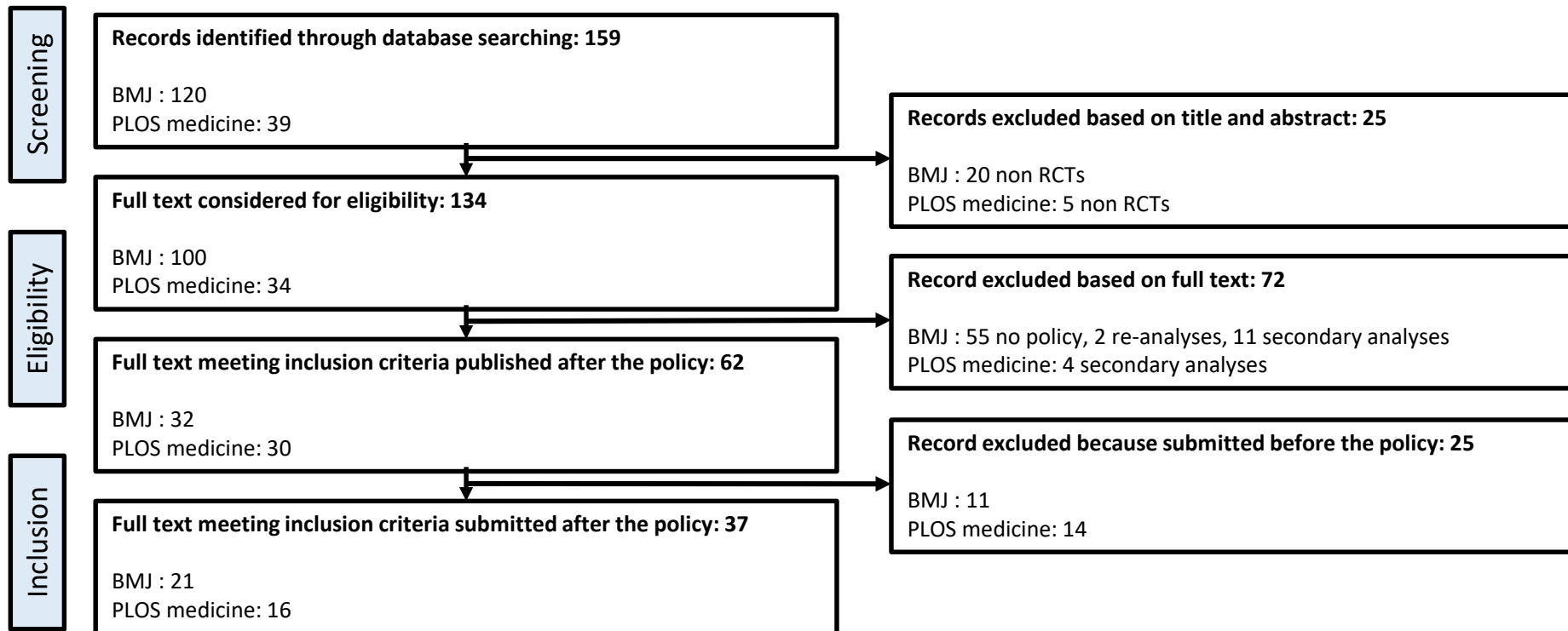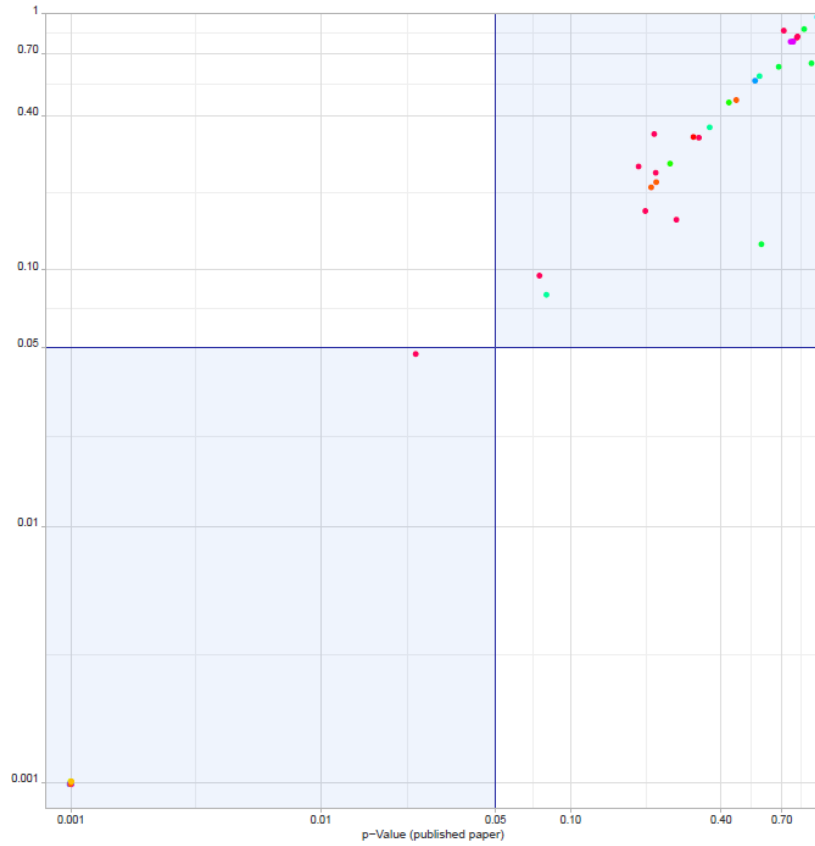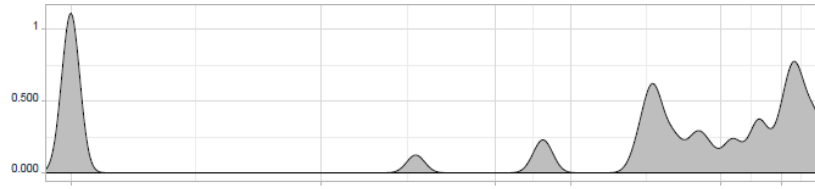
RESULTS
The efficacy of paroxetine and imipramine was not statistically or clinically significantly different from placebo for any prespecified primary or secondary efficacy outcome. HAM-D scores decreased by 10.7 (least squares mean) (95% confidence interval 9.1 to 12.3), 9.0 (7.4 to 10.5), and 9.1 (7.5 to 10.7) points, respectively, for the paroxetine, imipramine and placebo groups (P=0.20). There were clinically significant increases in harms, including suicidal ideation and behaviour and other serious adverse events in the paroxetine group and cardiovascular problems in the imipramine group.

CONCLUSIONS
Neither paroxetine nor high dose imipramine showed efficacy for major depression in adolescents, and there was an increase in harms with both drugs. Access to primary data from trials has important implications for both clinical practice and research, including that published conclusions about efficacy and safety should not be read as authoritative. The reanalysis of Study 329 illustrates the necessity of making primary trial data and protocols available to increase the rigour of the evidence base.

46% retrieval rate for raw data of randomized trials under full data sharing policy

Screening

Records identified through database searching: 159

BMJ : 120
PLOS medicine: 39

Records excluded based on title and abstract: 25

BMJ : 20 non RCTs
PLOS medicine: 5 non RCTs

Full text considered for eligibility: 134

BMJ : 100
PLOS medicine: 34

Eligibility

Record excluded based on full text: 72

BMJ : 55 no policy, 2 re-analyses, 11 secondary analyses
PLOS medicine: 4 secondary analyses

Full text meeting inclusion criteria published after the policy: 62

BMJ : 32
PLOS medicine: 30

Record excluded because submitted before the policy: 25

BMJ : 11
PLOS medicine: 14

Inclusion

Full text meeting inclusion criteria submitted after the policy: 37

BMJ : 21
PLOS medicine: 16

Naudet et al, BMJ 2018

Precision on top? E.g. N-of-1 trials were placed at the top in the mid-90s

Why were N-of-1 trials largely abandoned 25 years ago?

- Not good if the disease does not have a steady natural history

- Not good if there is carry over effect

- Not good if there are priming effects and if effects depend on previous choices

- Not good if the disease has a fatal outcome and a relatively short course

- Not good if there is poor/unpredictable compliance/adherence/tolerability

Big data (my definition)

- Data that carries the least possible information content per unit

- The more insignificant the content of information per unit, the bigger the big data

- The exact opposite of Bradford Hill ("back of the envelope") type of information

"The end of theory: The data deluge makes the scientific method obsolete" (WIRED)

VIEWPOINT

Stealth Research
Is Biomedical Innovation Happening Outside the Peer-Reviewed Literature?

Ioannidis, JAMA, 2015; and JAMA 2016

Stealth research: Lack of peer-reviewed evidence from healthcare unicorns

Ioana A. Cristea[1,2] | Eli M. Cahan[3,4] | John P. A. Ioannidis[1,5,6,7,8]

Key messages

- Start-ups are widely accepted as key vehicles of innovation and disruption in healthcare, positioned to make revolutionary discoveries.
- Most of the highest-valued start-ups in healthcare have a limited or non-existent participation and impact in the publicly available scientific literature.
- The system of peer-reviewed publishing, while imperfect, is indispensable for validating innovative products and technologies in biomedicine.
- Healthcare products not subjected to peer-review but based on internal data generation alone may be problematic and non-trustworthy.

In the Era of Precision Medicine and Big Data, Who Is Normal?

Arjun K. Manrai, PhD
Department of
Biomedical Informatics,
Harvard Medical
School, Boston,
Massachusetts.

Chirag J. Patel, PhD
Department of
Biomedical Informatics,
Harvard Medical
School, Boston,
Massachusetts.

John P. A. Ioannidis, MD, DSc
Stanford Prevention
Research Center,
Stanford University,
Stanford, California;
Departments of
Medicine, Health

The definition of "normal" values for common laboratory tests often governs the diagnosis, treatment, and overall management of tested individuals. Some test results may depend on demographic traits of the tested population including age, race, and sex. Ideally, laboratory test results should be interpreted in reference to a population of "similar" "healthy" individuals. In many settings, it is unclear exactly who these individuals are. How much population stratification and what criteria for healthy individuals are optimal? In particular, with the evolution of medicine into fully personalized or "precision" medicine and the availability of large-scale data sets, there may be interest in trying to match each person to an increasingly granular normal reference population. Is this precision feasible to obtain in reliable ways and will it improve practice?

There are limited systematic analyses of baseline variation across demographically diverse population

lation strata be overcome as the normal population becomes more precise and personalized?

It is essential to answer these questions for widely used clinical laboratory tests such as complete blood count and blood chemistries before delving into more rare tests. Such tests are a routine entry point for invasive and expensive follow-up tests and procedures, yet remain poorly characterized across strata. Data sets sufficiently capacious to study stratified variation at scale include select research cohorts, electronic health records, and insurance claims data sets. Although some data sets may be queried with relative ease (eg, electronic health records at an investigator's institution or public claims data), how generalizable findings are to other clinical settings is unclear.[4]

Challenges of Precision Medicine and Big Data
Defining Normality
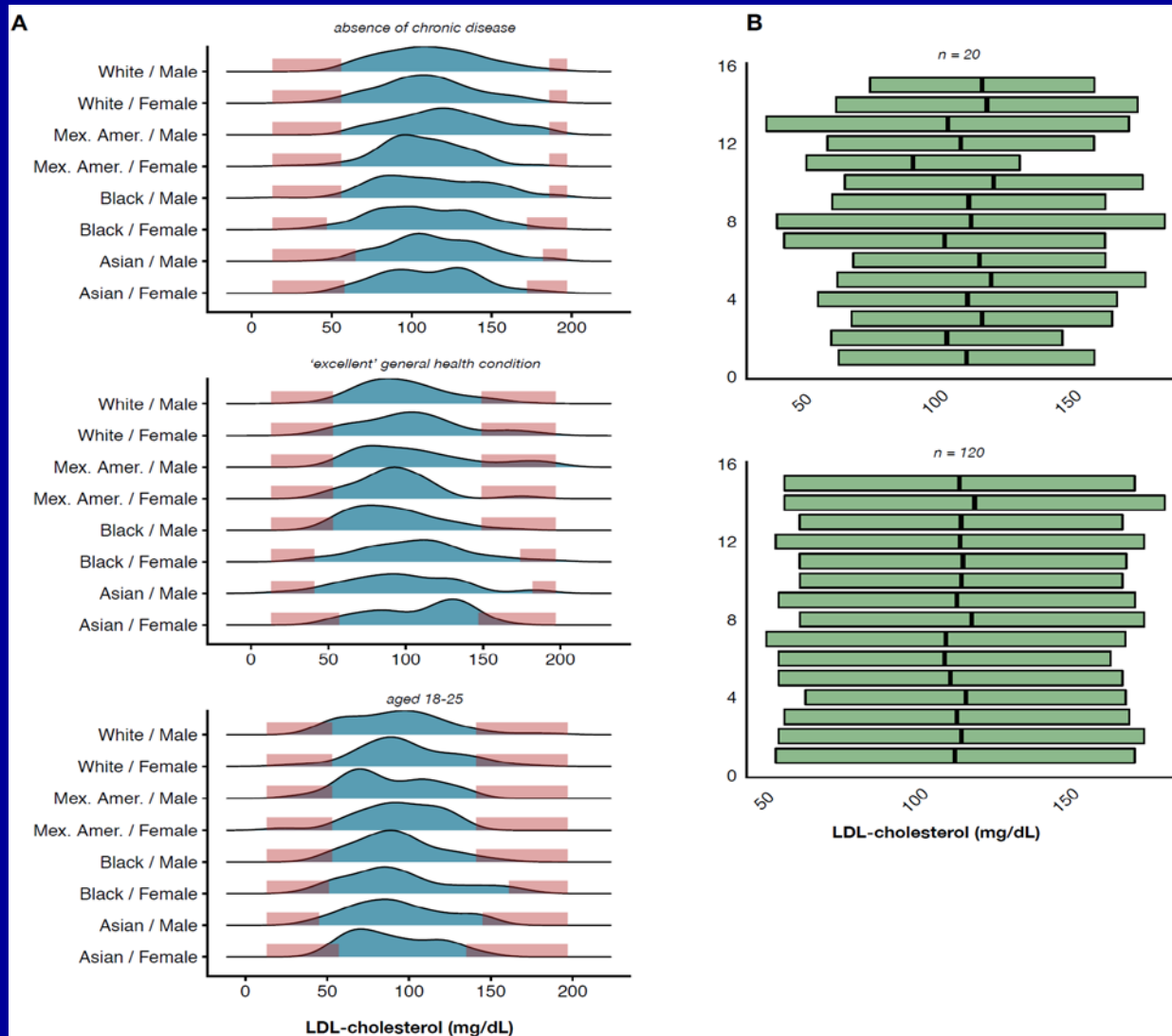
Figure: Sensitivity of reference ranges to the method of defining 'normal' individuals and multiplicity.
A Distributions of LDL-cholesterol (mg/dL) computed using the CDC NHANES 2013-2014 cohort based on individuals without a set of chronic diseases (top panel), rated in 'excellent' general health condition (middle panel), and aged 18-25 (lower panel). The highlighted portion in red indicates the outer 2.5% of each tail of the distribution. **B** Repeated samplings from only the White / Female population without a set of chronic diseases with n = 20 samples (top panel) and n = 120 samples (bottom panel). While the choices of n = 120 individuals (the current CLSI guideline) provides greater stability, considerable variability exists for both n = 20 and n = 120 even when sampling the same subpopulation.

Prevalence and outcomes of incidental imaging findings: umbrella review

Jack W O'Sullivan,[1] Tim Muntinga,[1] Sam Grigg,[2] John P A Ioannidis[3,4,5,6,7]

ABSTRACT

OBJECTIVE
To provide an overview of the evidence on prevalence and outcomes of incidental imaging findings.

DESIGN
Umbrella review of systematic reviews.

DATA SOURCES
Searches of MEDLINE, EMBASE up to August 2017; screening of references in included papers.

ELIGIBILITY CRITERIA
Criteria included systematic reviews and meta-analyses of observational studies that gave a prevalence of incidental abnormalities ("incidentalomas"). An incidental imaging finding was defined as an imaging abnormality in a healthy, asymptomatic patient or an imaging abnormality in a symptomatic patient, where the abnormality was not apparently related to the patient's symptoms. Primary studies that measured the prevalence of incidentalomas in patients with a history of malignancy were also considered in sensitivity analyses.

RESULTS
20 systematic reviews (240 primary studies) were identified from 7098 references from the database search. Fifteen systematic reviews provided data to quantify the prevalence of incidentalomas, whereas 18 provided data to quantify the outcomes of incidentalomas (13 provided both). The prevalence of incidentalomas varied substantially between imaging tests; it was less than 5% for chest computed tomography for incidental pulmonary embolism in patients with and without cancer and whole body positron emission tomography (PET) or PET/computed tomography (for patients with and without cancer). Conversely, incidentalomas occurred in more than a third of images in cardiac magnetic resonance imaging (MRI), chest computed tomography (for incidentalomas of thorax, abdomen, spine, or heart), and computed tomography colonoscopy (for extra-colonic incidentalomas). Intermediate rates occurred with MRI of the spine (22%) and brain (22%). The rate of malignancy in incidentalomas varied substantially between organs; the prevalence of malignancy was less than 5% in incidentalomas of the brain, parotid, and adrenal gland. Extra-colonic, prostatic, and colonic incidentalomas were malignant between 10% and 20% of the time, whereas renal, thyroid, and ovarian incidentalomas were malignant around a quarter of the time. Breast incidentalomas had the highest percentage of malignancy (42%, 95% confidence interval 31% to 54%). Many assessments had high between-study heterogeneity (15 of 20 meta-analyses with I^2 >50%).

CONCLUSIONS
There is large variability across different imaging techniques both in the prevalence of incidentalomas and in the prevalence of malignancy for specific organs. This umbrella review will aid clinicians and patients weigh up the pros and cons of requesting imaging scans and will help with management decisions after an incidentaloma diagnosis. Our results can underpin the creation of guidelines to assist these decisions.

SYSTEMATIC REVIEW REGISTRATION
PROSPERO: CRD42017075679.

Introduction

Incidentalomas, incidental imaging findings serendipitously diagnosed in an asymptomatic patient or symptomatic patient undergoing imaging for an unrelated reason,[1-3] are fast becoming a modern medical crisis.[4] The rapid rise in demand for imaging,[5 6] coupled with rapidly advancing image resolution is

Precision based on prediction: too much of a good thing?

Prediction models for cardiovascular disease risk in the general population: systematic review

Johanna A A G Damen,[1,2] Lotty Hooft,[1,2] Ewoud Schuit,[1,2,3] Thomas P A Debray,[1,2] Gary S Collins,[4] Ioanna Tzoulaki,[5] Camille M Lassale,[5] George C M Siontis,[6] Virginia Chiocchia,[4,7] Corran Roberts,[4] Michael Maia Schlüssel,[4] Stephen Gerry,[4] James A Black,[8] Pauline Heus,[1,2] Yvonne T van der Schouw,[1] Linda M Peelen,[1] Karel G M Moons[1,2]
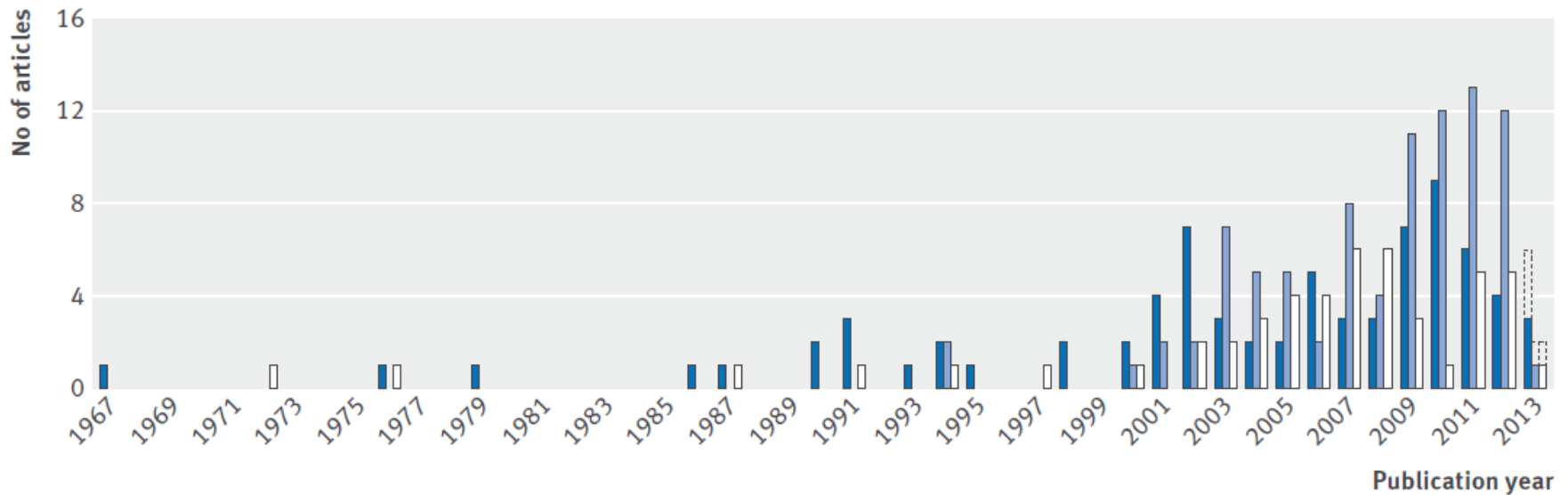
Fig 2 | Numbers of articles in which only one or more models were developed (dark blue), only one or more models were externally validated (light blue), or one or more models were developed and externally validated (white), ordered by publication year (up to June 2013). Predictions of the total numbers in 2013 are displayed with dotted lines

Transparency versus complexity in predictive modeling

Dzok and Ioannidis, Trends in Neuroscience 2019
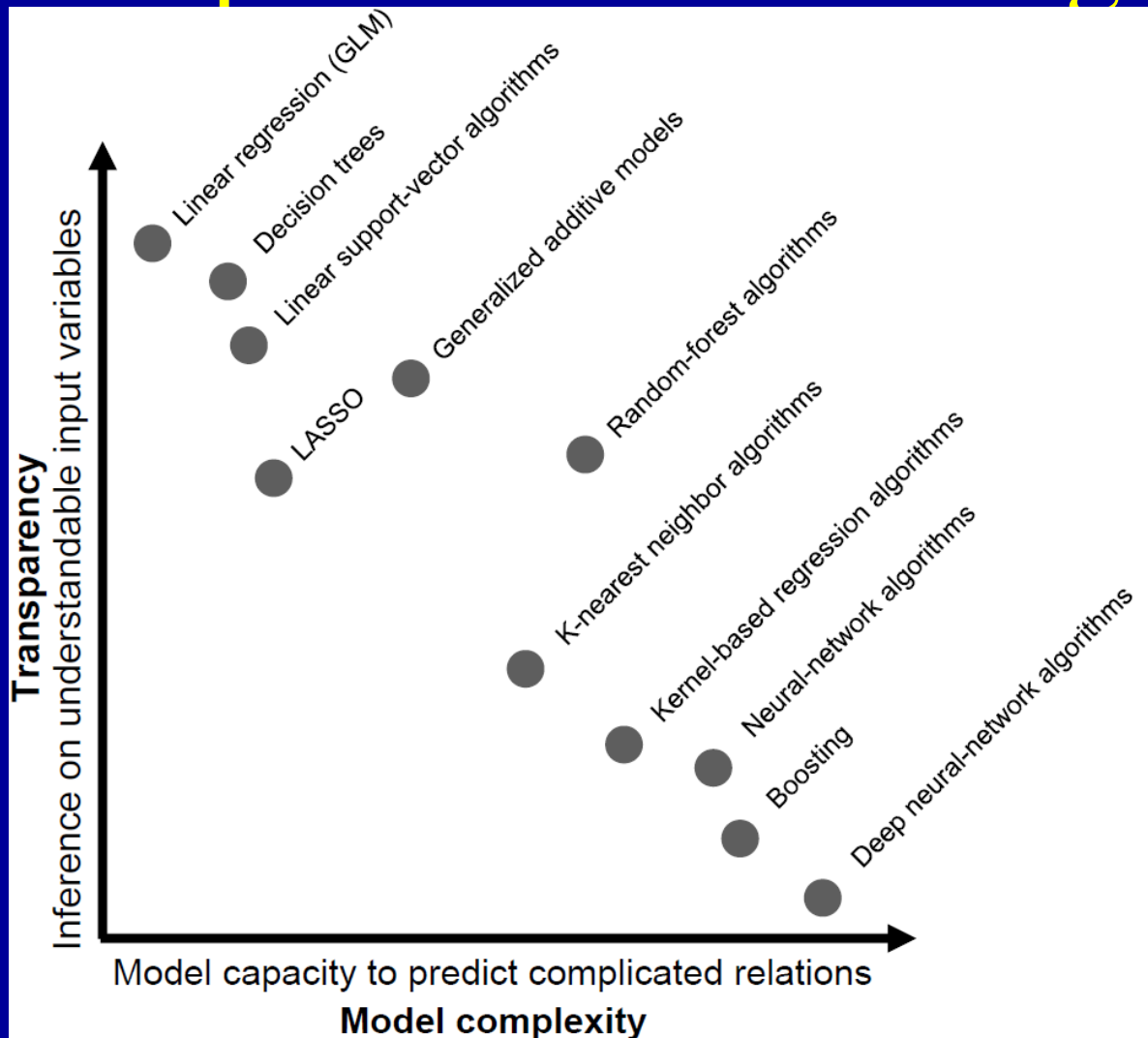
81 EHR-based predictive models

Goldstein et al., JAMIA 2016
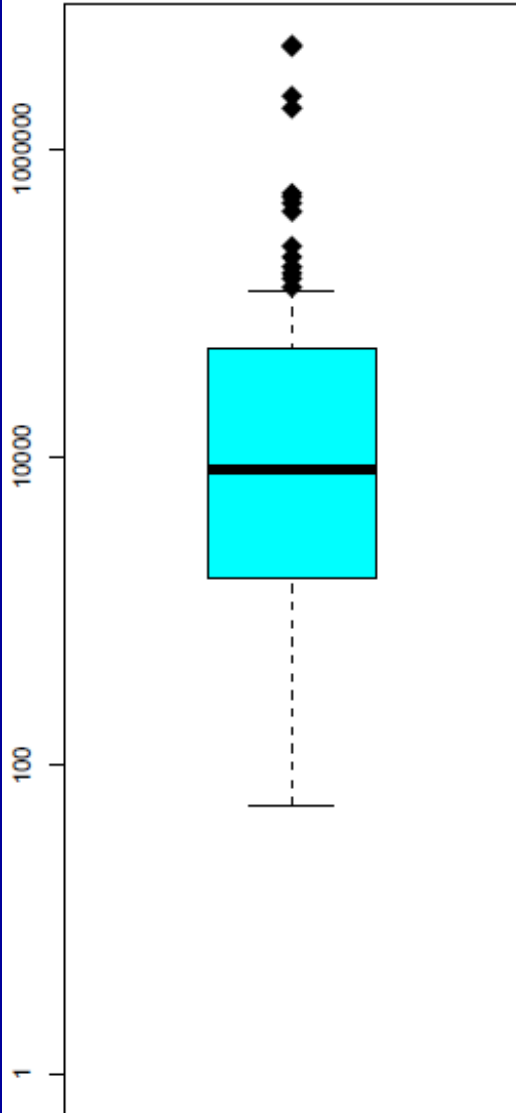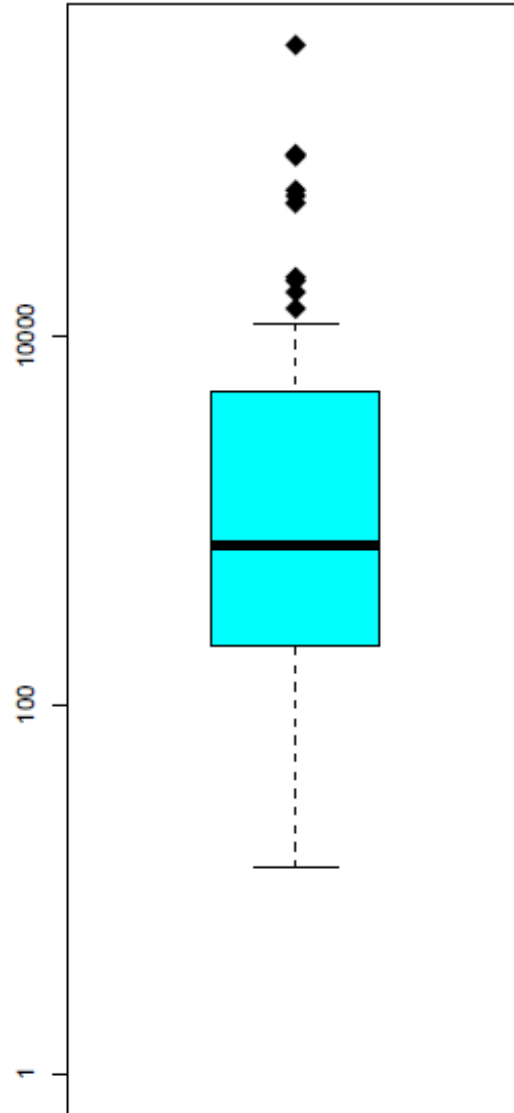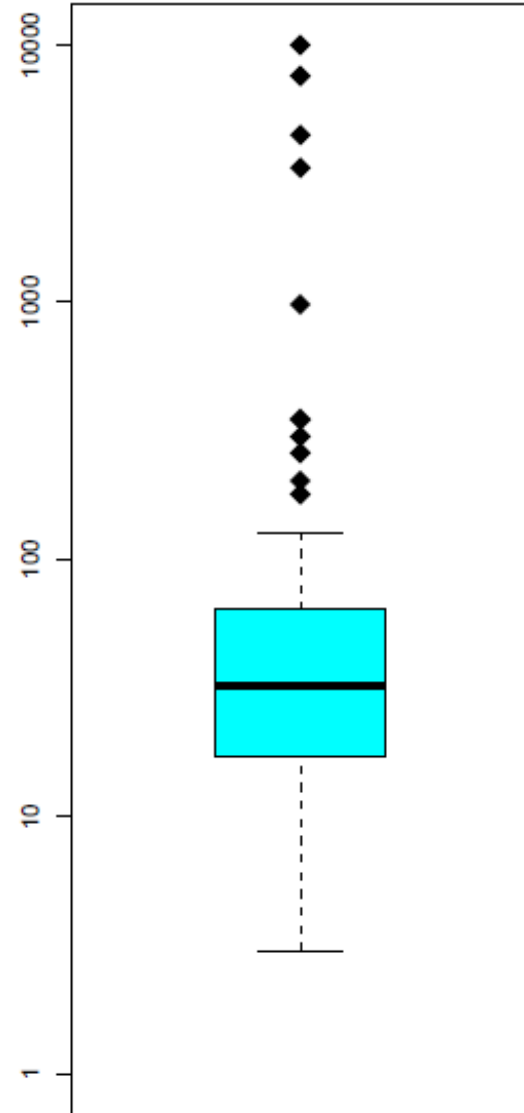
Stratified medicine: Month of birth and benefit from endarterectomy

Month of birth	Events/patients		ARR (%)	95% CI
	Surgical	Medical		
May–Jun	6/83	18/47	33·4	18·2 to 48·6
Jul–Aug	8/84	16/58	20·7	7·0 to 34·4
Sept–Oct	10/87	7/34	9·6	−6·2 to 25·3
Nov–Dec	6/56	9/39	11·2	−5·2 to 27·6
Jan–Feb	9/73	6/43	0·1	−13·1 to 13·2
Mar–Apr	12/64	6/53	−7·7	−20·8 to 5·3
Total	51/447	62/274	11·6	5·6 to 17·6

Heterogeneity: p<0·0001

% absolute risk reduction (95% CI)

Figure 3: Effect of carotid endarterectomy in patients with ⩾70% symptomatic stenosis in ECST[126] according to month of birth in six 2 month periods

Sex based subgroup differences in randomized controlled trials: empirical evidence from Cochrane meta-analyses

Joshua D Wallach,[1] Patrick G Sullivan,[1] John F Trepanowski,[2] Ewout W Steyerberg,[3] John P A Ioannidis[4]

ABSTRACT

OBJECTIVE
To evaluate the frequency, validity, and relevance of statistically significant (P<0.05) sex-treatment interactions in randomized controlled trials in Cochrane meta-analyses.

DESIGN
Meta-epidemiological study.

DATA SOURCES
Cochrane Database of Systematic Reviews (CDSR) and PubMed.

ELIGIBILITY CRITERIA FOR STUDY SELECTION
Reviews published in the CDSR with sex-treatment subgroup analyses in the forest plots, using data from randomized controlled trials.

DATA EXTRACTION
Information on the study design and sex subgroup data were extracted from reviews and forest plots that met inclusion criteria. For each statistically significant sex-treatment interaction, the potential for biological plausibility and clinical significance was considered.

RESULTS
Among the 41 reviews with relevant data, there were 109 separate treatment-outcome analyses ("topics"). Among the 109 topics, eight (7%) had a statistically significant sex-treatment interaction. The 109 topics included 311 randomized controlled trials (162 with both sexes, 46 with males only, 103 with females only). Of the 162 individual randomized controlled trials that included both sexes, 15 (9%) had a statistically significant sex-treatment interaction. Of four topics where the first published randomized controlled trial had a statistically significant sex-treatment interaction, no meta-analyses that included other randomized controlled trials retained the statistical significance and no meta-analyses showed statistical significance when data from the first published randomized controlled trial were excluded.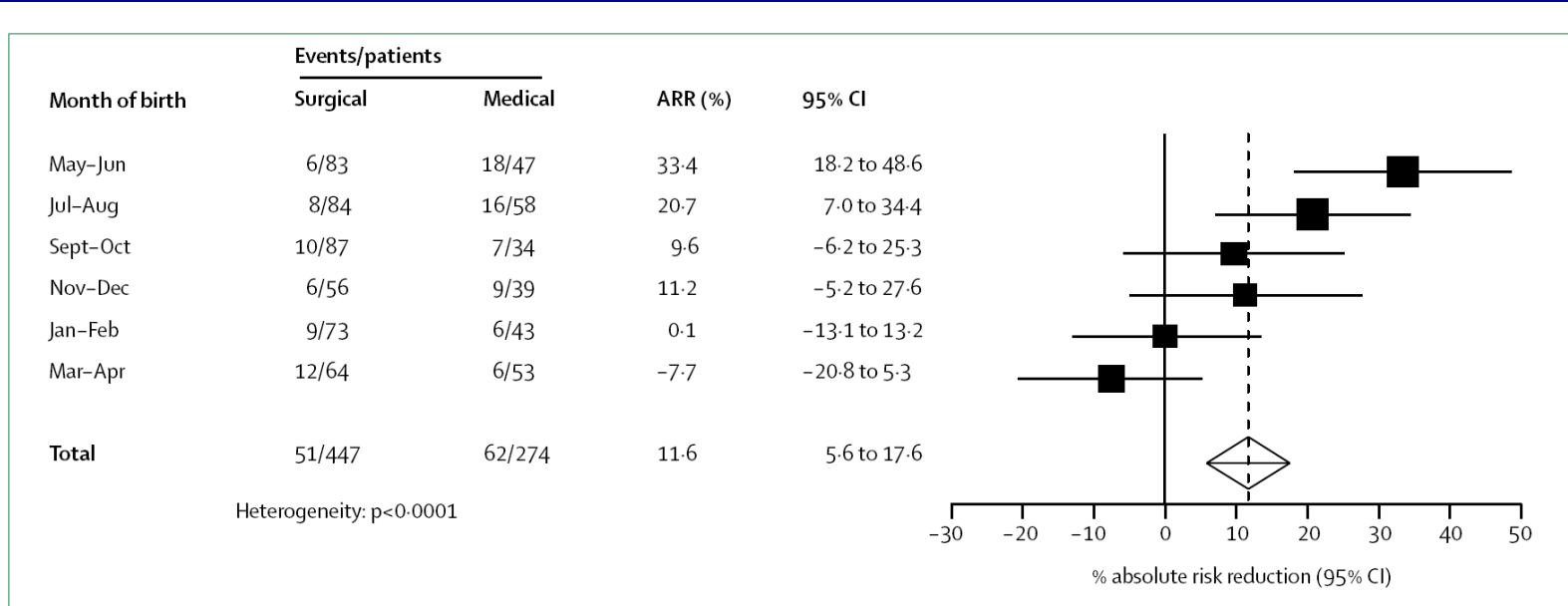 Of the eight statistically significant sex-treatment interactions from the overall analyses, only three were discussed by the CDSR reviewers for a potential impact on different clinical management for males compared with females. None of these topics had a sex-treatment interaction that influenced treatment recommendations in recent guidelines. UpToDate, an online physician-authored clinical decision support resource, suggested differential management of men and women for one of these sex-treatment interactions.

CONCLUSION
Statistically significant sex-treatment interactions are only slightly more frequent than what would be expected by chance and there is little evidence of subsequent corroboration or clinical relevance of sex-treatment interactions.

Introduction

Subgroup analyses in randomized controlled trials are commonly used to determine whether treatment effects vary across certain patient characteristics, such as whether an effect is different between males and females.[1-5] It has been proposed that results from these analyses can be used to tailor patient care ("stratified medicine" and "precision medicine").[6-8] In particular, male and female subgroups are often compared for their responses to a broad range of interventions owing to differences that might exist between the sexes in physiology, pharmacokinetics, and pharmacodynamics.[9-12] For example, it is speculated that women might respond differently from men to some drugs and might have more adverse events in response to certain drugs.[9-15] Though there is substantial interest about sex

Subgroup differences in large-scale MIPDs: few and with few support

Schuit et al, Int J Epidemiol 2018
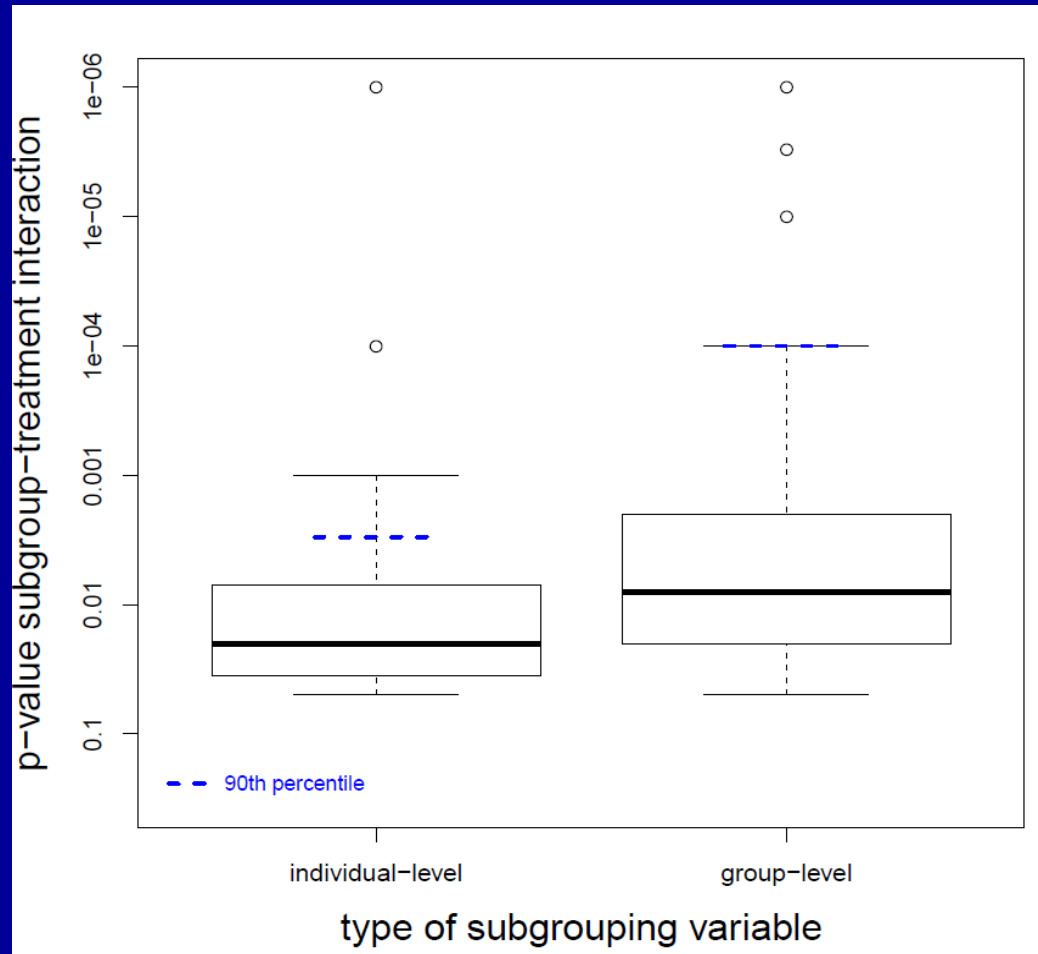
Treatment effect modifications for individual and group level subgrouping variables: typically small

Accelerated approvals

Idiopathic thrombocytopenic purpura:
Eltrombopag

Multiple sclerosis:
Natalizumab

Pulmonary arterial hypertension:
Treprostinil

Cancer:
Alemtuzumab
Bortezomib
Brentuximab
Carfilzomib
Cetuximab
Clofarabine
Crizotinib
Gefitinib
Gemtuzumab
Ibritumomab
Ibrutinib
Imatinib
Lenalidomide
Nelarabine
Nilotinib
Ofatumumab
Omecatexine
Oxaliplatin
Panitumumab
Pomalidomide
Ponatinib
Pralatrexate

Iron overload due to blood transfusions:
Deferasirox
Deferiprone

Fabry disease:
Agalsidase

HIV:
Darunavir
Enfuvirtide
Etravirine
Lopinavir
Maraviroc
Raltegravir
Tenofovir
Tipranavir

Tuberculosis:
Bedaquiline

Accelerated approvals 2000-2013, from Naci et al. Milbank Q 2017

Approximately 10% of approvals based on non-RCT data

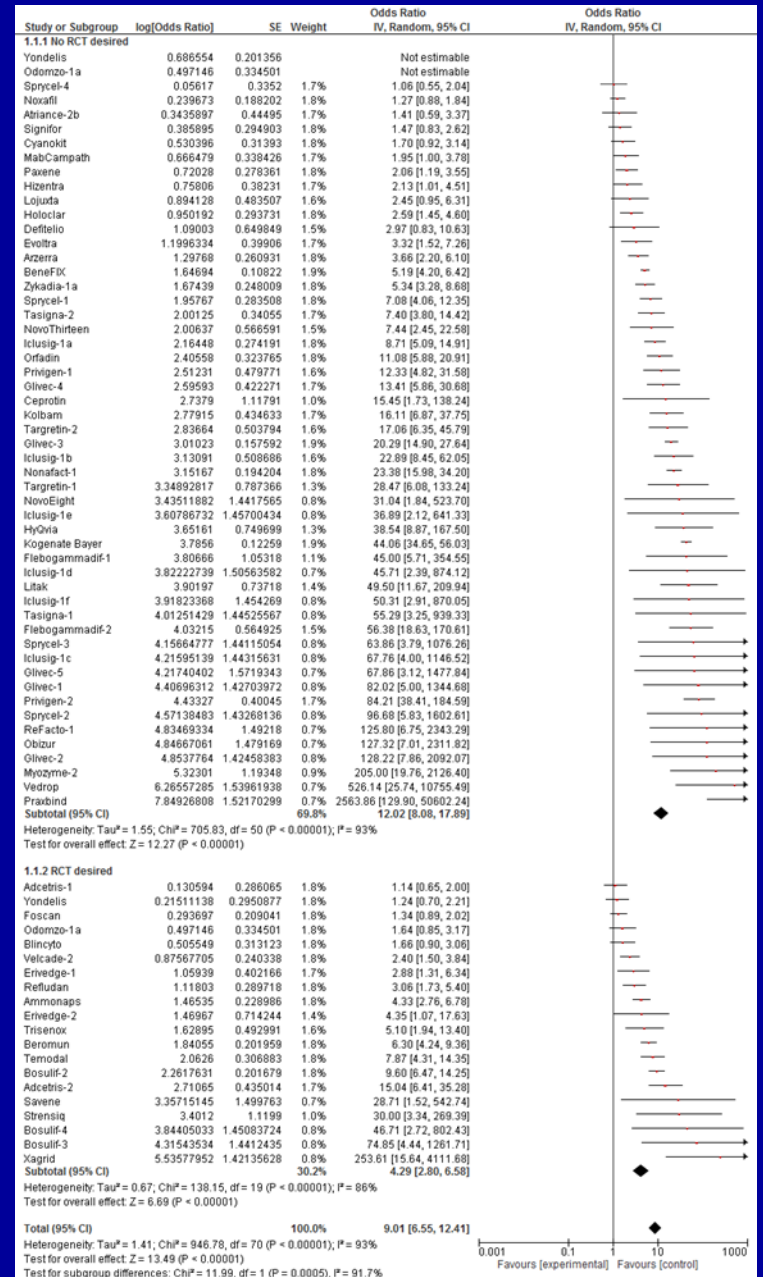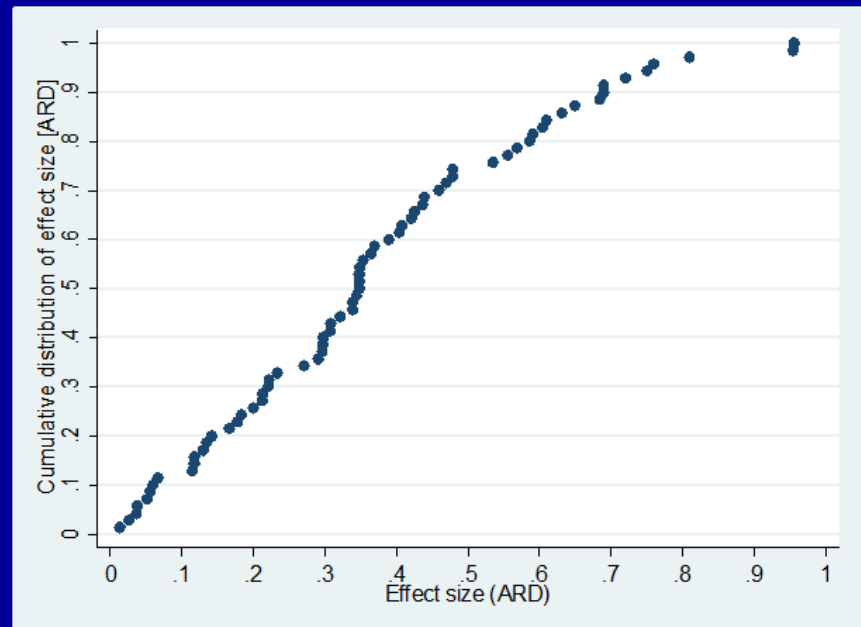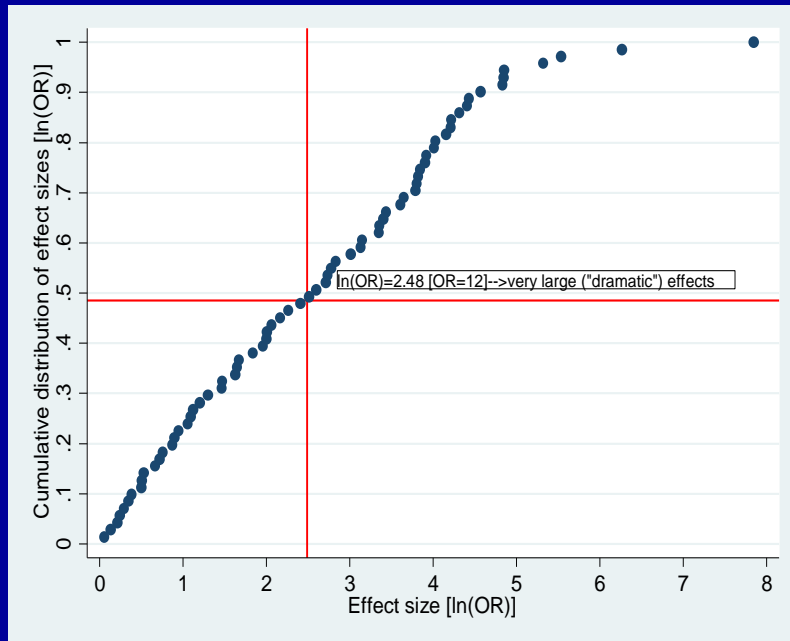| Study or Subgroup	log[Odds Ratio]	SE	Weight	Odds Ratio IV, Random, 95% CI
1.1.1 No RCT desired				
Yondelis	0.686554	0.201356		Not estimable
Odomzo-1a	0.497146	0.334501		Not estimable
Sprycel-4	0.05617	0.3352	1.7%	1.06 [0.55, 2.04]
Noxafil	0.239673	0.188202	1.8%	1.27 [0.88, 1.84]
Atriance-2b	0.3435897	0.44495	1.7%	1.41 [0.59, 3.37]
Signifor	0.385895	0.294903	1.8%	1.47 [0.83, 2.62]
Cyanokit	0.530396	0.31393	1.8%	1.70 [0.92, 3.14]
MabCampath	0.666479	0.338426	1.7%	1.95 [1.00, 3.78]
Paxene	0.72028	0.278361	1.8%	2.06 [1.19, 3.55]
Hizentra	0.75806	0.38231	1.7%	2.13 [1.01, 4.51]
Lojuxta	0.894128	0.483507	1.6%	2.45 [0.95, 6.31]
Holoclar	0.950192	0.293731	1.8%	2.59 [1.45, 4.60]
Defitelio	1.09003	0.649849	1.5%	2.97 [0.83, 10.63]
Evoltra	1.1996334	0.39906	1.7%	3.32 [1.52, 7.26]
Arzerra	1.29768	0.260931	1.8%	3.66 [2.20, 6.10]
BeneFIX	1.64694	0.10822	1.9%	5.19 [4.20, 6.42]
Zykadia-1a	1.67439	0.248009	1.8%	5.34 [3.28, 8.68]
Sprycel-1	1.95767	0.283508	1.8%	7.08 [4.06, 12.35]
Tasigna-2	2.00125	0.34055	1.7%	7.40 [3.80, 14.42]
NovoThirteen	2.00637	0.566591	1.5%	7.44 [2.45, 22.58]
Iclusig-1a	2.16448	0.274191	1.8%	8.71 [5.09, 14.91]
Orfadin	2.40558	0.323765	1.8%	11.08 [5.88, 20.91]
Privigen-1	2.51231	0.479771	1.6%	12.33 [4.82, 31.58]
Glivec-4	2.59593	0.422271	1.7%	13.41 [5.86, 30.68]
Ceprotin	2.7379	1.11791	1.0%	15.45 [1.73, 138.24]
Kolbam	2.77915	0.434633	1.7%	16.11 [6.87, 37.75]
Targretin-2	2.83664	0.503794	1.6%	17.06 [6.35, 45.79]
Glivec-3	3.01023	0.157592	1.9%	20.29 [14.90, 27.64]
Iclusig-1b	3.13091	0.508686	1.6%	22.89 [8.45, 62.05]
Nonafact-1	3.15167	0.194204	1.8%	23.38 [15.98, 34.20]
Targretin-1	3.34892817	0.787366	1.3%	28.47 [6.08, 133.24]
NovoEight	3.43511882	1.4417565	0.8%	31.04 [1.84, 523.70]
Iclusig-1e	3.60786732	1.45700434	0.8%	36.89 [2.12, 641.33]
HyQvia	3.65161	0.749699	1.3%	38.54 [8.87, 167.50]
Kogenate Bayer	3.7856	0.12259	1.9%	44.06 [34.65, 56.03]
Flebogammadif-1	3.80666	1.05318	1.1%	45.00 [5.71, 354.55]
Iclusig-1d	3.82222739	1.50563582	0.7%	45.71 [2.39, 874.12]
Litak	3.90197	0.73718	1.4%	49.50 [11.67, 209.94]
Iclusig-1f	3.91823368	1.454269	0.8%	50.31 [2.91, 870.05]
Tasigna-1	4.01251429	1.44525567	0.8%	55.29 [3.25, 939.33]
Flebogammadif-2	4.03215	0.564925	1.5%	56.38 [18.63, 170.61]
Sprycel-3	4.15664777	1.44115054	0.8%	63.86 [3.79, 1076.26]
Iclusig-1c	4.21595139	1.44315631	0.8%	67.76 [4.00, 1146.52]
Glivec-5	4.21740402	1.5719343	0.7%	67.86 [3.12, 1477.84]
Glivec-1	4.40696312	1.42703972	0.8%	82.02 [5.00, 1344.68]
Privigen-2	4.43327	0.40045	1.7%	84.21 [38.41, 184.59]
Sprycel-2	4.57138483	1.43268136	0.8%	96.68 [5.83, 1602.61]
ReFacto-1	4.83469334	1.49218	0.7%	125.80 [6.75, 2343.29]
Obizur	4.84667061	1.479169	0.7%	127.32 [7.01, 2311.82]
Glivec-2	4.8537764	1.42458383	0.8%	128.22 [7.86, 2092.07]
Myozyme-2	5.32301	1.19348	0.9%	205.00 [19.76, 2126.40]
Vedrop	6.26557285	1.53961938	0.7%	526.14 [25.74, 10755.49]
Praxbind	7.84926808	1.52170299	0.7%	2563.86 [129.90, 50602.24]
Subtotal (95% CI)			69.8%	12.02 [8.08, 17.89]
Heterogeneity: Tau² = 1.55; Chi² = 705.83, df = 50 (P < 0.00001); I² = 93%				
Test for overall effect: Z = 12.27 (P < 0.00001)				
1.1.2 RCT desired				
Adcetris-1	0.130594	0.286065	1.8%	1.14 [0.65, 2.00]
Yondelis	0.21511138	0.2950877	1.8%	1.24 [0.70, 2.21]
Foscan	0.293697	0.209041	1.8%	1.34 [0.89, 2.02]
Odomzo-1a	0.497146	0.334501	1.8%	1.64 [0.85, 3.17]
Blincyto	0.505549	0.313123	1.8%	1.66 [0.90, 3.06]
Velcade-2	0.87567705	0.240338	1.8%	2.40 [1.50, 3.84]
Erivedge-1	1.05939	0.402166	1.7%	2.88 [1.31, 6.34]
Refludan	1.11803	0.289718	1.8%	3.06 [1.73, 5.40]
Ammonaps	1.46535	0.228986	1.8%	4.33 [2.76, 6.78]
Erivedge-2	1.46967	0.714244	1.4%	4.35 [1.07, 17.63]
Trisenox	1.62895	0.492991	1.6%	5.10 [1.94, 13.40]
Beromun	1.84055	0.201959	1.8%	6.30 [4.24, 9.36]
Temodal	2.0626	0.306883	1.8%	7.87 [4.31, 14.35]
Bosulif-2	2.2617631	0.201679	1.8%	9.60 [6.47, 14.25]
Adcetris-2	2.71065	0.435014	1.7%	15.04 [6.41, 35.28]
Savene	3.35715145	1.499763	0.7%	28.71 [1.52, 542.74]
Strensiq	3.4012	1.1199	1.0%	30.00 [3.34, 269.39]
Bosulif-4	3.84405033	1.45083724	0.8%	46.71 [2.72, 802.43]
Bosulif-3	4.31543534	1.4412435	0.8%	74.85 [4.44, 1261.71]
Xagrid	5.53577952	1.42135628	0.8%	253.61 [15.64, 4111.68]
Subtotal (95% CI)			30.2%	4.29 [2.80, 6.58]
Heterogeneity: Tau² = 0.67; Chi² = 138.15, df = 19 (P < 0.00001); I² = 86%				
Test for overall effect: Z = 6.69 (P < 0.00001)				
Total (95% CI)			100.0%	9.01 [6.55, 12.41]
Heterogeneity: Tau² = 1.41; Chi² = 946.78, df = 70 (P < 0.00001); I² = 93%				
Test for overall effect: Z = 13.49 (P < 0.00001)				
Test for subgroup differences: Chi² = 11.99, df = 1 (P = 0.0005), I² = 91.7%				

Odds Ratio IV, Random, 95% CI

0.001 0.1 1 10 1000
Favours [experimental] Favours [control]

What are dramatic enough effects to warrant licensing without randomized data?

Djulbegovic, et al, J Clin Epidemiology, 2018

Biomarker-driven precision trial designs

	Enrichment	Randomize-all	Adaptive design	Umbrella	Basket
Histology	dependent	dependent	dependent	dependent	independent
Number of targeted therapies	1	1	≥ 1	> 1	≥ 1
Number of biomarkers	1	1	≥ 1	> 1	≥ 1
Type of biomarkers	Bm+	Bm+ and Bm−	Bm+ and Bm−	Bm+ if exploratory Bm+ and Bm− if confirmatory	Usually Bm+
Biomarker credentials (a priori knowledge)	very strong	+/−	+/−	strong	very strong
Biomarker assay	single, locally	single, locally	single, locally	multiplex, centralized	single, locally
Provides information on the Biomarker-treatment benefit association (is the biomarker predictive?)	−	+/−	+/−	+	−
Number of patients required to screen	Prevalence-dependent	Prevalence-dependent	Prevalence-dependent	Prevalence-dependent	Prevalence-dependent
Sufficiently large sample size (depends on the rarity of the mutation)*	+	+ +	+/−	+ + +	+ +
Overlap of patients	−	+/−	+/−	+	−
Statistical complexity	+	+	+ + +	+ +	+ +
Tradeoff between power versus sample size	−	+	+ +	+ + +	+ + +
Subgroup analyses – multiplicity	−	−	+ + +	++	+ + +
Type 1 error problems	+	+	+ +	+ +	+ + +
Flexibility†	−	−	+ + +	+	+
Time efficiency and cost savings	− −	− −	+	+ +	+ +

Umbrella and basket trials in oncology

- As of July 2018, in ClinicalTrials.gov:
- 30 umbrella and 27 basket trials registered
- Only 2 and 9 of them respectively are randomized
- This includes 3 trials with adaptive randomization
- Five of them published

Janiaud, Serghiou, Ioannidis, Cancer Treatment Reviews 2019

OXFORD

INVITED REVIEW

Evidence-based medicine and big genomic data

John P.A. Ioannidis[1],* and Muin J. Khoury[2]

Table 1. An evidence framework for genetic testing: seven steps and some challenges for each step with the advent of BGD

Step	Challenges
Define genetic test scenarios on the basis of the clinical setting, the purpose of the test, the population, the outcomes of interest and comparable alternative methods	Unfortunately most research to-date has not used this approach of starting from the clinical problem that needs to be solved, thus there is little evidence on comparable alternative methods that fits to this framework
For each genetic test scenario, conduct an initial structured assessment to determine whether the test should be covered, denied, or subject to additional evaluation	It is unclear what exactly this initial structured assessment would entail, if it not a full systematic review. While method for rapid reviews and scoping reviews do exist or get developed, it is unclear how well they would work in the case of big genomics. The proposed step seems like an effort to quickly get rid of a large number of tests in an environment where there would be a difficult to handle mass of big data, but it is unclear if cutting corners will help or make things worse
Conduct or support evidence-based systematic reviews for genetic test scenarios that require additional evaluation	Welcome emphasis on systematic review approach. Systematic reviews however, have major problems when conducted retrospectively with fragmented data subject to publication biases. Given the strong tradition of genetics in data sharing, there is an opportunity to promote a model of large-scale international collaboration with prospective, ongoing, continuously updated reviews, as new data accumulate
Conduct or support a structured process to produce clinical guidance for a genetic test scenario	This clause anticipates that there are many contextual issues that go beyond the strict evidence review, for example social issues, net benefits and harms, and aggregate costs
	Some of these may be difficult to define, they may be setting-dependent, and they may carry substantial subjectivity. There is extensive evidence about how to produce guidelines and also about caveats in the process. Given the massive information, generating and updating guidelines for BGD will be a major challenge
Publicly share resulting decisions and justification about evaluated genetic test scenarios, and retain decisions in a repository	A repository is useful to the extent that it can be comprehensive, systematic and also allow user-friendly navigation so that one can readily find the most appropriate guidance. Experience with traditional guidelines repositories exist (e.g. the Guidelines Clearinghouse), but it is unclear if the same concept would work with the massive and rapidly evolving BGD
Implement timely review and revision of decisions on the basis of new data	As above, this would have the best chances of success, if evidence is incorporated in real-time based on some international collaboration and sharing with accumulation of all relevant data. Still, reviewing and revising all decisions will require enormous resources and it is questionable whether the process can be automated and objective or will continue to require subjective calls
Identify evidence gaps to be addressed by research	This is a traditional major role of systematic reviews. More reliable and up-to-date systematic reviews would have the best chance to do this task well

Table 2. Randomized controlled trials on providing multigenetic score information versus control without such genetic score information

Author (ref)	N	SNPs	Phenotype	Main results
Godino (35)	580	23	Type 2 DM	No effect on physical activity, self-reported diet, self-reported weight, worry and anxiety
Grant (36)	116	36	Type 2 DM	No effect on self-reported motivation, program attendance or mean weight loss
Kullo (37)	216	28	CHD	Improved LDL-C, more statin uptake, no effects on diet or physical activity
Knowles (38)	94	19	CHD	No effects on LDL-C, other CHD risk factors, weight loss, diet, physical activity, risk perceptions, and psychological outcomes

UK minister blunder

- In a [promotional video] released online, Hancock confided he had a higher than average risk for prostate cancer. "My risk by age 75 is almost 15%," he worried. The minister immediately booked an appointment with a doctor to get a further blood test.

- "The truth is, genomics might have saved my life," he said.

The systematic review and meta-analysis epidemic

Ioannidis, Milbank Q 2016
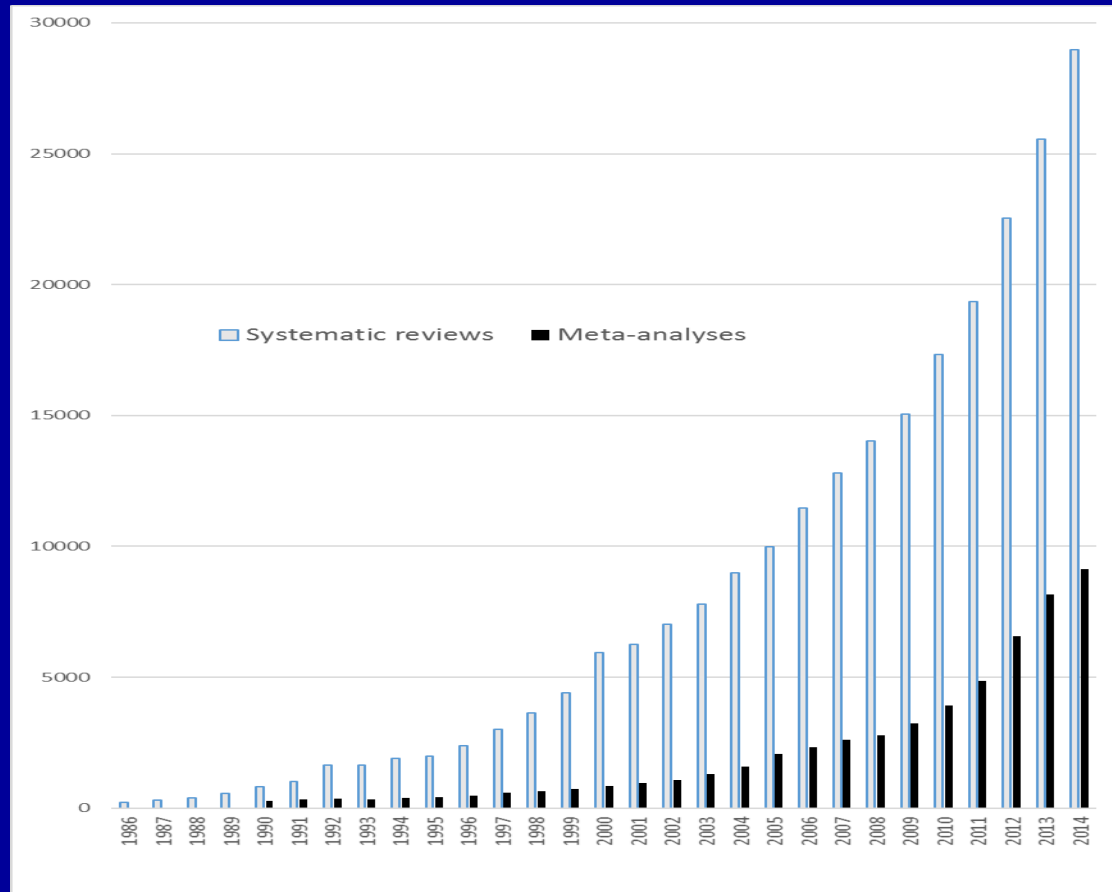
Is useful? Is it precise?

- Systematic reviews and meta-analyses have become the most powerful, influential tool of EBM

- Therefore they have been hijacked to serve various agendas

- Most systematic reviews and meta-analyses are not useful

- Hardly any of them can lead to precision medicine

Genetic meta-analyses from China
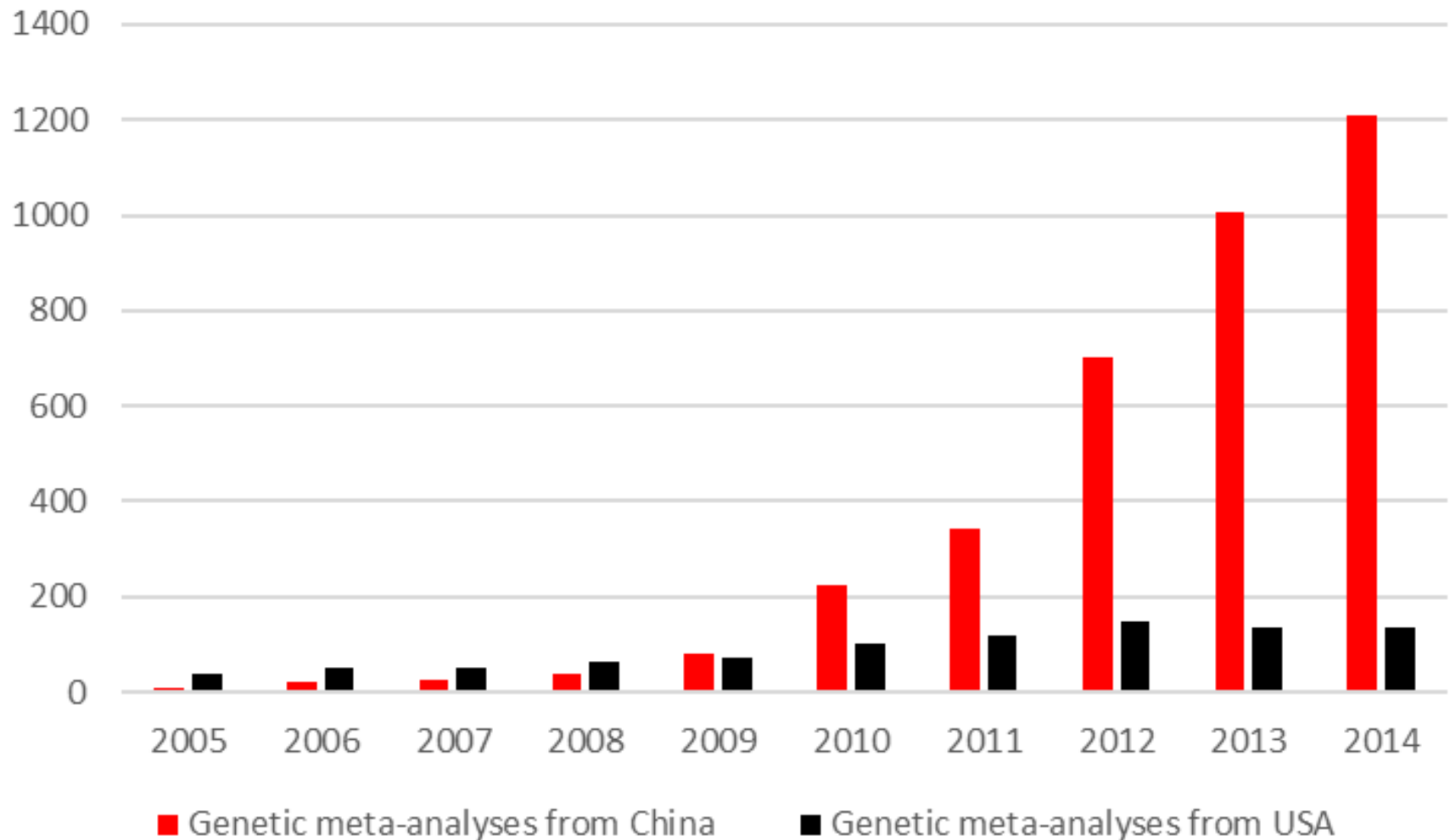
The meta-pie
(see Ioannidis, Milbank Quarterly 2016)

VIEWPOINT

Inconsistent Guideline Recommendations for Cardiovascular Prevention and the Debate About Zeroing in on and Zeroing LDL-C Levels With PCSK9 Inhibitors

John P. A. Ioannidis, MD, DSc
Stanford Prevention Research Center, Department of Medicine and Departments of Health Research and Policy, Biomedical Data Science, and Statistics, and Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California.

Evidence for the benefits of cardiovascular prevention, with lifestyle changes or with medications,[1] is strong. However, recently released guidelines[2-4] from the United States, Europe, and Canada have differing recommendations regarding which patients to treat with medications and whether to tailor treatment aiming for specific targets. Low-density lipoprotein cholesterol (LDL-C) levels are the focal point in this debate.

These guidelines vary on their proposed risk thresholds for treatment and on whether they single out lipid levels as a key factor to guide initiation and desirable targets of therapy. The US Preventive Services Task Force (USPSTF) guidelines[2] recommend treatment in the presence of 1 major risk factor and a greater than 10% 10-year risk of cardiovascular events (grade B; ie, offer or provide this service). The recommendations for treating patients at a 7.5% to 10% 10-year risk are more selective (grade C; ie, offer or provide this service for selected patients depending on individual circumstances), and LDL-C levels are not assigned a special role.[2] The European guidelines[3] use SCORE to calculate the 10-year risk of cardiovascular death (not just any events) and offer different treatment recommendations for different LDL-C levels. The guidelines aim for lowering LDL-C levels to below 100 mg/dL in high-risk patients and for a greater than 50% reduction in LDL-C regardless of risk. The Canadian guidelines[4] use LDL-C (or non–high-density lipoprotein cholesterol or apolipoprotein B) as targets, aiming for a greater than 50% reduction in LDL-C levels. The guidelines recommend offering treatment to all patients with a 10-year risk of cardiovascular events exceeding 20% and to several groups of patients in the 10% to 19% risk window, guided by lipid levels and other risk factors.

According to one report, following the USPSTF guidelines, compared with following the 2013 American College of Cardiology/American Heart Association (ACC/AHA) guidelines,[5] would lead to treatment of an estimated 10 million fewer individuals in the primary prevention population in the United States.[6] However, the USPSTF guidelines would recommend statins for more people than the European guidelines and even more people than the Canadian guidelines.

This diversity in recommendations probably reflects remaining gaps in the available evidence. The evidence report accompanying the USPSTF guidelines[1] summarized 19 randomized trials that evaluated the effects of statins vs placebo or no statins among more than 70 000 adults, and more than 2300 deaths were recorded during follow-up. Most of these studies involved exclusively primary prevention populations and demonstrated that use of low- or moderate-dose statin therapy was associated with an approximately 30% relative risk reduction in cardiovascular events and in cardiovascular deaths and a 10% to 15% relative risk reduction in all-cause mortality. Results were consistent across all subgroups evaluated based on different risk factors, including age, sex, race/ethnicity, and lipid levels. These findings would argue against singling out LDL-C from other risk factors. Evidence for benefits of cholesterol lowering in secondary prevention (ie, for patients with a history of cardiovascular disease, such as symptomatic coronary artery disease or ischemic stroke) is also very strong.

However, the follow-up of patients in randomized trials of statins has been 5 years or less, whereas for many patients, cholesterol-lowering therapy may be life-long. Most events or prevented events thus would occur after the period most trials have included. Moreover, these randomized trials have included almost exclusively patients with 2 or more risk factors, although many patients for whom drug treatment is recommended by some guidelines may not have 2 risk factors. For instance, the ACC/AHA[5] risk calculation leads to treatment as men become older even if they have no risk factors other than their unmodifiable age and sex. To many, it sounds absurd that there is no such thing as healthy aging and that everyone eventually will need some medication. The estimated number needed to treat values for the summary treatment effects of the 19 trials are 250 to prevent 1 death and 72 to prevent 1 composite cardiovascular event. If the relative risk reduction is the same across all risk levels (a picture consistent with the available trial data), the number needed to treat increases proportionally as the risk threshold becomes lower.

The one outcome for which clinical trials have shown large inconsistency (with heterogeneity $I^2 = 86\%$) is withdrawals from treatment—in other words, tolerability of statins has varied substantially across trials.[1] Randomized trials[1] suggest no significant excess of major adverse events with the widely used statins at least during the limited available follow-up. However, skepticism persists about the ability of these trials to capture adverse events reliably. In clinical practice, many clinicians and patients suggest that adverse events are more common than the rates described in trials. The introduction and aggressive marketing of new lipid-lowering drugs that have even less evidence about their safety add more questions about long-term tolerability and adverse events. Moreover, manufacturers of lipid-lowering drugs have sponsored or

Corresponding Author: John P. A. Ioannidis, MD, DSc, Stanford Prevention Research Center, 1265 Welch Rd, Medical School Office Bldg, Room X306, Stanford, CA 94305 (jioannid @stanford.edu).

Getting from individual, precise effects much larger population effects for the same pathway?

A precise poem by Marianne Moore

I learn that we are precisionists
not citizens of Pompeii arrested in action

Fossilized precision after the volcano of information erupted

What Happens When Underperforming Big Ideas in Research Become Entrenched?

Michael J. Joyner, MD
Laboratory of Human
Integrative Physiology
and Department of
Anesthesiology, Mayo
Clinic, Rochester,
Minnesota.

**Nigel Paneth, MD,
MPH**
Departments of
Epidemiology and
Biostatistics and
Pediatrics and Human
Development, College
of Human Medicine,
Michigan State
University,
East Lansing.

**John P. A. Ioannidis,
MD, DSc**
Stanford Prevention
Research Center,
Department of
Medicine and Meta-
Research Innovation
Center at Stanford,
Stanford University,
Stanford, California.

**Corresponding
Author:** John P. A.
Ioannidis, MD, DSc,
Stanford Prevention
Research Center,
Department of
Medicine and
Meta-Research
Innovation Center at
Stanford, Stanford
University, 1265 Welch
Rd, Medical School
Office Bldg, Room
X306, Stanford, CA
94305 (jioannid
@stanford.edu).

For several decades now the biomedical research community has pursued a narrative positing that a combination of ever-deeper knowledge of subcellular biology, especially genetics, coupled with information technology will lead to transformative improvements in health care and human health. In this Viewpoint, we provide evidence for the extraordinary dominance of this narrative in biomedical funding and journal publications; discuss several prominent themes embedded in the narrative to show that this approach has largely failed; and propose a wholesale reevaluation of the way forward in biomedical research.

Primacy of the Narrative

In 2016 approximately $15 billion of the $26 billion of extramural research funding sponsored by the National Institutes of Health (NIH) could be linked to some version of search terms that include gene, genome, stem cells, or regenerative medicine.[1] These topics have also increased geometrically in their representation among published articles. Between 1974 and 2014 the annual number of published articles indexed in PubMed increased by 410% (from 234 613 to 1 196 110), but those identified with genome increased by 2127% (2705 to 60 246). Between 1994 and 2014, the annual number of articles indexed in PubMed increased by 175% (from 435 376 to 1 196 110), but articles identified with gene therapy or stem cell increased by 874% (2635 to 25 662) and 752% (3452 to 29 196). Apparently a large number of scientists either believe in the potential of these topics or feel compelled to work on them, recognizing that these topics constitute a major locus of important science, financial support, recognition, and prospects for a successful career.

Exploring Some Key Examples

In 1999, Collins[2] envisioned a genetic revolution in medicine facilitated by the Human Genome Project and described 6 major themes: (1) common diseases will be explained largely by a few DNA variants with strong associations to disease, (2) this knowledge will lead to improved diagnosis; (3) such knowledge will also drive preventive medicine; (4) pharmacogenomics will improve therapeutic decision making; (5) gene therapy will treat multiple diseases; and (6) a substantial increase in novel targets for drug development and therapy will ensue. These 6 ideas have more recently been branded as personalized or precision medicine.[3] Similarly, there is the increasing interest in and expectation that stem cell therapy—a seventh theme—can treat common diseases.[3]

To avoid the misconception that big ideas are all related to biological sciences, an eighth theme is the emphasis in the narrative on the clinical and research value of converting medical records to electronic formats. As of April 2016, the Centers for Medicare & Medicaid Services had paid $34 billion in financial incentives to service providers for implementing electronic health record (EHR) systems.[4] EHRs are an important aspect of this narrative because they are thought to provide the structural underpinnings of precision medicine. It has been suggested by some that some combination of these 8 big ideas will yield substantial cost savings for health care.

Expectations that a few DNA variants explain most common diseases have faded as the genetic architecture of most diseases has proved to be formidably complex. Apparently, hundreds or even tens of thousands of genetic variants are involved in each common disease. The function of these variants is difficult to decipher. Very few genes have found undisputed roles in preventive efforts or pharmacogenetic testing.

Continued enthusiasm for gene therapy ignores what is known from classic single-gene disorders such as sickle cell anemia. The complex biological processes set in motion by a single amino acid substitution that leads to painful crises, stroke, and other complications are not predictable from the genomic defect, but only by appreciating the complexity of biological systems at the level of tissues and organs. Sixty years after the discovery of the genetic defect, no targeted therapy has emerged for sickle cell anemia.

The complex and adaptive nature of most tumors thwarts the optimistic projections for molecularly targeted therapy for cancer. A randomized trial of targeted therapy based on molecular profiling for advanced cancers from diverse anatomical locations showed no improvement in progression-free survival.[5] The NCI-Molecular Analysis for Therapy Choice (NCI-MATCH) trial links patients with cancer to drugs targeted against their cancer DNA mutations. So far, just 2.5% of screened patients have been assigned to a trial intervention group. Even though this fraction should increase as the number of trial treatment groups is increased, even if effectiveness is demonstrated, the rarity of the targeted mutations means that this approach will help only a minority of patients with cancer.[6]

The prospects of effective treatment based on stem cells have been challenged in comprehensive reviews of the available trials. For instance, in congestive heart failure, improvements in cardiac function have been observed only in industry-sponsored studies, and a positive relationship has been noted between effect size and the number of experimental design flaws.[7] To its credit, the International Society for Stem Cell Research has issued "anti-hype" guidelines that "[h]ighlight the responsibility of all groups communicating stem cell science and medicine—scientists, clinicians, industry, science

Reversing the paradigm

It is unlikely we will be able to fuel precision medicine (or health) on individuals until we can obtain large-scale, coordinated evidence on large populations

Precision medicine for individual patients should use population group averages and larger, not smaller, groups

Benjamin Djulbegovic[1] | John P. A. Ioannidis[2]

Concluding comments

- Most medical evidence is either problematic/spurious/false or has no utility for medical and shared decision making

- Precision medicine (and health) aims to satisfy one of the main pillars of EBM, to deal with individuals

- Precision medicine, by definition, is likely to have minimal impact on life expectancy and other major population outcomes

- Precision medicine is using some interesting designs, some of which are not new and others which are novel, but both types are over-hyped probably as to their potential

- Too much personalized information is not necessarily good for your health and it may even be harmful

- A synergy between large-scale evidence and precision approaches would be useful to tell us what we can learn from each

Special thanks

Daniele Fanelli
Steve Goodman
Shanil Ebrahim
Despina Contopoulos-Ioannidis
Georgia Salanti
Chirag Patel
Lars Hemkens
Ann Hsing
Lamberto Manzoli
Maria Elena Flacco
George Siontis
Denes Szucs
Kostas Siontis
Vangelis Evangelou
Kristin Sainani
Muin Khoury
Orestis Panagiotou
Florence Bourgeois

Special thanks

Joseph Lau
Malcolm MacLeod
Marcus Munafo
David Allison
Josh Wallach
Fotini Karassa
Athina Tatsioni
Evi Ntzani
Ioanna Tzoulaki
Demos Katritsis
Nikos Patsopoulos
Fainia Kavvoura
Brian Nosek
Victoria Stodden
Ele Zeggini
Belinda Burford
Kostas Tsilidis
Jodi Prochaska

Special thanks

Charitini Stavropoulou
Evropi Theodoratou
Nikos Pandis
Huseyin Naci
Vanesa Bellou
Antony Doufas
Lazaros Belbasis
Chris Doucouliagos
Stelios Serghiou
Anna Chaimani
Fotini Chatzinasiou
Stephania Papatheodorou
Florian Naudet
Tom Hardwicke
Perrine Janiaud
Ioana-Alina Cristea
Shannon Brownlee
Vikas Saini
Matthias Egger
Patrick Bossuyt
Andre Uitterlinden
Doug Altman
Deb Zarin
Katherine Flegal

Special thanks

Shanthi Kapaggoda
Ewoud Schuit
Stefania Boccia
David Chavalarias
Jennifer Ware
Viswam Nair
Stephan Bruns
Dorothy Bishop
Tom Trikalinos
Kristina Sundquist
Johanna Int'hout
Kevin Boyack
Brett Thombs
Raj Manrai
Nazmus Saquib
Elizabeth Iorns
Abraham Verghese
Euan Ashley