VALIDATION STUDIES: ISPE GUIDELINE

PROFESSOR VERA EHRENSTEIN
DEPARTMENT OF CLINICAL EPIDEMIOLOGY, AARHUS UNIVERSITY

Misclassified present-day Bahamas as Spain

Christopher Columbus

https://www.biography.com/history-culture

Misclassified Cuba as mainland China

Christopher Columbus

Amerigo Vespucci

https://www.biography.com/history-culture

MISCLASSIFICATION IN EPIDEMIOLOGY

Wrong classification of study subjects:

- Classifying treated as untreated and vice versa
- Classifying cases as noncases and vice versa
- Classifying smokers as nonsmokers and vice versa
- More complex scenarios with > 2 levels

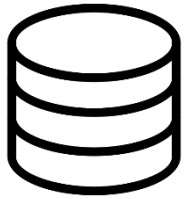
a.k.a.: measurement error

a.k.a.: information bias

SOURCES OF ERROR

At patient level

- Coding errors



ICD-10 G40 Epilepsy	69	100.0%
Epilepsy	52	75.4%
Seizures	5	7.2%
Suspected seizures	5	7.2%
Coding error	2	2.9%
Asphyxia	1	1.4%
Mental retardation	1	1.4%
Unspecified neurologic problem	1	1.4%
Heart failure	2	2.9%

MEDICAL CHART REVIEW

https://www.ncbi.nlm.nih.gov/pubmed/16451724

ISPE MANUSCRIPT: THE MOTIVATION

Received: 15 May 2020 | Accepted: 22 December 2020
DOI: 10.1002/pds.5193

Check for updates

ORIGINAL ARTICLE

WILEY

Outcomes in patients with lung cancer treated with crizotinib and erlotinib in routine clinical practice: A post-authorization safety cohort study conducted in Europe and in the United States

Vera Ehrenstein[1] | Kui Huang[2] | Johnny Kahlert[1] | Shahram Bahmanyar[3] | Pär Karlsson[3] | Lukas Löfling[3] | Anthony P. Nunes[4] | Cheryl Enger[4] | Irene D. Bezemer[5] | Josephina G. Kuiper[5] | Fabian Hoti[6] | Rosa Juuti[6] | Pasi Korhonen[6] | Jingping Mo[2] | Stephen E. Schachterle[2,7] | Keith D. Wilner[2] | Mikael Rørth[1] | Henrik T. Sørensen[1]

QTc prolongation (ECG records reviewed by Adjudication Committee on request)
If the abstraction form includes evidence of any of the following, then event is definite:

Severe/symptomatic

- Clinical diagnosis of polymorphic ventricular tachycardia rate lasting for more than 5 seconds or ventricular fibrillation, documented by an ECG recording;
- Syncope recorded in case notes;
- Clinical diagnosis of torsade de pointes or TdP documented by an ECG recording;
- Clinical diagnosis of sudden death or sudden cardiac death;

Mild/asymptomatic

- ECG showing Fridericia rate-corrected QT interval ($QT/RR^{1/3}$) >450 ms in men or >470 ms in women;
- QTcF (Fridericia correction) change $\geq$ 60 msec from baseline;
- Absolute QT or QTc of >500 msec.

Exclude: Known coronary syndromes or known congenital long QT syndromes

Vision disorders

If the abstraction form includes evidence of any of the following then event is definite:

- Clinical diagnosis of vision disturbances including blurred vision, photophobia, photopsia, palinopsia (ask about multiple sclerosis), visual illusion, reduced visual acuity, diplopia, visual impairment, visual field defect, and vitreous floaters, maculopathy, retinal edema, retinal hemorrhage.

Exclude:

- Refractive error, amblyopia, corneal disorder (abnormal sensation in eye, anterior chamber collapse, anterior chamber opacity, aqueous humour leakage, asthenopia, chemical burns of eye, chemical eye injury, contact lens intolerance, corneal suture, corneal sutures removal, deposit eye, dry eye, eye burns, eye inflammation, eye injury, eye irritation, eye laser surgery, eye operation complication, eye penetration, flat anterior chamber of eye, foreign body in eye, foreign body sensation in eyes, hypoaesthesia eye, ocular toxicity, slit-lamp tests abnormal, superficial injury of eye, thermal burns of eye, vitamin A deficiency eye disorder, xerophthalmia, acquired corneal dystrophy, allergic keratitis, arcus lipoides, atopic keratoconjunctivitis, benign neoplasm of cornea, biopsy cornea, biopsy cornea abnormal, bowman's membrane disorder, corneal abrasion, corneal bleeding,corneal cyst, corneal decompensation, corneal defect, corneal degeneration, corneal deposits, corneal diameter decreased, corneal diameter increased, corneal disorder, corneal endothelial cell loss, corneal endotheliitis, corneal epithelial microcysts, corneal epithelium defect, corneal erosion, corneal exfoliation, corneal flap complication, corneal graft rejection, corneal hypertrophy, corneal implant, corneal infiltrates, corneal lesion, corneal lesion removal, corneal light reflex test abnormal, corneal oedema, corneal opacity, corneal operation, corneal perforation, corneal pigmentation, corneal reflex decreased, corneal scar, corneal staining, corneal striae, corneal thickening, corneal thinning, corneal touch, corneal transplant, corneoconjunctival intraepithelial neoplasia, dellen, detached Descemet's membrane, diffuse lamellar keratitis, injury corneal, iridocorneal endothelial syndrome, Kayser-Fleischer ring, keratectomy, keratitis, keratitis interstitial, keratitis sclerosing, keratoconus, keratomalacia, keratometry, keratomileusis, keratopathy, keratorhexis, keratotomy, limbal hyperaemia, limbal swelling, macrocornea, malignant neoplasm of cornea, microcornea, neoplasm of cornea unspecified malignancy, neurotrophic keratopathy, photokeratitis, photorefractive keratectomy, punctate keratitis, Terrien's marginal degeneration, topography corneal abnormal, ulcerative keratitis, vital dye staining cornea present, vitamin A deficiency related corneal disorder), visual impairing cataracts, uncontrolled diabetes, brain tumor, age-related macular degeneration, toxic maculopathy (eg, Chloroquine or Tamoxifen)

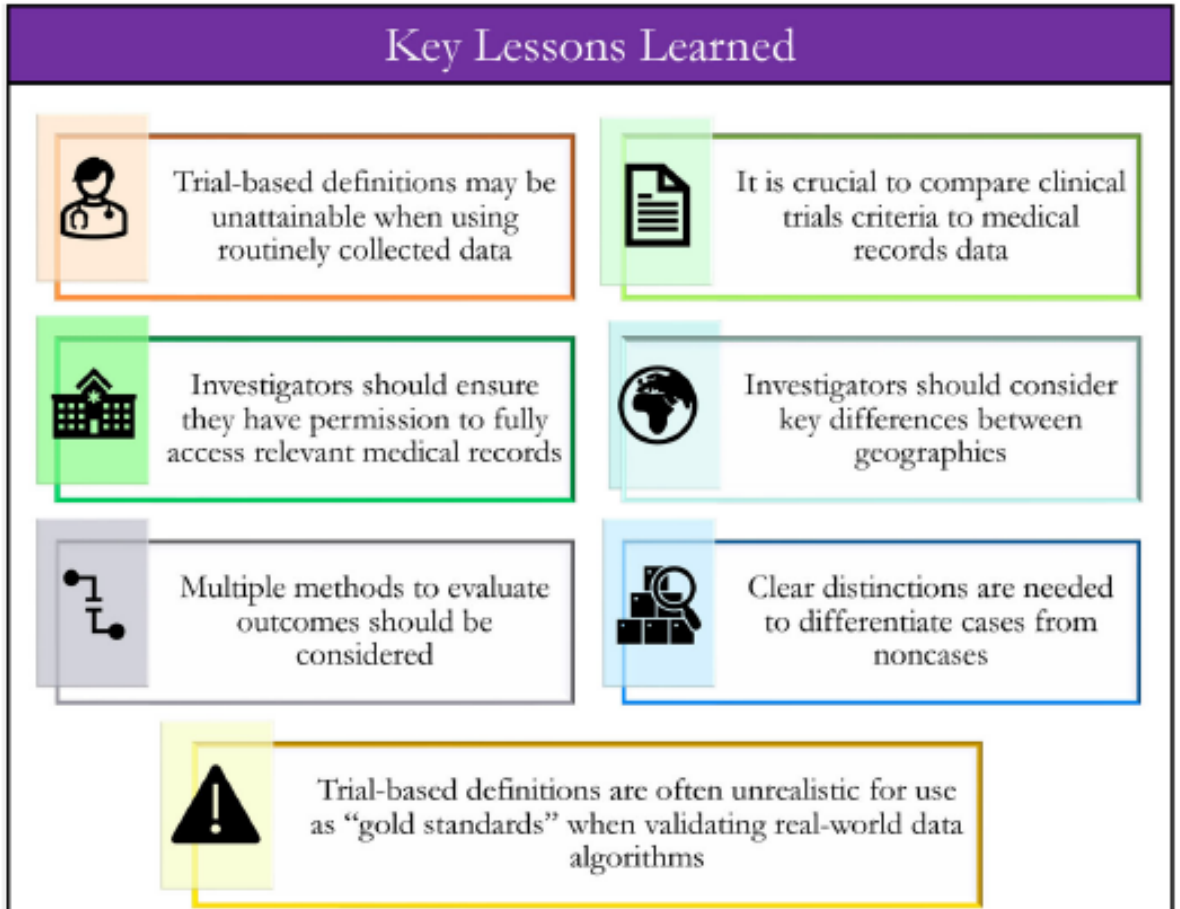
Received: 19 April 2022 | Revised: 18 November 2022 | Accepted: 29 November 2022
DOI: 10.1002/pds.5582

WILEY

COMMENTARY

Validation of safety outcomes in routinely collected data: Lessons learned from a multinational postapproval safety study

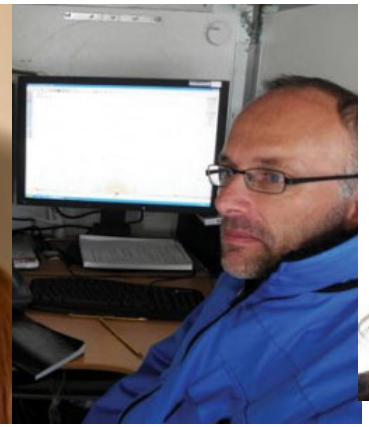
Patrick J. Arena[1,2] | Kui Huang[1] | Lukas Löfling[3,4] | Shahram Bahmanyar[3] | Jingping Mo[5] | Stephen E. Schachterle[1] | Anthony P. Nunes[6,7] | Elisabeth Smits[8] | Rosa Juuti[9] | Fabian Hoti[9] | Pasi Korhonen[9] | Kasper Adelborg[10] | Jens Sundbøll[10] | Torben Riis Rasmussen[11] | Anders Løkke[12,13] | Vera Ehrenstein[9]



American Journal of Epidemiology, 2024, 00, 1–13
https://doi.org/10.1093/aje/kwae071
Advance access publication date May 17, 2024
Practice of Epidemiology

Validation of algorithms in studies based on routinely collected health data: general principles

Vera Ehrenstein*[,1], Maja Hellfritzsch[2,3], Johnny Kahlert[1], Sinéad M. Langan[4], Hisashi Urushihara[5], Danica Marinac-Dabic[6], Jennifer L. Lund[7], Henrik Toft Sørensen[1], Eric I. Benchimol[8,9,10]





American Journal of Epidemiology, 2024, 00, 1–13
https://doi.org/10.1093/aje/kwae071
Advance access publication date May 17, 2024
Practice of Epidemiology

Validation of algorithms in studies based on routinely collected health data: general principles

Vera Ehrenstein*[,1], Maja Hellfritzsch[2,3], Johnny Kahlert[1], Sinéad M. Langan[4], Hisashi Urushihara[5], Danica Marinac-Dabic[6], Jennifer L. Lund[7], Henrik Toft Sørensen[1], Eric I. Benchimol[8,9,10]

[1]Department of Clinical Epidemiology, Department of Clinical Medicine, Aarhus University and Aarhus University Hospital, 8200 Aarhus N, Denmark
[2]Research Unit of Clinical Pharmacology, Pharmacy, and Environmental Medicine, University of Southern Denmark, 5230 Odense M, Denmark
[3]Department of Cardiology, Gødstrup Hospital, 7400 Herning, Denmark
[4]Department of Non-communicable Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London WC1E 7HT, United Kingdom
[5]Division of Drug Development and Regulatory Science, Faculty of Pharmacy, Keio University, Tokyo 105-8512, Japan
[6]Office of Clinical Evidence and Analysis, Center for Devices and Radiological Health, US Food and Drug Administration, Silver Spring, MD 20993, United States
[7]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, United States
[8]Division of Gastroenterology, Hepatology and Nutrition and Child Health Evaluative Sciences, SickKids Research Institute, The Hospital for Sick Children, Toronto, ON M5G 1X8, Canada
[9]Department of Paediatrics and Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON M5T 3M6, Canada
[10]Institute for Clinical Evaluative Sciences, Toronto, ON M4N 3M5, Canada
*Corresponding author: Vera Ehrenstein, Department of Clinical Epidemiology, Department of Clinical Medicine, Aarhus University and Aarhus University Hospital, Olof Palmes Allé 43-45, 8200 Aarhus N, Denmark (ve@clin.au.dk)

Abstract

Clinicians, researchers, regulators, and other decision-makers increasingly rely on evidence from real-world data (RWD), including data routinely accumulating in health and administrative databases. RWD studies often rely on algorithms to operationalize variable definitions. An algorithm is a combination of codes or concepts used to identify persons with a specific health condition or characteristic. Establishing the validity of algorithms is a prerequisite for generating valid study findings that can ultimately inform evidence-based health care. In this paper, we aim to systematize terminology, methods, and practical considerations relevant to the conduct of validation studies of RWD-based algorithms. We discuss measures of algorithm accuracy, gold/reference standards, study size, prioritization of accuracy measures, algorithm portability, and implications for interpretation. Information bias is common in epidemiologic studies, underscoring the importance of transparency in decisions regarding choice and prioritizing measures of algorithm validity. The validity of an algorithm should be judged in the context of a data source, and one size does not fit all. Prioritizing validity measures within a given data source depends on the role of a given variable in the analysis (eligibility criterion, exposure, outcome, or covariate). Validation work should be part of routine maintenance of RWD sources.

This article is part of a Special Collection on Pharmacoepidemiology.

Key words: algorithms; data quality; information bias; measurement error; misclassification; routinely collected health data; real-world data; validity.

Introduction

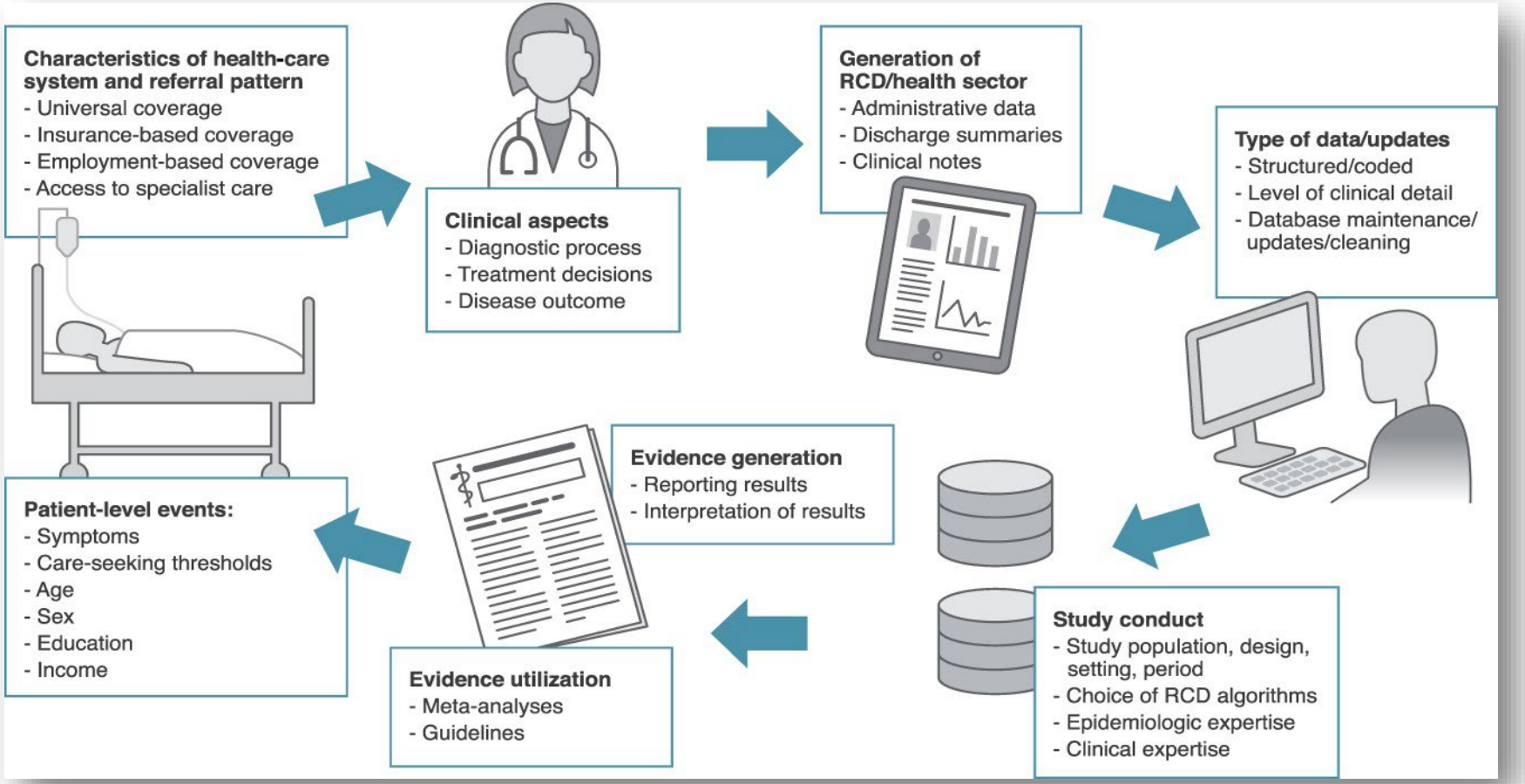
Real-world data (RWD) and real-world evidence (RWE) originate outside phase I-III (preapproval) interventional trials.[1,2] Routinely collected health data (RCD) accrue as a by-product of health-care delivery and encompass electronic health records; health administrative data; claims; electronic records of diagnoses, treatments, procedures, or devices; and patient-reported events.[3] The role of RCD-based RWE is increasing in decision-making by clinicians, regulators, and health-care payers.[1,2,4–11] This development predicts an ever-strengthening role of RWE in research on disease prevention, etiology, and clinical course.[12]

Advantages of RWD include accrual independent of research questions, population-wide coverage, sustainability, structured format, and standard coding.[1,12] Measurement error, inevitably introduced along the path from the point of care to an analytical data point, may translate into systematic errors (or information bias) on the level of study results and, ultimately, to misleading clinical practice.[13] Large-scale RWD-based studies yielding precise but potentially biased results may exacerbate the fallacy of mistaking precision for validity.[14] Validation studies of RWD and reporting guidelines have helped improve transparency[15–19]; however, there remains an unmet need for education and awareness about the conduct of validation studies in RWD-based research.[12,20–27] A validation study aims to quantify the validity of a measurement instrument. In (pharmaco)epidemiologic studies, operational definitions of the study variables—exposure, outcomes, covariates—are the measurement instruments (commonly called algorithms) whose validity is estimated against a chosen standard.[26]

This paper outlines universal principles of validation "best practice," including the basic terminology, the underlying theory, and the practical approaches. After defining an algorithm,

Received: February 28, 2023. Accepted: May 14, 2024

Bookmarks

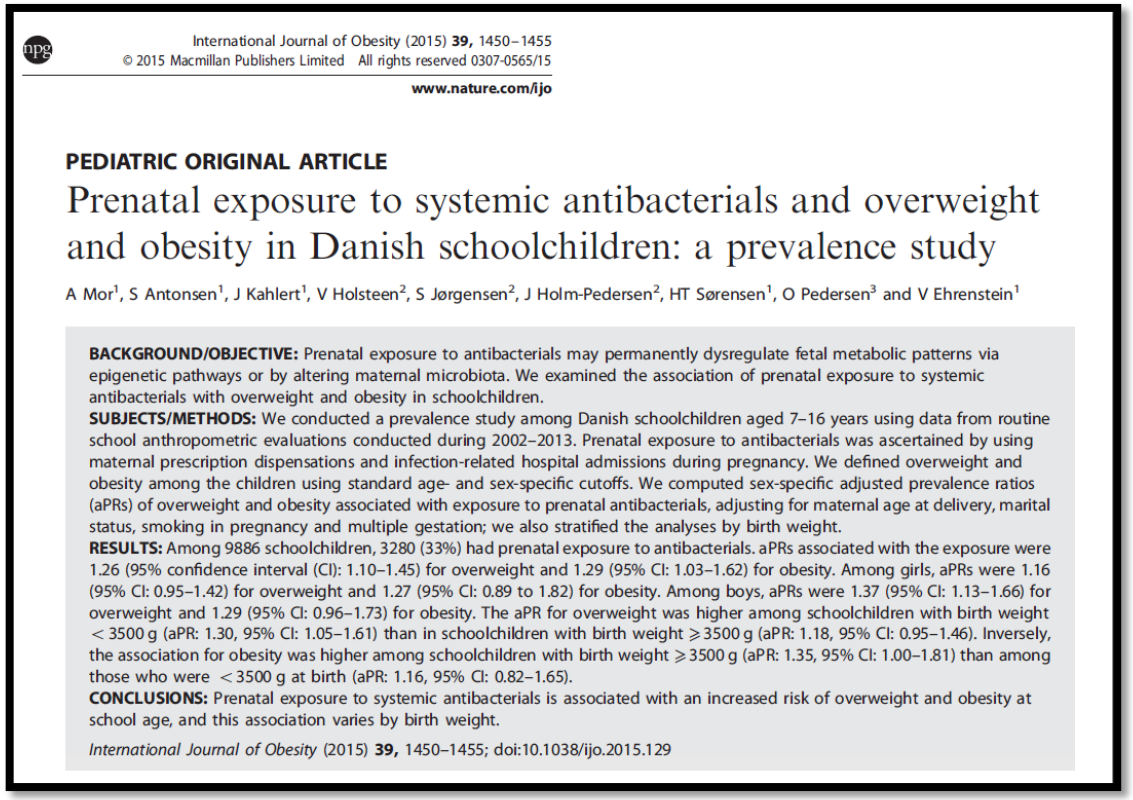


Am J Epidemiol, Volume 193, Issue 11, November 2024, Pages 1612–1624, https://doi.org/10.1093/aje/kwae071

THE DIAGNOSIS PARADIGM

ALGORITHM/PHENOTYPE: BACTERIAL INFECTION

Algorithm	Health state of interest
Dispensing with ATC J01 (antibacterials for systemic use)	Bacterial infection
Dispensing with ATC J01 (antibacterials for systemic use)	Exposure to a systemic antibacterial agent



International Journal of Obesity (2015) **39,** 1450–1455

www.nature.com/ijo

PEDIATRIC ORIGINAL ARTICLE

Prenatal exposure to systemic antibacterials and overweight and obesity in Danish schoolchildren: a prevalence study

A Mor[1], S Antonsen[1], J Kahlert[1], V Holsteen[2], S Jørgensen[2], J Holm-Pedersen[2], HT Sørensen[1], O Pedersen[3] and V Ehrenstein[1]

BACKGROUND/OBJECTIVE: Prenatal exposure to antibacterials may permanently dysregulate fetal metabolic patterns via epigenetic pathways or by altering maternal microbiota. We examined the association of prenatal exposure to systemic antibacterials with overweight and obesity in schoolchildren.
SUBJECTS/METHODS: We conducted a prevalence study among Danish schoolchildren aged 7–16 years using data from routine school anthropometric evaluations conducted during 2002–2013. Prenatal exposure to antibacterials was ascertained by using maternal prescription dispensations and infection-related hospital admissions during pregnancy. We defined overweight and obesity among the children using standard age- and sex-specific cutoffs. We computed sex-specific adjusted prevalence ratios (aPRs) of overweight and obesity associated with exposure to prenatal antibacterials, adjusting for maternal age at delivery, marital status, smoking in pregnancy and multiple gestation; we also stratified the analyses by birth weight.
RESULTS: Among 9886 schoolchildren, 3280 (33%) had prenatal exposure to antibacterials. aPRs associated with the exposure were 1.26 (95% confidence interval (CI): 1.10–1.45) for overweight and 1.29 (95% CI: 1.03–1.62) for obesity. Among girls, aPRs were 1.16 (95% CI: 0.95–1.42) for overweight and 1.27 (95% CI: 0.89 to 1.82) for obesity. Among boys, aPRs were 1.37 (95% CI: 1.13–1.66) for overweight and 1.29 (95% CI: 0.96–1.73) for obesity. The aPR for overweight was higher among schoolchildren with birth weight < 3500 g (aPR: 1.30, 95% CI: 1.05–1.61) than in schoolchildren with birth weight $\geqslant 3500$ g (aPR: 1.18, 95% CI: 0.95–1.46). Inversely, the association for obesity was higher among schoolchildren with birth weight $\geqslant 3500$ g (aPR: 1.35, 95% CI: 1.00–1.81) than among those who were < 3500 g at birth (aPR: 1.16, 95% CI: 0.82–1.65).
CONCLUSIONS: Prenatal exposure to systemic antibacterials is associated with an increased risk of overweight and obesity at school age, and this association varies by birth weight.

International Journal of Obesity (2015) **39,** 1450–1455; doi:10.1038/ijo.2015.129

ALGORITHM/PHENOTYPE: SERIOUS INFECTION

	A	B	C	D	E	F	G	H	I	J	K
1	InfectionType	ICD10	ICD10Text	Exclusion	Exclusiontext	Notes	DiagnosesTypes	PatientType			
2	Intra-abdominal infe	A00	Cholera				C_ADIAG	0 (Inpatient 1+ overnight stay)			
3	Intra-abdominal infe	A01	Typhoid and paratyphoid fevers				C_ADIAG	0 (Inpatient 1+ overnight stay)			
4	Intra-abdominal infe	A02	Other salmonella infections	A021, A022C	Excluding A021 Salmonella sepsis, A022C S	Included in other types	C_ADIAG	0 (Inpatient 1+ overnight stay)			
5	Sepsis	A021	Salmonella sepsis				C_ADIAG	0 (Inpatient 1+ overnight stay)			
6	Infections of CNS	A022C	Salmonella meningitis				C_ADIAG	0 (Inpatient 1+ overnight stay)			
7	Intra-abdominal infe	A03	Shigellosis				C_ADIAG	0 (Inpatient 1+ overnight stay)			
8	Intra-abdominal infe	A04	Other bacterial intestinal infections				C_ADIAG	0 (Inpatient 1+ overnight stay)			
9	Intra-abdominal infe	A05	Other bacterial foodborne intoxications, not elsewhere classified				C_ADIAG	0 (Inpatient 1+ overnight stay)			
10	Intra-abdominal infe	A06	Amoebiasis				C_ADIAG	0 (Inpatient 1+ overnight stay)			
11	Intra-abdominal infe	A07	Other protozoal intestinal diseases				C_ADIAG	0 (Inpatient 1+ overnight stay)			
12	Intra-abdominal infe	A08	Viral and other specified intestinal infections				C_ADIAG	0 (Inpatient 1+ overnight stay)			
13	Intra-abdominal infe	A09	Other gastroenteritis and colitis of infectious and unspecified origin				C_ADIAG	0 (Inpatient 1+ overnight stay)			
14	Infections of CNS	A170	Tuberculous meningitis				C_ADIAG	0 (Inpatient 1+ overnight stay)			
15	Other	A20	Plague	A203	Exclude A203 Plague meningitis	Included in other types	C_ADIAG	0 (Inpatient 1+ overnight stay)			
16	Infections of CNS	A203	Plague meningitis				C_ADIAG	0 (Inpatient 1+ overnight stay)			
17	Other	A21	Tularaemia				C_ADIAG	0 (Inpatient 1+ overnight stay)			
18	Other	A22	Anthrax	A227	Exclude A227 Anthrax sepsis	Included in other types	C_ADIAG	0 (Inpatient 1+ overnight stay)			
19	Sepsis	A227	Anthrax sepsis				C_ADIAG	0 (Inpatient 1+ overnight stay)			
20	Other	A23	Brucellosis				C_ADIAG	0 (Inpatient 1+ overnight stay)			
21	Other	A24	Glanders and melioidosis				C_ADIAG	0 (Inpatient 1+ overnight stay)			
22	Other	A25	Rat-bite fevers				C_ADIAG	0 (Inpatient 1+ overnight stay)			
23	Other	A26	Erysipeloid				C_ADIAG	0 (Inpatient 1+ overnight stay)			
24	Other	A27	Leptospirosis				C_ADIAG	0 (Inpatient 1+ overnight stay)			
25	Other	A28	Other zoonotic bacterial diseases, not elsewher	A282B		Included in other types	C_ADIAG	0 (Inpatient 1+ overnight stay)			
26	Sepsis	A282B	Yersinia sepsis				C_ADIAG	0 (Inpatient 1+ overnight stay)			
27	Other	A30	Leprosy [Hansen disease]				C_ADIAG	0 (Inpatient 1+ overnight stay)			
28	Other	A31	Infection due to other mycobacteria				C_ADIAG	0 (Inpatient 1+ overnight stay)			

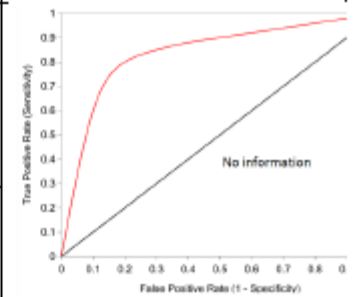
Table 1. Glossary of important epidemiologic terms used in this paper.

Term	Definition/elaboration
Algorithm	An algorithm is defined as a combination of people with a specific health condition, *health state*, used to classify persons with algorithm may be as simple as a single meeting algorithm criteria are classified
Alloyed gold standard	See *reference standard*.
Area under the curve	Area under the ROC curve, used as a single values of a continuous variable. See Tab
Completeness	The completeness of a data source is the p population captured in that data source
Confounder	A confounder can be conceptualized as a c development. Thus, a confounder cann study, a characteristic that fulfills the a exposure categories *and* (2) predicts the
Covariate	A characteristic of members of the study p term is often used for predictors of the associated with the exposure of interes
Diagnostic odds ratio	The ratio of the odds of algorithm positivity those without the health state of intere
Exposure	A health state that, in a given study, explici treatment allocation in randomized tri status (medicinal, surgical, device), but socioeconomic status). Conceptually, ex
External validation	Validation of events of a health state identi
Gold standard	In an ideal sense, a method that classifies i standards rarely exist (see *reference stan* gold standards may be time-dependent methods previously considered gold sta
Health state	A generic term used in this paper for any e health state may play the role of expos
Information bias	Discrepancy between the true value of a gi value. Errors may be inherent in a meas (eg, data entry errors). *Information bias* i
Internal validation	Validation of events of a health state identi
Measurement error	See *information bias*.
Misclassification	Incorrect classification by an *algorithm* of a body mass index in the normal range n or weight).
Negative predictive value	Proportion of patients who are truly negati algorithm. See Table S2 for details.
Outcome	A health state that, in a given study, plays outcome of treatment (benefit or risk), infarction), or death. Conceptually, expo
Positive predictive value	Proportion of patients who are truly positiv algorithm. See Table S2 for details.
Real-world data	Data that originate in the course of routine clinical practice, usually by contrast with data originating in phase I-III clinical trials.
Real-world evidence	Evidence generated using RWD.
ROC curve	For algorithms that classify patients into health states (eg, presence/absence/severity), a plot of sensitivity (or true positive proportion) against 1 – specificity (or false-positive proportion). See Table S2 for details.
Reference standard	A method or data source that is expected to classify individuals with respect to a health state of interest better than the algorithm being validated. Alternative term: *alloyed gold standard*.
Routinely collected health data	Data that accrue routinely as a by-product of health-care delivery or administration of the health-care system. Routinely collected data may also include nonhealth characteristics such as income, education, and employment, if they are important in studying health states. Routinely collected health data are a type of RWD.
Sensitivity	Proportion of persons with a given health state according to a gold/reference standard who are classified as positive by an (RWD) algorithm. See Table S2 for details.
Specificity	Proportion of persons without a given health state according to a gold/reference standard who are classified as negative by an (RWD) algorithm. See Table S2 for details.
Validation cohort	Study population used to evaluate measures of algorithm validity.
Validity (internal)	Correspondence between an estimate and the true value of a parameter. In epidemiologic studies, a valid (unbiased) estimate may estimate the occurrence of a health event, or an association between exposure and outcome.

Table S1. Examples of RWD-based algorithms and validation studies of RWD-based algorithms c

Health state or event being validated	Country	RWD source(s)	RWD-based algorithm	Gold/reference standard
In-hospital chemotherapy treatment	Denmark	Danish National Patient Registry	Agent-specific treatment codes	In-hospital pharmacy production system and chart review.
Start, duration and end of prescribed medicine	Finland	Finnish Prescription Registry	Different approaches to estimation of drug use periods based on prescription data. Three fixed methods were included and one data-driven method (PRE2DUP)	Expert review of purchase history.
Cardiac interventions	Denmark	Danish National Patient Registry	Surgical and procedure codes according to the NSCP classification	Chart review.
Osteonecrosis of the jaw	Denmark	Danish National Patient Registry	ICD-10 codes recorded at departments of oral and maxillofacial surgery in cancer patients	Chart review and an independent sample of confirmed cases.
Colorectal cancer recurrence	Denmark	Danish National Patient Registry and the Danish Pathology Registry	Combination of codes compatible with metastatic disease, chemotherapy administration, and cancer recurrence in patients previously diagnosed with non-metastatic colorectal cancer	Independent sample of patients with known recurrence status.
Frailty in the elderly	United States	Administrative claims	Medicare claims-based algorithm of dependency in activities of daily living (or dependency) developed as a proxy for frailty	A reference standard measure of phenotypic frailty.
Hepatic encephalopathy	United States	Administrative claims	ICD-10 codes and treatment proxies after switch to ICD-10 in 2015 rendered algorithms	Independent prospectively acquired sample containing both true-positive and

Table S2. Measures of algorithm validity

Measure	Range	Description	Formulae
Sensitivity	[0, 1]	Proportion of persons with a given health state according to gold/reference standard who are classified as such by an algorithm.	Health state status based on gold/reference standard: Positive / Negative Health state status based on algorithm: Positive — A (True positive) / C (False negative); Negative — B (False positive) / D (True negative) $Sensitivity = \frac{A}{(A+B)} =$
Specificity	[0, 1]	Proportion of persons without a given health state according to gold/reference standard who are classified as such by an algorithm.	$Specificity = \frac{D}{(C+D)} =$
Positive predictive value (PPV)	[0, 1]	Proportion of patients who truly have the health state among all those who are classified as positive by the algorithm.	$PPV = \frac{A}{(A+C)} =$
Negative predictive value (NPV)	[0, 1]	Proportion of patients who truly do not have the health state among all those who are classified as negative by the algorithm. Chubak et al and Benchimol et al. provide formulae connecting sensitivity and specificity, PPV and NPV with prevalence of a health state.[12]	$NPV = \frac{D}{(B+D)} =$
Receiver operating characteristic (ROC) curve	Definition of an RWD-based algorithm may be based on a cut-off value of a measured continuous variable (e.g., hyponatremia based on serum potassium levels in the study by Holland-Bill et al[9]). Generally, more extreme cut-off values of a continuous variable leads to increase in proportion of both false-positives, but also true-positives. For algorithms that classify patients into health states (e.g., presence/absence/severity) an ROC is a plot of sensitivity (or true-positive proportion) on the y-axis against (1-specificity) or false-positive proportion on the x-axis. For a (hypothetical) perfect algorithm, sensitivity=specificity=1.[4]	 Example of a receiver operating characteristic (ROC) curve and area under the curve (AUC). (Sørensen and Vandenbroucke (in press.[9])
Area under the curve (AUC)	AUC is derived from the ROC curve and is a single measure, frequently used to indicate performance of diagnostic tests, or algorithms, as they are used to 'diagnose' presence of a health state in the study population. With varying thresholds for case definition, the points (1-Sensitivity, Specificity) are plotted on the plain with 1-Sensitivity on the x-axis and Specificity on the y-axis to construct the curve. An algorithm with sensitivity = specificity=1 ROC=1; an algorithm that would classify patients not better than a coin toss would have an AUC=0.5.[4]	
Diagnostic odds ratio (DOR)	[0, ∞)	An odds ratio of dichotomous tests in diagnostic application and frequently used for meta-analysis of diagnostic tests. For the purposes of validation studies, DOR can be calculated as the ratio of the odds of positivity in those with a health state of interest relative to the odds of positivity in those without the health state of interest.[4] DOR does not depend on prevalence of health state, with higher values indicating better discriminatory test performance and used in combination with sensitivity and specificity. A value of 1 means that the test does not discriminate between cases and non-cases. Higher DOR corresponds to higher probability of an algorithm to be positive among true cases than in non-cases of a given health state.	$DOR = \frac{A\ (True\ Positive)}{B\ (False\ Negative)} / \frac{C\ (False\ Positive)}{D\ (True\ Negative)}$
Kappa statistic	[-1, 1]	Kappa statistic is used to quantify interrater variability.[4] In evaluating an RWD-based algorithm, it can be used to quantify agreement between two algorithms or data sources, none of which can be considered better (a reference standard) relative to the other. Based on the value of kappa statistic, agreement may be qualified on a scale 'less than chance' to 'almost perfect (Viera and Garrett 2005[7]). Kappa — Agreement	Health state status according to data source 1: Positive / Negative Health state status according to data source 2: Positive — a / b / n_1; Negative — c / d / n_0 n_1 n_0 n Expected agreement between two data sources/algorithm, none of which is superior to the other in classifying a given health state into present/absent

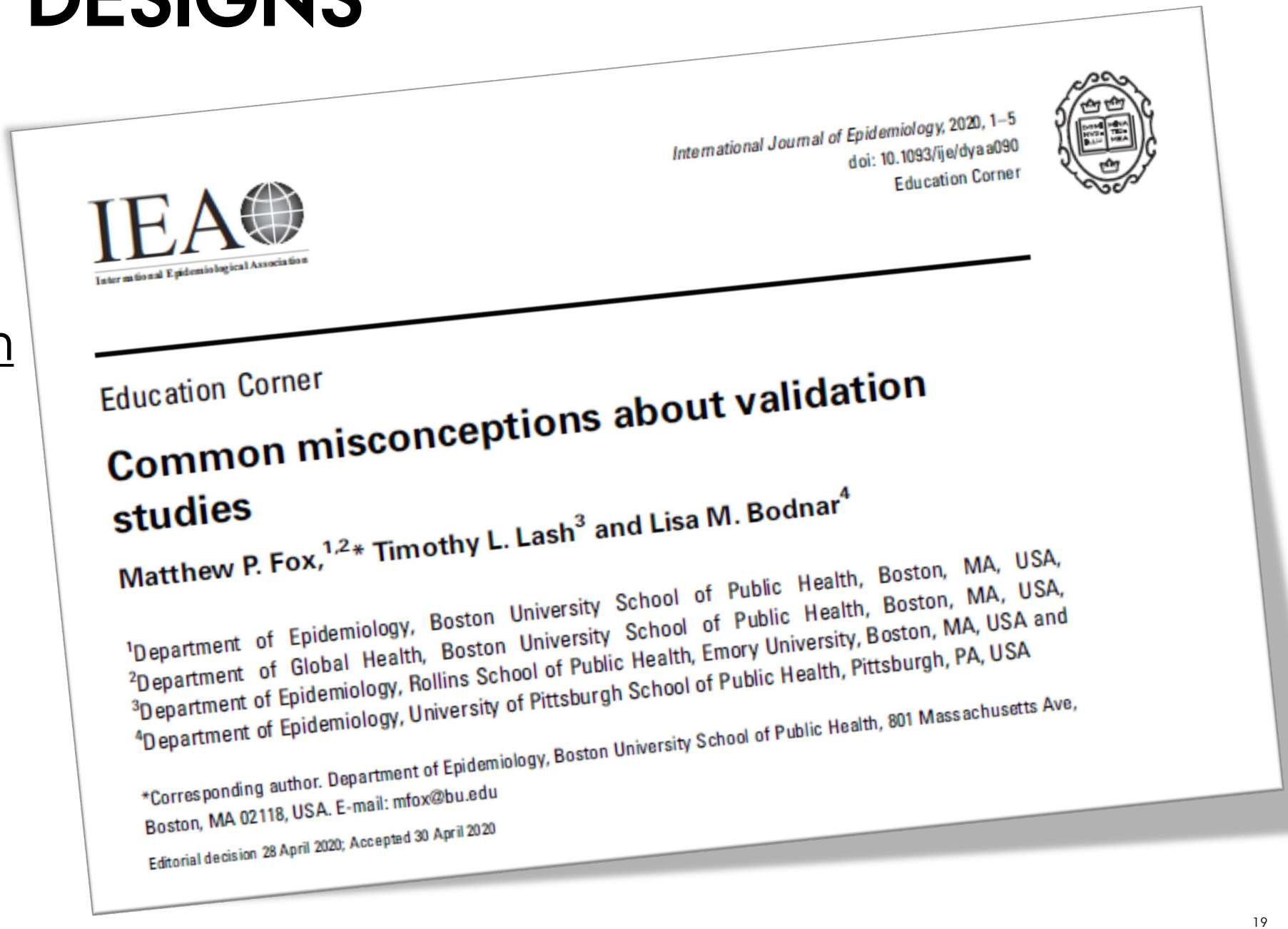
		<0 — Less than chance 0.01–0.20 — Slight 0.21–0.40 — Fair 0.41–0.60 — Moderate 0.61–0.80 — Substantial 0.81–0.99 — Almost perfect	$Observed\ agreement\ p_{obs} = \frac{(a+d)}{n}$ $Expected\ agreement\ p_{exp} = \left[\left(\frac{n_1}{n}\right) \times \left(\frac{m_1}{n}\right)\right] + \left[\left(\frac{n_0}{n}\right) \times \left(\frac{m_0}{n}\right)\right]$ $Kappa\ statistic\ \kappa = \frac{(p_{obs} - p_{exp})}{(1 - p_{exp})}$

VALIDATION 'DESIGNS'

1: select on algorithm

2: select on truth

3: take them all



International Journal of Epidemiology, 2020, 1–5
doi: 10.1093/ije/dyaa090
Education Corner

IEA
International Epidemiological Association

Education Corner

Common misconceptions about validation studies

Matthew P. Fox,[1,2]* Timothy L. Lash[3] and Lisa M. Bodnar[4]

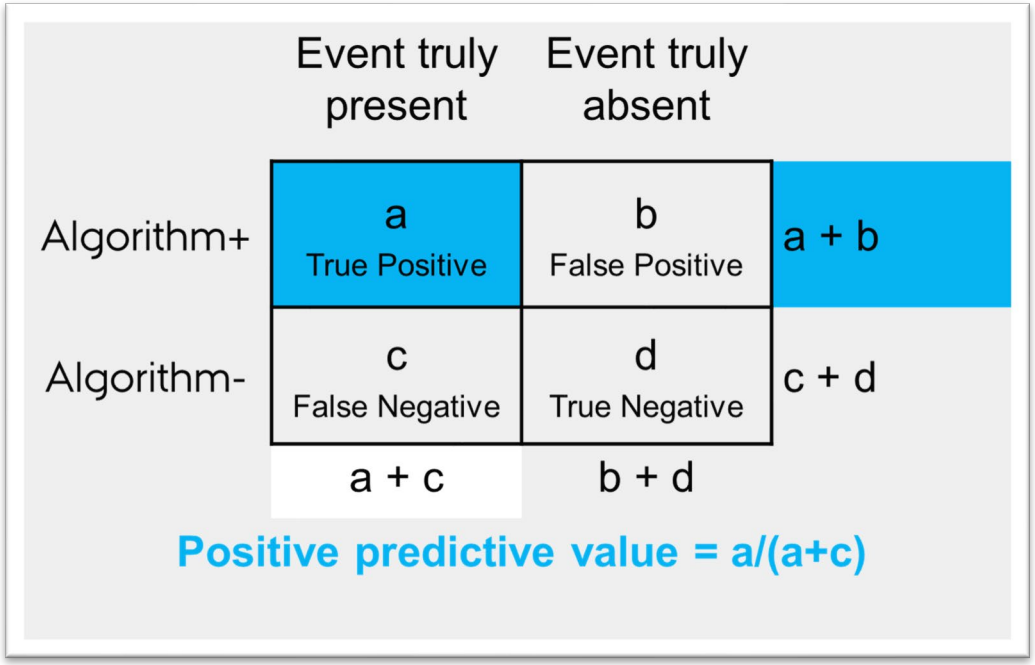
[1]Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA,
[2]Department of Global Health, Boston University School of Public Health, Boston, MA, USA,
[3]Department of Epidemiology, Rollins School of Public Health, Emory University, Boston, MA, USA and
[4]Department of Epidemiology, University of Pittsburgh School of Public Health, Pittsburgh, PA, USA

*Corresponding author. Department of Epidemiology, Boston University School of Public Health, 801 Massachusetts Ave, Boston, MA 02118, USA. E-mail: mfox@bu.edu

Editorial decision 28 April 2020; Accepted 30 April 2020

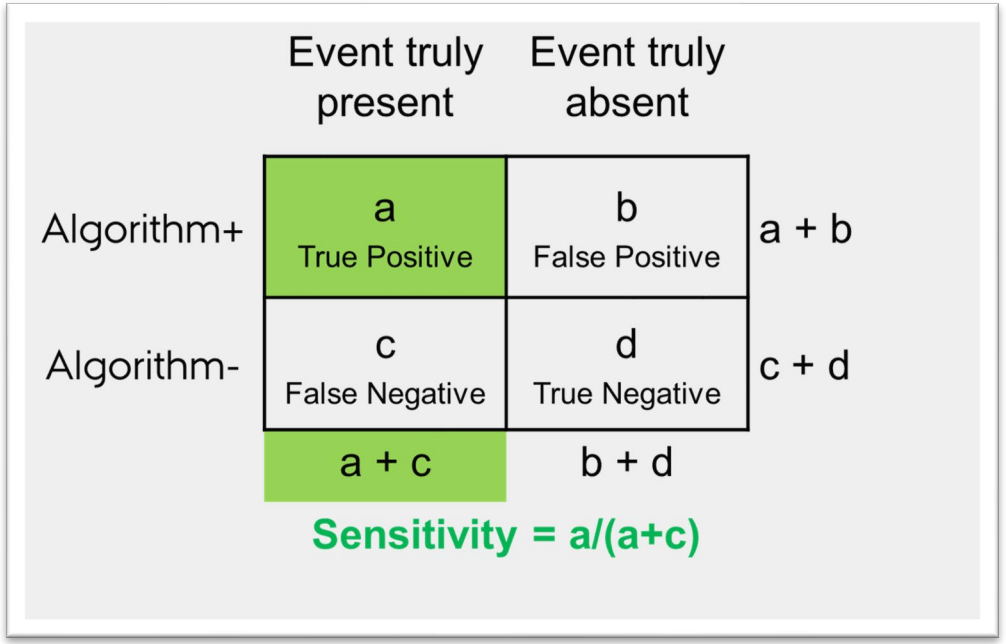
SCENARIO 1: SELECT ON ALGORITHM

- Select **algoirthm-positive** (and negative) persons
- Estimate **PPV** (and NPV)
  - Any component
  - Each component



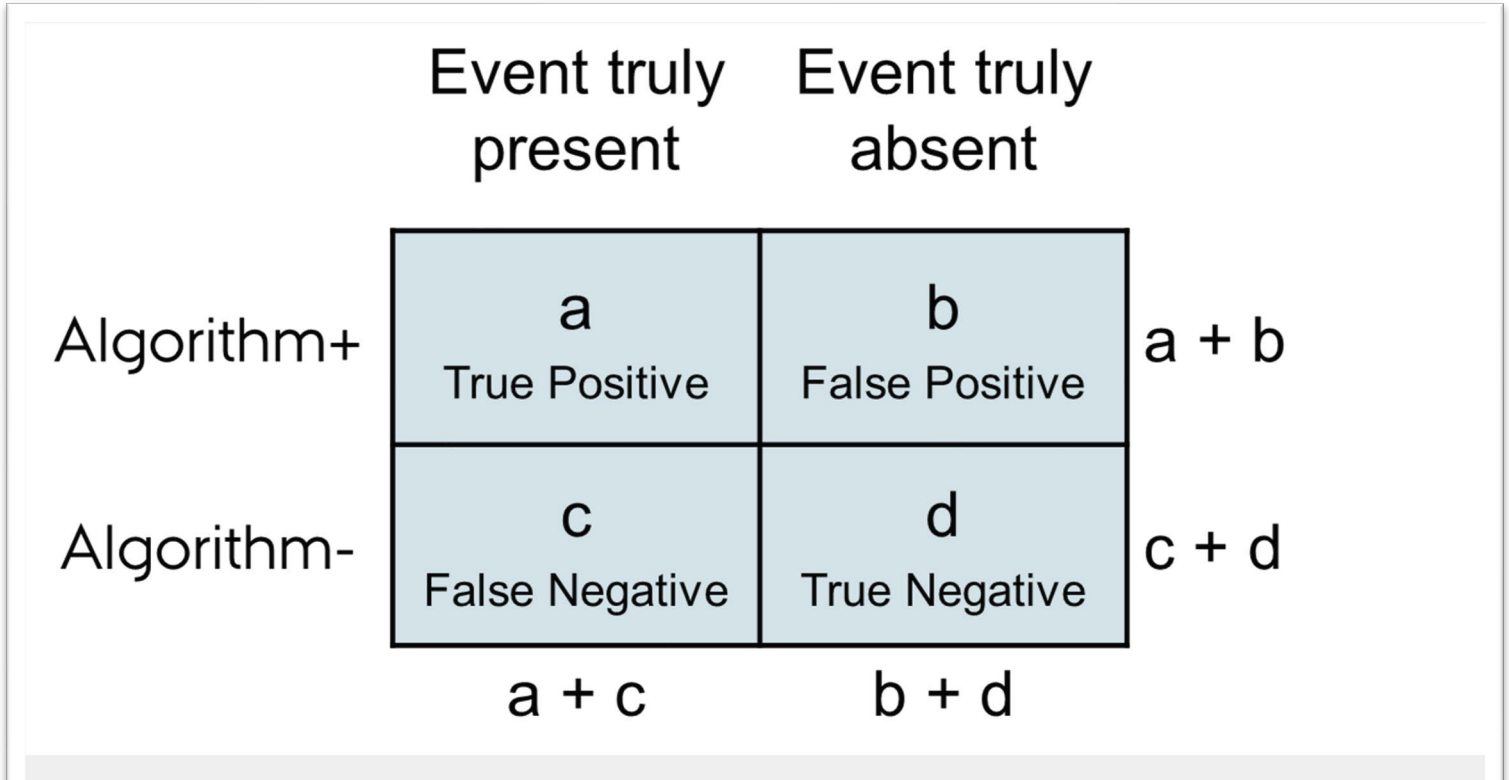
SCENARIO 2: SELECT ON GOLD STANDARD

- Identify an independent sample of **gold-standard-positive events** (and nonevents)
- Estimate **sensitivity** (and specificity)
  - Any component
  - Each component



SCENARIO 3: SELECT INDEPENDENTLY OF BOTH

- Select population regardless of algorithm/gold standard status
- Estimate **sensitivity, specificity, PPV, NPV**
  - Any component
  - Each component



GOLD STANDARD

- Perfect (better) measure of event of interest
- **Case vs. Non-case definition**
- Examples
  - Medical charts
  - Patient histories e-review
  - Another database

MEDICAL CHART REVIEW

- Ascertainment time windows
- Key record types required for identifying the gold/standard (eg, laboratory or diagnostic results)
- Key specialists/departments who should provide those records
- Combine or decouple chart abstraction and case adjudication
- Obtaining additional information, resolving disagreements, and quantifying interrater variability
- Staff skills and need for training in standardized chart abstraction (a pilot medical record review ahead of a large-scale effort will inform the process and help use resources efficiently)
- Whether/to what extent to adopt the event definitions from RCTs

Routinely collected data: the importance of high-quality diagnostic coding to research

Stuart G. Nicholls PhD, Sinéad M. Langan MSc PhD, Eric I. Benchimol MD PhD

■ Cite as: *CMAJ* 2017 August 21;189:E1054-5. doi: 10.1503/cmaj.170807

medRxiv

THE PREPRINT SERVER FOR HEALTH SCIENCES

Cold Spring Harbor Laboratory · BMJ · Yale

Follow this preprint

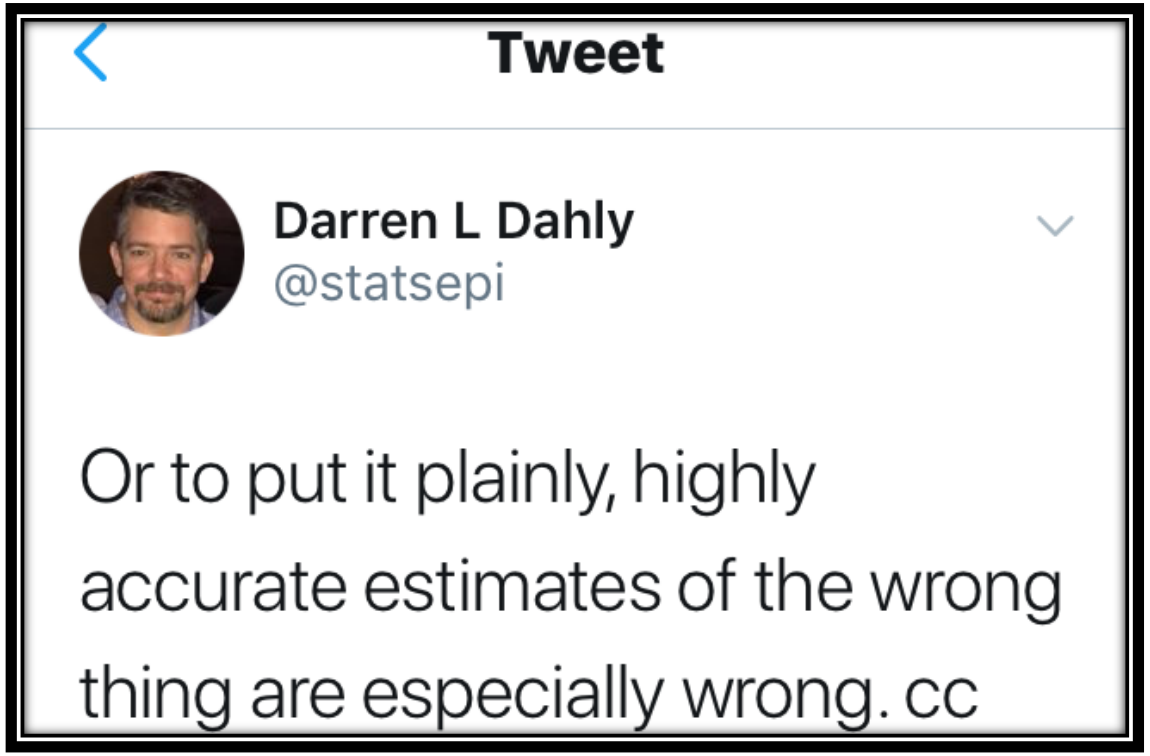
Trends of use of drugs with suggested shortages and their alternatives across 52 real world data sources and 18 countries in Europe and North America

Marta Pineda-Moncusí, Alexandros Rekkas, Álvaro Martínez Pérez, Angela Leis, Carlos Lopez Gomez, Eric Fey, Erwin Bruninx, Filip Maljković, Francisco Sánchez-Sáez, Jordi Rodeiro, Loretta Zsuzsa Kiss, Michael Franz, Miguel-Angel Mayer, Neva Eleangovan, Pericàs Pulido Pau, Pantelis Natsiavas, Selçuk Şen, Steven Cooper, Sulev Reisberg, Katrin Manlik, Beatriz del Pino, Albert Prats Uribe, Ali Yağız Üresin, Ana Danilović Bastić, Ana Maria Rodrigues, Ãngela Afonso, Anna Palomar-Cros, Annelies Verbiest, Antonella Delmestri, Barış Erdoğan, Carina Dinkel-Keuthage, Carmen Olga Torre, Caroline de Beukelaar, Caroline Eteve-Pitsaer, Cátia F. Gonçalves, Costantino de Palma, Cristina Gavina, Daniel Dedman, David Brendan Price, Denisa Gabriela Balan, Dirk Enders, Edward Burn, Elisa Henke, Elyne Scheurwegs, Emma Callewaert, Encarnación Pérez Martínez, Eng Hooi Tan, Fabian Prasser, Francois Antonini, Frank Staelens, Fredrik Nyberg, Geoffray Agard, Gianluigi Galli, Gianmario Candore, Gianny Mestdach, Hadas Shachaf, Harri Rantala, Huiqi Li, Ines Reinecke, Irene López-Sánchez, Jaime E Poquet-Jornet, Javier de la Cruz Bertolo, Jelle Evers, João Firmino-Machado, Jonas Wastesson, Juan Luis Cruz Bermúdez, Juan Manuel Ramírez-Anguita, Kimmo Porkka, Kristina Johnell, Laurent Boyer, Lieselot Cool, Luca Moscetti, Manon Merkelbach, Mariana Canelas-Pais, Massimo Dominici, Máté Szilcz, Matteo Puntoni, Mees Mosseveld, Mina Tadrous, Miquel Oltra-Sastre, Mona Bové, Nadav Rappoport, Noelia García Barrio, Otto Ettala, Paolo Baili, Paula Rubio Mayo, Peter Prinsen, Raeleesha Norris, Ravinder Claire, Reut Sherman Yackob, Roberto Lillini, Salvador Garcia-Torrens, Sampo Kukkurainen, Silvia Lazzarelli, Talita Duarte-Salles, Tiago Taveira-Gomes, Tim Jansen, Ulrich Keilholz, Wai Yi Man, Xintong Li, Zsolt Bagyura, Daniel Prieto-Alhambra, Peter R. Rijnbeek, Theresa Burkard

doi: https://doi.org/10.1101/2024.08.28.24312695

Increasing study size

- Reduces random error
- Does not reduce systematic error



https://www.medrxiv.org/content/10.1101/2024.08.28.24312695v3

Clinical Epidemiology

Dovepress
open access to scientific and medical research

Open Access Full Text Article

EDITORIAL

Helping everyone do better: a call for validation studies of routinely recorded health data

Vera Ehrenstein[1]
Irene Petersen[1,2]
Liam Smeeth[3]
Susan S Jick[4]
Eric I Benchimol[5,6]
Jonas F Ludvigsson[7,8]
Henrik Toft Sørensen[1]

This article wa
Clinical Epidem
12 April 2016
Number of tim

There has
electronic
and diseas
and Drug
of validat
focusing o
papers of t

EPIDEMIOLOGY

Articles & Issues | Collections | Multimedia | For Authors | Journal Info

Next Article >

EPIDEMIOLOGY Announces the "Validation Study" Submission Category

Lash, Timothy L.; Olshan, Andrew F.

Epidemiology: September 2016 - Volume 27 - Issue 5 - p 613–614
doi: 10.1097/EDE.0000000000000532
Commentary

Received: 7 January 2018 | Revised: 5 October 2018 | Accepted: 11 October 2018
DOI: 10.1002/pds.4694

COMMENTARY

WILEY

***Pharmacoepidemiology and Drug Safety*'s special issue on validation studies**

Danielle S. Chun | Jennifer L. Lund | Til Stürmer

Epidemiology, University of North Carolina at Chapel Hill Gillings School of Global Public Health, Chapel Hill, North Carolina, USA

Correspondence
D. S. Chun, Epidemiology, University of North Carolina at Chapel Hill Gillings School of Global Public Health, Chapel Hill, NC, USA.
Email: danielle.chun@unc.edu

..research accomplishes more good and brings greater health to humanity than more intensive doctoring would ever accomplish…

NOTES FROM BEYOND

My Interview With John Snow

Kenneth J. Rothman

Epidemiology. 2004;15

A huge responsibility!

CONSIDERATIONS FOR RESEARCHER

Are there validated algorithms for your variables/database?

- If yes, is their validity satisfactory given
  - Study aims
  - Variable roles

If no, is a bespoke validation study possible (resources)?

If resources are limited, what is the best solution?

AARHUS
UNIVERSITY