Precision Medicine and Big Data: Reconciling the Differences

Howard Bauchner, MD
Editor-in-Chief, JAMA and The JAMA Network
Senior Vice President, Scientific Publications
Professor of Pediatrics and Public Health
Boston University School of Medicine (on leave)

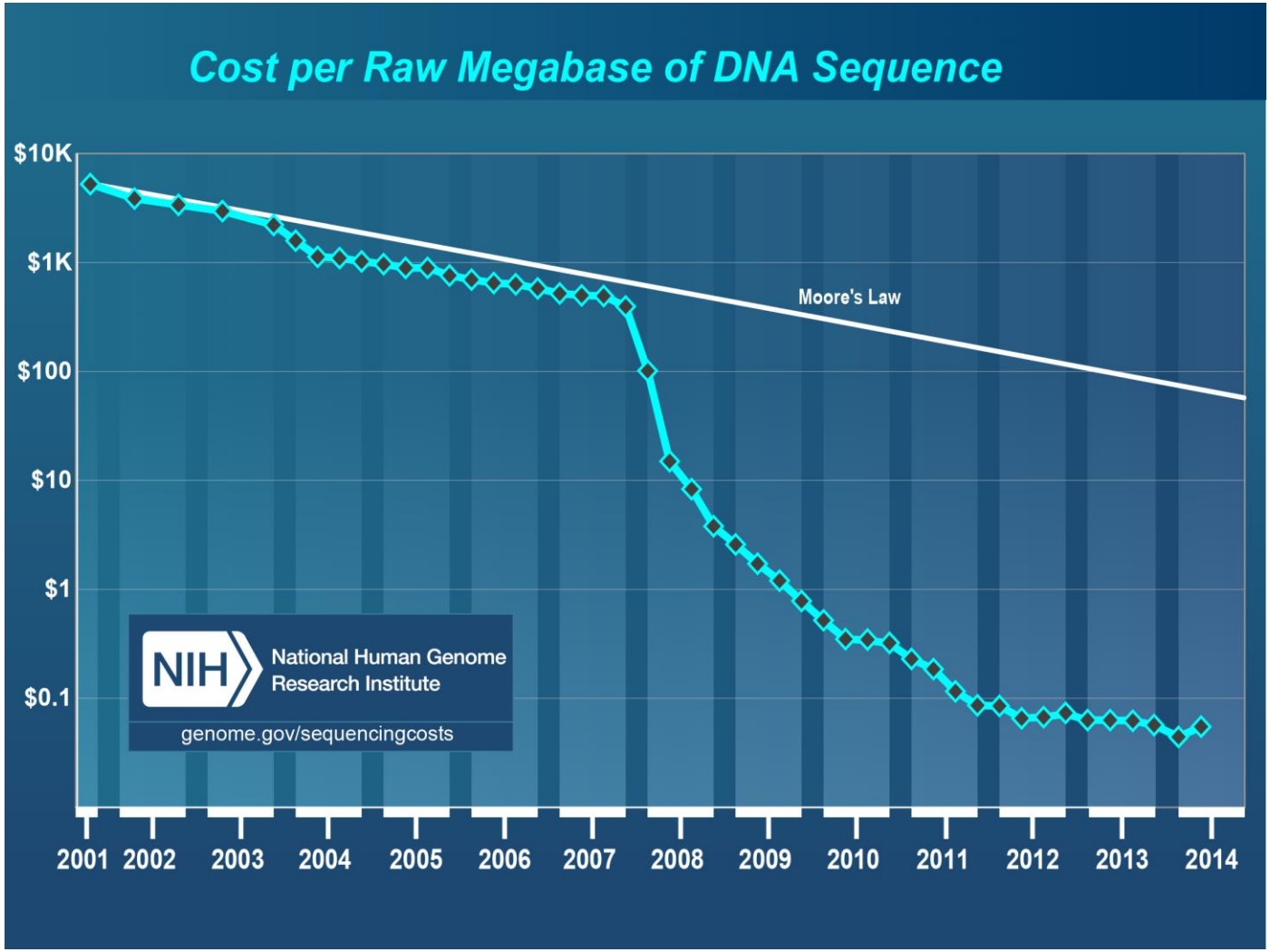
The JAMA Network

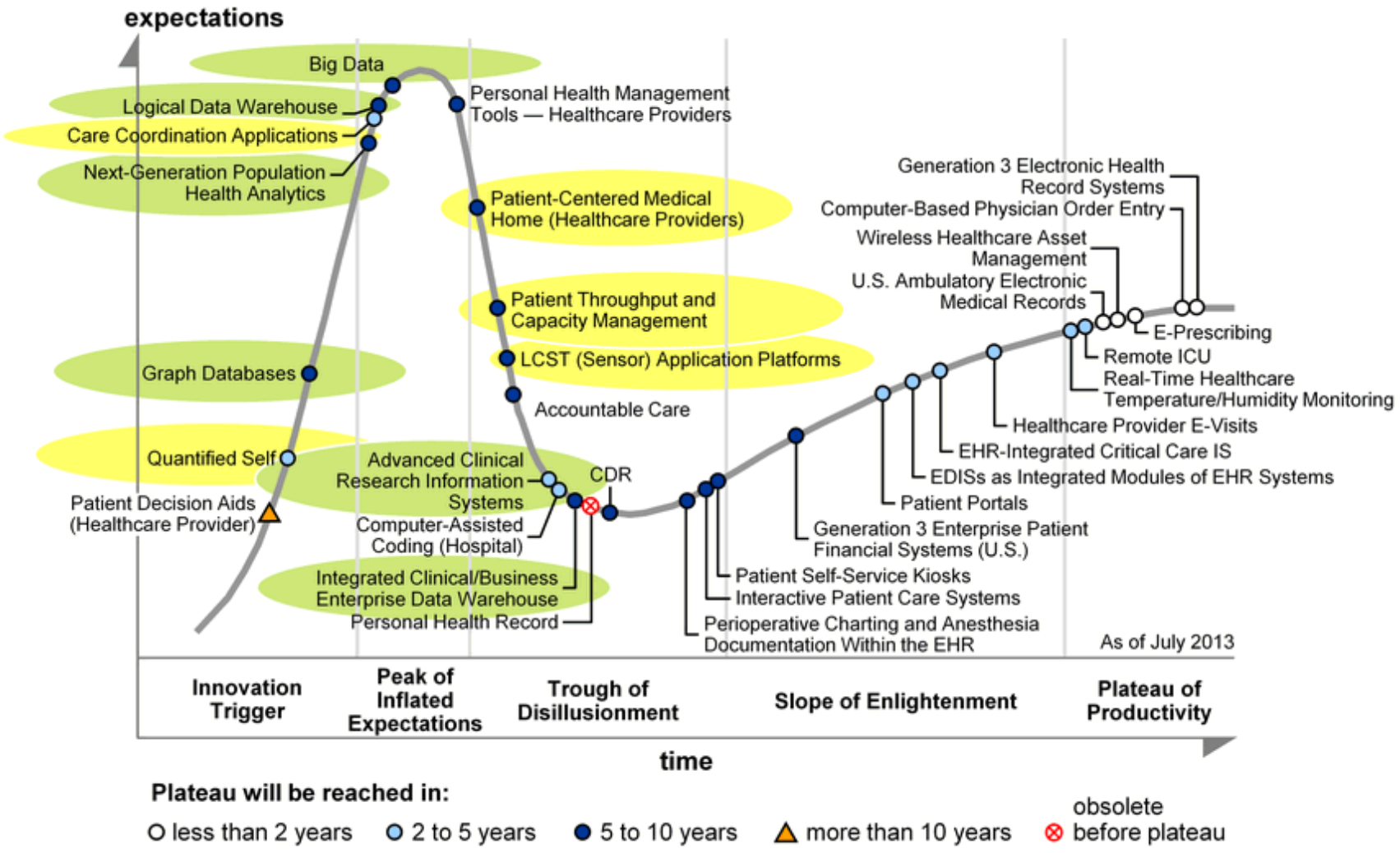
JAMA
THE JOURNAL OF THE AMERICAN MEDICAL ASSOCIATION

Talking Points

- Precision Medicine
  - Personalized – older term – treatments developed for an individual
  - Precision – focus on which approaches, genetic, environmental, will be more effective for individuals
- Big Data
- Observations

The President's Council of Advisors on Science and Technology noted that personalized medicine "refers to the tailoring of medical treatment to the individual characteristics of each patient. It does not literally mean the creation of drugs or medical devices that are unique to a patient, but rather the ability to classify individuals into subpopulations that differ in their susceptibility to a particular disease or their response to a specific treatment. Preventive or therapeutic interventions can then be concentrated on those who will benefit, sparing expense and side effects for those who will not."

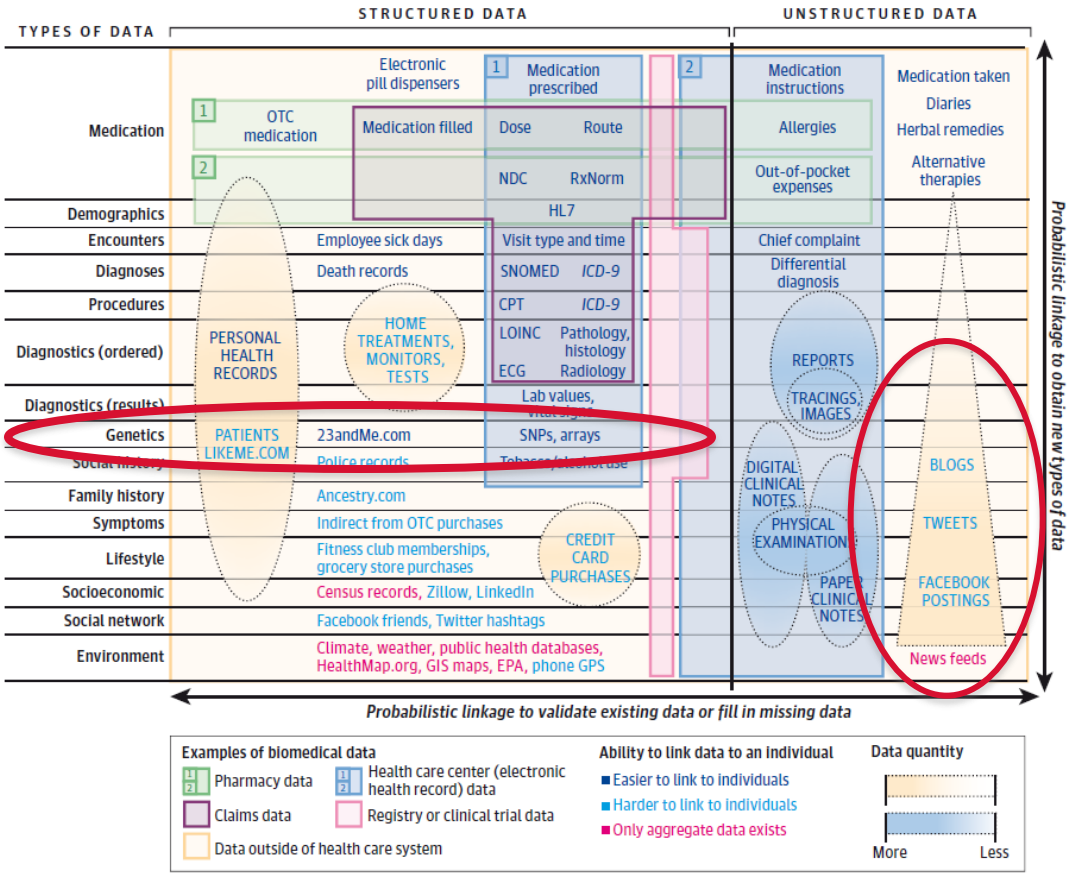




What is Big Data?

The term 'Big Data' is meant to capture the opportunities and challenges facing all biomedical researchers in accessing, managing, analyzing, and integrating datasets of diverse data types [e.g., imaging, phenotypic, molecular (including various '–omics'), exposure, health, behavioral, and the many other types of biological and biomedical and behavioral data] that are increasingly larger, more diverse, and more complex, and that exceed the abilities of currently used approaches to manage and analyze effectively. Big Data emanate from three sources: (1) a small number of groups that produce very large amounts of data, usually as part of projects specifically funded to produce important resources for use by the entire research community; (2) individual investigators who produce large datasets, often empowered by the use of readily available new technologies; and (3) an even greater number of sources that each produce small datasets (e.g. research data or clinical data in electronic health records) whose value can be amplified by aggregating or integrating them with other data.

Figure. The Tapestry of Potentially High-Value Information Sources That May be Linked to an Individual for Use in Health Care



CPT indicates current procedural terminology; ECG, electrocardiography; EPA, US Environmental Protection Agency; GIS, geographic information systems; GPS, global positioning system; HL7, Health Level 7 coding standard; *ICD-9*, *Institutional Classification of Diseases, Ninth Revision*; LOINC, Logical Observation Identifiers Names and Codes; NDC, National Drug Code; OTC, over-the-counter; SNOMED, Systematized Nomenclature of Medicine; SNP, single-nucleotide polymorphism.

The President's Council of Advisors on Science and Technology noted that personalized medicine "refers to the tailoring of medical treatment to the individual characteristics of each patient. It does not literally mean the creation of drugs or medical devices that are unique to a patient, but rather the ability to classify individuals into subpopulations that differ in their susceptibility to a particular disease or their response to a specific treatment. Preventive or therapeutic interventions can then be concentrated on those who will benefit, sparing expense and side effects for those who will not."

Original Investigation

Genetic Variants Associated With Phenytoin-Related Severe Cutaneous Adverse Reactions

Wen-Hung Chung, MD, PhD; Wan-Chun Chang, MS; Yun-Shien Lee, PhD; Ying-Ying Wu, MS; Chih-Hsun Yang, MD; Hsin-Chun Ho, MD; Ming-Jing Chen, MD; Jing-Yi Lin, MD; Rosaline Chung-Yee Hui, MD, PhD; Ji-Chen Ho, MD; Wei-Ming Wu, MD, PhD; Ting-Jui Chen, MD; Tony Wu, MD, PhD; Yih-Ru Wu, MD, PhD; Mo-Song Hsih, MD; Po-Hsun Tu, MD; Chen-Nen Chang, MD, PhD; Chien-Ning Hsu, PhD; Tsu-Lan Wu, PhD; Siew-Eng Choon, MD; Chao-Kai Hsu, MD, PhD; Der-Yuan Chen, MD, PhD; Chin-San Liu, MD, PhD; Ching-Yuang Lin, MD, PhD; Nahoko Kaniwa, PhD; Yoshiro Saito, PhD; Yukitoshi Takahashi, MD, PhD; Ryosuke Nakamura, PhD; Hiroaki Azukizawa, MD, PhD; Yongyong Shi, PhD; Tzu-Hao Wang, MD, PhD; Shiow-Shuh Chuang, MD, PhD; Shih-Feng Tsai, MD, PhD; Chee-Jen Chang, PhD; Yu-Sun Chang, PhD; Shuen-Iu Hung, PhD; for the Taiwan Severe Cutaneous Adverse Reaction Consortium and the Japan Pharmacogenomics Data Science Consortium

IMPORTANCE The antiepileptic drug phenytoin can cause cutaneous adverse reactions, ranging from maculopapular exanthema to severe cutaneous adverse reactions, which include drug reactions with eosinophilia and systemic symptoms, Stevens-Johnson syndrome, and toxic epidermal necrolysis. The pharmacogenomic basis of phenytoin-related severe cutaneous adverse reactions remains unknown.

OBJECTIVE To investigate the genetic factors associated with phenytoin-related severe cutaneous adverse reactions.

DESIGN, SETTING, AND PARTICIPANTS Case-control study conducted in 2002-2014 among 105 cases with phenytoin-related severe cutaneous adverse reactions (n=61 Stevens-Johnson syndrome/toxic epidermal necrolysis and n=44 drug reactions with eosinophilia and systemic symptoms), 78 cases with maculopapular exanthema, 130 phenytoin-tolerant control participants, and 3655 population controls from Taiwan, Japan, and Malaysia. A genome-wide association study (GWAS), direct sequencing of the associated loci, and replication analysis were conducted using the samples from Taiwan. The initial GWAS included samples of 60 cases with phenytoin-related severe cutaneous adverse reactions and 412 population controls from Taiwan. The results were validated in (1) 30 cases with severe cutaneous adverse reactions and 130 phenytoin-tolerant controls from Taiwan, (2) 9 patients with Stevens-Johnson syndrome/toxic epidermal necrolysis and 2869 population controls from Japan, and (3) 6 cases and 374 population controls from Malaysia.

MAIN OUTCOMES AND MEASURES Specific genetic factors associated with phenytoin-related severe cutaneous adverse reactions.

RESULTS The GWAS discovered a cluster of 16 single-nucleotide polymorphisms in *CYP2C* genes at 10q23.33 that reached genome-wide significance. Direct sequencing of *CYP2C* identified missense variant rs1057910 (*CYP2C9*3*) that showed significant association with phenytoin-related severe cutaneous adverse reactions (odds ratio, 12; 95% CI, 6.6-20; P=1.1 × 10^{-17}). The statistically significant association between *CYP2C9*3* and phenytoin-related severe cutaneous adverse reactions was observed in additional samples from Taiwan, Japan, and Malaysia. A meta-analysis using the data from the 3 populations showed an overall odds ratio of 11 (95% CI, 6.2-18; z=8.58; P < .00001) for *CYP2C9*3* association with phenytoin-related severe cutaneous adverse reactions. Delayed clearance of plasma phenytoin was detected in patients with severe cutaneous adverse reactions, especially *CYP2C9*3* carriers, providing a functional link of the associated variants to the disease.

CONCLUSIONS AND RELEVANCE This study identified *CYP2C* variants, including *CYP2C9*3*, known to reduce drug clearance, as important genetic factors associated with phenytoin-related severe cutaneous adverse reactions.

JAMA. 2014;312(5):525-534. doi:10.1001/jama.2014.7859

Supplemental content at jama.com

Author Affiliations: Author affiliations are listed at the end of this article.

Group Information: The Taiwan Severe Cutaneous Adverse Reactions Consortium comprises the physicians, pharmacists, and other investigators from Taipei Medical University Hospital, Wan Fang Hospital, Chung Shan University Hospital, Taichung Veterans General Hospital, Changhua Christian Hospital, China Medicine University Hospital, National Cheng Kung University Hospital, National Taiwan University Hospital, and Chang Gung Memorial Hospitals.

Corresponding Authors: Wen-Hung Chung, MD, PhD, Department of Dermatology, Chang Gung Memorial Hospital, No. 5, Fusing St, Taoyuan, 333 Taiwan (chung1@cgmh.org.tw); Shuen-Iu Hung, PhD, Institute of Pharmacology, School of Medicine, National Yang-Ming University, No. 155, Section 2, Linong St, Taipei, 112 Taiwan (sihung@ym.edu.tw).

Original Investigation

Using Multiplexed Assays of Oncogenic Drivers in Lung Cancers to Select Targeted Drugs

Mark G. Kris, MD; Bruce E. Johnson, MD; Lynne D. Berry, PhD; David J. Kwiatkowski, MD; A. John Iafrate, MD; Ignacio I. Wistuba, MD; Marileila Varella-Garcia, PhD; Wilbur A. Franklin, MD; Samuel L. Aronson, ALM, MA; Pei-Fang Su, PhD; Yu Shyr, PhD; D. Ross Camidge, MD, PhD; Lecia V. Sequist, MD; Bonnie S. Glisson, MD; Fadlo R. Khuri, MD; Edward B. Garon, MD; William Pao, MD, PhD; Charles Rudin, MD, PhD; Joan Schiller, MD; Eric B. Haura, MD; Mark Socinski, MD; Keisuke Shirai, MD; Heidi Chen, PhD; Giuseppe Giaccone, MD; Marc Ladanyi, MD; Kelly Kugler, BA; John D. Minna, MD; Paul A. Bunn, MD

IMPORTANCE Targeting oncogenic drivers (genomic alterations critical to cancer development and maintenance) has transformed the care of patients with lung adenocarcinomas. The Lung Cancer Mutation Consortium was formed to perform multiplexed assays testing adenocarcinomas of the lung for drivers in 10 genes to enable clinicians to select targeted treatments and enroll patients into clinical trials.

OBJECTIVES To determine the frequency of oncogenic drivers in patients with lung adenocarcinomas and to use the data to select treatments targeting the identified driver(s) and measure survival.

DESIGN, SETTING, AND PARTICIPANTS From 2009 through 2012, 14 sites in the United States enrolled patients with metastatic lung adenocarcinomas and a performance status of 0 through 2 and tested their tumors for 10 drivers. Information was collected on patients, therapies, and survival.

INTERVENTIONS Tumors were tested for 10 oncogenic drivers, and results were used to select matched targeted therapies.

MAIN OUTCOMES AND MEASURES Determination of the frequency of oncogenic drivers, the proportion of patients treated with genotype-directed therapy, and survival.

RESULTS From 2009 through 2012, tumors from 1007 patients were tested for at least 1 gene and 733 for 10 genes (patients with full genotyping). An oncogenic driver was found in 466 of 733 patients (64%). Among these 733 tumors, 182 tumors (25%) had the *KRAS* driver; sensitizing *EGFR*, 122 (17%); *ALK* rearrangements, 57 (8%); other *EGFR*, 29 (4%); 2 or more genes, 24 (3%); *ERBB2* (formerly *HER2*), 19 (3%); *BRAF*, 16 (2%); *PIK3CA*, 6 (<1%); *MET* amplification, 5 (<1%); *NRAS*, 5 (<1%); *MEK1*, 1 (<1%); *AKT1*, 0. Results were used to select a targeted therapy or trial in 275 of 1007 patients (28%). The median survival was 3.5 years (interquartile range [IQR], 1.96-7.70) for the 260 patients with an oncogenic driver and genotype-directed therapy compared with 2.4 years (IQR, 0.88-6.20) for the 318 patients with any oncogenic driver(s) who did not receive genotype-directed therapy (propensity score–adjusted hazard ratio, 0.69 [95% CI, 0.53-0.9], $P = .006$).

CONCLUSIONS AND RELEVANCE Actionable drivers were detected in 64% of lung adenocarcinomas. Multiplexed testing aided physicians in selecting therapies. Although individuals with drivers receiving a matched targeted agent lived longer, randomized trials are required to determine if targeting therapy based on oncogenic drivers improves survival.

TRIAL REGISTRATION clinicaltrials.gov Identifier: NCT01014286

JAMA. 2014;311(19):1998-2006. doi:10.1001/jama.2014.3741

Editorial page 1975

Author Video Interview at jama.com

Supplemental content at jama.com

CME Quiz at jamanetworkcme.com and CME Questions page 2019

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Mark G. Kris, MD, Memorial Sloan Kettering Cancer Center, 300 E 66th St, New York, NY 10065 (krism@mskcc.org).

Original Investigation

Molecular Findings Among Patients Referred for Clinical Whole-Exome Sequencing

Yaping Yang, PhD; Donna M. Muzny, MS; Fan Xia, PhD; Zhiyv Niu, PhD; Richard Person, PhD; Yan Ding, PhD; Patricia Ward, MS; Alicia Braxton, MS; Min Wang, PhD; Christian Buhay, BS; Narayanan Veeraraghavan, PhD; Alicia Hawes, BS; Theodore Chiang, PhD; Magalie Leduc, PhD; Joke Beuten, PhD; Jing Zhang, PhD; Weimin He, PhD; Jennifer Scull, PhD; Alecia Willis, PhD; Megan Landsverk, PhD; William J. Craigen, MD, PhD; Mir Reza Bekheirnia, MD; Asbjorg Stray-Pedersen, MD, PhD; Pengfei Liu, PhD; Shu Wen, PhD; Wendy Alcaraz, PhD; Hong Cui, PhD; Magdalena Walkiewicz, PhD; Jeffrey Reid, PhD; Matthew Bainbridge, PhD; Ankita Patel, PhD; Eric Boerwinkle, PhD; Arthur L. Beaudet, MD; James R. Lupski, MD, PhD; Sharon E. Plon, MD, PhD; Richard A. Gibbs, PhD; Christine M. Eng, MD

IMPORTANCE Clinical whole-exome sequencing is increasingly used for diagnostic evaluation of patients with suspected genetic disorders.

OBJECTIVE To perform clinical whole-exome sequencing and report (1) the rate of molecular diagnosis among phenotypic groups, (2) the spectrum of genetic alterations contributing to disease, and (3) the prevalence of medically actionable incidental findings such as *FBN1* mutations causing Marfan syndrome.

DESIGN, SETTING, AND PATIENTS Observational study of 2000 consecutive patients with clinical whole-exome sequencing analyzed between June 2012 and August 2014. Whole-exome sequencing tests were performed at a clinical genetics laboratory in the United States (laboratory accredited by the College of American Pathologists and the US Centers for Disease Control and Prevention Clinical Laboratory Improvement Amendments of 1988). Results were reported by clinical molecular geneticists certified by the American Board of Medical Genetics. Tests were ordered by the patient's physician. The patients were primarily pediatric (1756 [88%]; mean age, 6 years; 888 females [44%], 1101 males [55%], and 11 fetuses [1% gender unknown]), demonstrating diverse clinical manifestations most often including nervous system dysfunction such as developmental delay.

MAIN OUTCOMES AND MEASURES Whole-exome sequencing diagnosis rate overall and by phenotypic category, mode of inheritance, spectrum of genetic events, and reporting of incidental findings.

RESULTS A molecular diagnosis was reported for 504 patients (25.2%) with 58% of the diagnostic mutations not previously reported. Molecular diagnosis rates for each phenotypic category were 143/526 (27.2%; 95% CI, 23.5%-31.2%) for the neurological group, 282/1147 (24.6%; 95% CI, 22.1%-27.2%) for the neurological plus other organ systems group, 30/83 (36.1%; 95% CI, 26.1%-47.5%) for the specific neurological group, and 49/244 (20.1%; 95% CI, 15.6%-25.8%) for the nonneurological group. The molecular diagnoses rendered 527 Mendelian disease patterns, which included 280 (53.1%) autosomal dominant, 181 (34.3%) autosomal recessive (including 5 with uniparental disomy), 65 (12.3%) X-linked, and 1 (0.2%) mitochondrial. Of 504 patients with a molecular diagnosis, 23 (4.6%) had blended phenotypes resulting from 2 single gene defects. About 30% of the positive cases harbored mutations in disease genes reported since 2011. There were 95 medically actionable incidental findings in genes unrelated to the phenotype but with immediate implications for management in 92 patients (4.6%), including 59 patients (3%) with mutations in genes recommended for reporting by the American College of Medical Genetics and Genomics.

CONCLUSIONS AND RELEVANCE Whole-exome sequencing provided a potential molecular diagnosis for 25% of a large cohort of patients referred for evaluation of suspected genetic conditions, including detection of rare genetic events and new mutations contributing to disease. The observed flexibility and yield of whole-exome sequencing may offer advantages over traditional molecular diagnosis approaches in certain patients.

JAMA. doi:10.1001/jama.2014.14601
Published online October 18, 2014.

Editorial
Related article
Supplemental content at jama.com
CME Quiz at jamanetworkcme.com and CME Questions

Author Affiliations: Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas (Yang, Xia, Niu, Person, Ward, Braxton, Leduc, Beuten, Zhang, He, Scull, Willis, Landsverk, Craigen, Bekheirnia, Liu, Wen, Alcaraz, Cui, Walkiewicz, Patel, Beaudet, Lupski, Plon, Gibbs, Eng); Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas (Muzny, Ding, Wang, Buhay, Veeraraghavan, Hawes, Chiang, Reid, Bainbridge, Boerwinkle, Lupski, Gibbs); Department of Pediatrics, Baylor College of Medicine, Houston, Texas (Craigen, Stray-Pedersen, Lupski, Plon); Human Genetics Center, University of Texas Health Science Center, Houston (Boerwinkle).

Corresponding Author: Christine M. Eng, MD, Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030 (ceng@bcm.edu)

Diagnostic Yield <5 years
Neurological – 30.4%
Neurological + other – 26.1%
Nonneurological – 24.4%

Some Findings
BRCA 2 – 9
BRCA 1 – 5
AF/Long QT – 5
Marfan/MVP - 4

Original Investigation

Association of Arrhythmia-Related Genetic Variants With Phenotypes Documented in Electronic Medical Records

Sara L. Van Driest, MD, PhD; Quinn S. Wells, MD, PharmD, MSCI; Sarah Stallings, PhD; William S. Bush, PhD, MS; Adam Gordon, PhD; Deborah A. Nickerson, PhD; Jerry H. Kim, MD; David R. Crosslin, PhD; Gail P. Jarvik, MD, PhD; David S. Carrell, PhD; James D. Ralston, MD, MPH; Eric B. Larson, MD, MPH; Suzette J. Bielinski, PhD; Janet E. Olson, PhD; Zi Ye, MD, PhD; Iftikhar J. Kullo, MD; Noura S. Abul-Husn, MD, PhD; Stuart A. Scott, PhD; Erwin Bottinger, MD; Berta Almoguera, PhD; John Connolly, PhD; Rosetta Chiavacci, BSN, CCRC; Hakon Hakonarson, MD, PhD; Laura J. Rasmussen-Torvik, PhD, MPH; Vivian Pan, MS, CGC; Stephen D. Persell, MD, MPH; Maureen Smith, MS, CGC; Rex L. Chisholm, PhD; Terrie E. Kitchner, CCRP; Max M. He, PhD; Murray H. Brilliant, PhD; John R. Wallace, MS; Kimberly F. Doheny, PhD; M. Benjamin Shoemaker, MD, MCSI; Rongling Li, MD, PhD, MPH; Teri A. Manolio, MD, PhD; Thomas E. Callis, PhD; Daniela Macaya, MQC; Marc S. Williams, MD; David Carey, PhD; Jamie D. Kapplinger, BA; Michael J. Ackerman, MD, PhD; Marylyn D. Ritchie, PhD; Joshua C. Denny, MD, MS; Dan M. Roden, MD

IMPORTANCE Large-scale DNA sequencing identifies incidental rare variants in established Mendelian disease genes, but the frequency of related clinical phenotypes in unselected patient populations is not well established. Phenotype data from electronic medical records (EMRs) may provide a resource to assess the clinical relevance of rare variants.

OBJECTIVE To determine the clinical phenotypes from EMRs for individuals with variants designated as pathogenic by expert review in arrhythmia susceptibility genes.

DESIGN, SETTING, AND PARTICIPANTS This prospective cohort study included 2022 individuals recruited for nonantiarrhythmic drug exposure phenotypes from October 5, 2012, to September 30, 2013, for the Electronic Medical Records and Genomics Network Pharmacogenomics project from 7 US academic medical centers. Variants in *SCN5A* and *KCNH2*, disease genes for long QT and Brugada syndromes, were assessed for potential pathogenicity by 3 laboratories with ion channel expertise and by comparison with the ClinVar database. Relevant phenotypes were determined from EMRs, with data available from 2002 (or earlier for some sites) through September 10, 2014.

EXPOSURES One or more variants designated as pathogenic in *SCN5A* or *KCNH2*.

MAIN OUTCOMES AND MEASURES Arrhythmia or electrocardiographic (ECG) phenotypes defined by *International Classification of Diseases, Ninth Revision (ICD-9)* codes, ECG data, and manual EMR review.

RESULTS Among 2022 study participants (median age, 61 years [interquartile range, 56-65 years]; 1118 [55%] female; 1491 [74%] white), a total of 122 rare (minor allele frequency <0.5%) nonsynonymous and splice-site variants in 2 arrhythmia susceptibility genes were identified in 223 individuals (11% of the study cohort). Forty-two variants in 63 participants were designated potentially pathogenic by at least 1 laboratory or ClinVar, with low concordance across laboratories (Cohen κ = 0.26). An *ICD-9* code for arrhythmia was found in 11 of 63 (17%) variant carriers vs 264 of 1959 (13%) of those without variants (difference, +4%; 95% CI, −5% to +13%; $P = .35$). In the 1270 (63%) with ECGs, corrected QT intervals were not different in variant carriers vs those without (median, 429 vs 439 milliseconds; difference, −10 milliseconds; 95% CI, −16 to +3 milliseconds; $P = .17$). After manual review, 22 of 63 participants (35%) with designated variants had any ECG or arrhythmia phenotype, and only 2 had corrected QT interval longer than 500 milliseconds.

CONCLUSIONS AND RELEVANCE Among laboratories experienced in genetic testing for cardiac arrhythmia disorders, there was low concordance in designating *SCN5A* and *KCNH2* variants as pathogenic. In an unselected population, the putatively pathogenic genetic variants were not associated with an abnormal phenotype. These findings raise questions about the implications of notifying patients of incidental genetic findings.

JAMA. 2016;315(1):47-57. doi:10.1001/jama.2015.17701

Editorial page 33

Supplemental content at jama.com

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Dan M. Roden, MD, Vanderbilt University Medical Center, 2215B Garland Ave, Room 1285, Nashville, TN 37232-0575 (dan.roden@vanderbilt.edu).

VIEWPOINT

Seven Questions for Personalized Medicine

Michael J. Joyner, MD
Department of Anesthesiology, Mayo Clinic, Rochester, Minnesota.

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

Corresponding Author: Michael J. Joyner, MD, Department of Anesthesiology, Mayo Clinic, 200 First Street SW, Rochester, MN 55905 (joyner.michael@mayo.edu).

Personalized or precision medicine maintains that medical care and public health will be radically transformed by prevention and treatment programs more closely targeted to the individual patient. These interventions will be developed by sequencing more genomes, creating bigger biobanks, and linking biological information to health data in electronic medical records (EMRs) or obtained by monitoring technologies. Yet the assumptions underpinning personalized medicine have largely escaped questioning. In this Viewpoint, we seek to stimulate a more balanced debate by posing 7 questions for the advocates of personalized medicine.

Does the Human Genome Contribute to Disease Risk Prediction?

Personalized medicine builds on the Human Genome Project, which was forecasted to revolutionize disease risk prediction, with projected relative risks as high as 6 for gene variants linked to specific diseases. However, the relative risks for the vast majority of gene variants rarely exceed 1.5, and these variants have added little useful predictive power to traditional risk prediction algorithms. Moreover, improved adherence with lifestyle interventions expected to result from the provision of genomic risk information to patients has not materialized.[1]

Will Gene-Based Drug Targeting and Development Fulfill Its Promise?

Personalized medicine predicts that therapies for cancer that target dysregulated "-omic" pathways will be transformative. Yet the benefit of such drugs on overall cancer survival has been limited, perhaps because of the adaptive nature of cancer. There is little evidence that targeted therapy will interrupt the cycle of expectation and disappointment that has typified many of the new approaches to cancer therapy. Most of the recent successes in cancer have resulted from the traditional public health measures of screening, early detection, and smoking reduction as well as some immunologic therapies.

For common disorders, the claim that genotype-based treatment schemes will be more effective with fewer adverse effects is not supported by negative findings for both tamoxifen[2] and warfarin.[3] Even though gene variant information can suggest new therapeutic targets, it will always have to be integrated into traditional drug discovery approaches.

Two much-publicized successes in disease gene identification were *BRCA1/2* for breast or ovarian cancer and mutations for cystic fibrosis (CF). Although finding a subgroup of the population that is at very high risk of cancer is important, no new therapy has resulted from discovery of the mutations. Instead, the 5% of patients with breast or ovarian cancer who are positive for *BRCA* are offered enhanced screening and preemptive surgery. In the 25 years since *BRCA1/2* was discovered, breast cancer mortality in the United States has declined by nearly one-third; however, little of this decline stems from the discovery of *BRCA1/2*. Moreover, *BRCA1/2* is a unique story because the gene variants account for such a substantial amount of the variance in outcome for a limited number of patients.

In CF, 2 drugs (ivacaftor and lumacaftor) have recently been developed based on the CF transmembrane conductance regulator gene (*CFTR*), but they are useful only in patients with specific *CFTR* mutations, in whom they increase, singly or in combination, maximum forced expiratory volume (FEV) by 5% to 10% and improve weight gain.[4] However, since the discovery of this gene in the 1980s, CF survival has improved substantially as a result of strict adherence to clinical management guidelines originating in pulmonary physiology and infectious diseases, but not genomics.

Although well-deserved recognition has accompanied these genetic discoveries, neither has been a significant factor in the substantial reduction in mortality from the 2 target diseases during the past 25 years. The commitment to screening technology and adherence to best practices has proven far more important to the lives of affected patients.

What Will EMRs Contribute?

The transition to EMRs has expanded the reach of medical record–based information, but has not markedly improved the quality of the data entered. Although examples of improved clinical practice driven by EMRs can be found, the quality and granularity of the data they record limit their use in research. The inherent variability of clinical data across institutions is magnified by institution-to-institution differences in EMR systems. A seamless, interoperable national EMR system is, at best, decades away for the United States and unlikely to include informative phenotypic data such as waist circumference, musculoskeletal fitness, and exercise capacity.

What Kinds of Studies Should Be Mounted in Personalized Medicine?

In recent years, terms such as *unsupervised*, *agnostic*, *discovery*, and *data mining* have been used to describe an approach to big data that proceeds without explicit hypotheses, with conclusions derived from the *P* values of discovered associations. Convenience samples are often used without an appreciation of how selection bias and other factors can distort exposure-outcome relationships. Much so-called discovery science presupposes that the individual is isolated from his or her social context and that cellular data are sufficient to predict disease. By contrast, successful population-based approaches to the study of disease, such as the

Precision Medicine

- Remains focused on genes
  - Pharmacogenomics
  - Genetics (biology) of cancer
- Potential advances in nutritional research
- Genes that suggest behavior change will be less valuable
- Will it be used to improve the effectiveness of an already successful drug or used to create success of a failed drug
- Ultimately are clinical trials necessary to assess the effectiveness of personalized medicine or are observational data sufficient
- Does cost and value matter – clinical (as determined by patients) vs statistical significance vs cost to society – more knowledge may no longer be sufficient

What is Big Data?

The term 'Big Data' is meant to capture the opportunities and challenges facing all biomedical researchers in accessing, managing, analyzing, and integrating datasets of diverse data types [e.g., imaging, phenotypic, molecular (including various '–omics'), exposure, health, behavioral, and the many other types of biological and biomedical and behavioral data] that are increasingly larger, more diverse, and more complex, and that exceed the abilities of currently used approaches to manage and analyze effectively. Big Data emanate from three sources: (1) a small number of groups that produce very large amounts of data, usually as part of projects specifically funded to produce important resources for use by the entire research community; (2) individual investigators who produce large datasets, often empowered by the use of readily available new technologies; and (3) an even greater number of sources that each produce small datasets (e.g. research data or clinical data in electronic health records) whose value can be amplified by aggregating or integrating them with other data.

Big Data

- Few examples because of lack of clear definition
- NIH now “vested” in BIG data
- Potentially unrealized value but data are siloed
- Utilization is largely observational (exception is Mendelian randomization, JAMA, Smith et al, 2014) so causality limited
- Recall the inability to highlight travel history in US, sadly EHRs are static and not dynamic systems

Mission Statement

The ability to harvest the wealth of information contained in biomedical Big Data will advance our understanding of human health and disease; however, lack of appropriate tools, poor data accessibility, and insufficient training, are major impediments to rapid translational impact. To meet this challenge, the National Institutes of Health (NIH) launched the Big Data to Knowledge (BD2K) initiative in 2012.

BD2K is a trans-NIH initiative established to enable biomedical research as a digital research enterprise, to facilitate discovery and support new knowledge, and to maximize community engagement.

The BD2K initiative addresses four major aims that, in combination, are meant to enhance the utility of biomedical Big Data:

To facilitate broad use of biomedical digital assets by making them discoverable, accessible, and citable.
To conduct research and develop the methods, software, and tools needed to analyze biomedical Big Data.
To enhance training in the development and use of methods and tools necessary for biomedical Big Data science.
To support a data ecosystem that accelerates discovery as part of a digital enterprise.

Patient-Centered Information Commons
Harvard University Medical School
PI: Isaac S. Kohane
Grant Number: 1U54HG007963-01

Investigators at the Patient-Centered Information Commons will develop systems to combine genetic, environmental, imaging, behavioral, and clinical data on individual patients from multiple sources into integrated sets. Computing across thousands of such individuals, will enable more accurate classification of individual disease or disease risk, and facilitate greater precision in patient disease prevention and treatment strategies.
- See more at:
http://bd2k.nih.gov/FY14/COE/COE.html#sthash.uJ9E205X.dpuf

Original Investigation

The Familial Risk of Autism

Sven Sandin, MSc; Paul Lichtenstein, PhD; Ralf Kuja-Halkola, MSc; Henrik Larsson, PhD; Christina M. Hultman, PhD; Abraham Reichenberg, PhD

IMPORTANCE Autism spectrum disorder (ASD) aggregates in families, but the individual risk and to what extent this is caused by genetic factors or shared or nonshared environmental factors remains unresolved.

OBJECTIVE To provide estimates of familial aggregation and heritability of ASD.

DESIGN, SETTING, AND PARTICIPANTS A population-based cohort including 2 049 973 Swedish children born 1982 through 2006. We identified 37 570 twin pairs, 2 642 064 full sibling pairs, 432 281 maternal and 445 531 paternal half sibling pairs, and 5 799 875 cousin pairs. Diagnoses of ASD to December 31, 2009 were ascertained.

MAIN OUTCOMES AND MEASURES The relative recurrence risk (RRR) measures familial aggregation of disease. The RRR is the relative risk of autism in a participant with a sibling or cousin who has the diagnosis (exposed) compared with the risk in a participant with no diagnosed family member (unexposed). We calculated RRR for both ASD and autistic disorder adjusting for age, birth year, sex, parental psychiatric history, and parental age. We estimated how much of the probability of developing ASD can be related to genetic (additive and dominant) and environmental (shared and nonshared) factors.

RESULTS In the sample, 14 516 children were diagnosed with ASD, of whom 5689 had autistic disorder. The RRR and rate per 100 000 person-years for ASD among monozygotic twins was estimated to be 153.0 (95% CI, 56.7-412.8; rate, 6274 for exposed vs 27 for unexposed); for dizygotic twins, 8.2 (95% CI, 3.7-18.1; rate, 805 for exposed vs 55 for unexposed); for full siblings, 10.3 (95% CI, 9.4-11.3; rate, 829 for exposed vs 49 for unexposed); for maternal half siblings, 3.3 (95% CI, 2.6-4.2; rate, 492 for exposed vs 94 for unexposed); for paternal half siblings, 2.9 (95% CI, 2.2-3.7; rate, 371 for exposed vs 85 for unexposed); and for cousins, 2.0 (95% CI, 1.8-2.2; rate, 155 for exposed vs 49 for unexposed). The RRR pattern was similar for autistic disorder but of slightly higher magnitude. We found support for a disease etiology including only additive genetic and nonshared environmental effects. The ASD heritability was estimated to be 0.50 (95% CI, 0.45-0.56) and the autistic disorder heritability was estimated to 0.54 (95% CI, 0.44-0.64).

CONCLUSIONS AND RELEVANCE Among children born in Sweden, the individual risk of ASD and autistic disorder increased with increasing genetic relatedness. Heritability of ASD and autistic disorder were estimated to be approximately 50%. These findings may inform the counseling of families with affected children.

Editorial page 1738

Author Audio Interview at jama.com

Supplemental content at jama.com

Author Affiliations: Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Sweden (Sandin, Lichtenstein, Kuja-Halkola, Larsson, Hultman); Department of Psychosis Studies, Institute of Psychiatry, King's College London, United Kingdom (Sandin);

Autism spectrum disorder (ASD) affects almost 1% of all children born in the United States[1] and is defined as impairment in social interaction and communication and the presence of restricted interests and repetitive behaviors. Autistic disorder is the most profound form of ASD.[1]

Family studies found that ASD aggregates in families and early twin studies estimated the proportion of the phenotypic variance due to genetic factors (the heritability) to be about 90%,[2-6] making it the most heritable of all developmental disorders. As a consequence, etiological research in ASD has focused predominantly on genetic factors.[7] Although recent twin studies support high heritability,[5,6] a large twin study[7] indicated a substantial role for shared environmental influences. Results of family studies also raise questions about the relative influence of genetic factors[8] and contribute to uncertainty regarding the etiology of ASD.

Previous studies have limitations. Twin studies often have small samples, limiting the reliability when studying rare diseases such as ASD. None of the previous studies represent a prospective, population-based, random sample, which raises concerns for potential biases introduced by population selection. Restricted follow-up time and possible differences in etiology for different ASD subtypes may also limit reliability.

Although heritability estimates provide a valuable metric for estimating the effects of genetic factors in the population, they do not provide any information on individual risk. Detailed etiological models will require accounting for risk on a population level, as well as providing quantitative information in a given individual, thus allowing for individualized disease prevention and treatment.[9] Recurrence risk expresses the risk of having another affected family member in an already-affected family. The relative recurrence risk (RRR) measures this recurrence in relation to disease in families without any affected members but can be interpreted and compared between groups that may differ in disease prevalence.

Consequently, there is a need for reliable estimates of heritability for ASD, as well as combining these population-based estimates with individual-level risk estimates to provide a more precise and complete picture of the etiology of ASD.

To that goal, we conducted a longitudinal cohort study of all births in Sweden between 1982 and 2006. Using all pairs of monozygotic and dizygotic twins, full siblings, half siblings, and cousins in the population, we determined the family clustering of ASD by estimating RRR within families and assessed the importance of genetic vs environmental factors associated with ASD.

Methods

Study Population

The study was approved by the ethics committee at the Karolinska Institutet, Stockholm, Sweden. Informed consent was waived by the ethics committee. A birth cohort of all children born alive in Sweden between January 1, 1982, and December 31, 2006, was established using data from Swedish national registers including the Swedish Medical Birth Register,[10] the Swedish Multi-generation Register,[11] the National Patient Register,[12-14] the Swedish Twin Registry,[15] and the Statistics Sweden Total Population Register for vital statistics. Single-child families were excluded from the cohort. Twin zygosity was obtained from the Swedish Twin Registry, and was determined by DNA analysis in 86% of same-sex twins. For the remainder, an algorithm based on 5 parent-reported items assessing twin similarity was used. The Swedish Multi-generation Register contains identifiers for the parents of all children born from 1932 onwards. This allowed us to determine family relations (full siblings, maternal and paternal half siblings, and cousins) using the unique identifiers of the parents and grandfathers. Cousins were derived for full siblings only. Further details are provided in eAppendix A in the Supplement. Data are collected routinely by Swedish government agencies and were merged and anonymized by an independent government agency (Statistics Sweden), and the code linking the personal identification numbers to the new case numbers was destroyed immediately after merging.

Ascertainment of Autism and Psychiatric Diagnosis

In Sweden, all infants and preschool children regularly undergo routine medical and developmental examinations. At age 4 years, a mandatory developmental assessment (motor, language, cognitive, and social development) is conducted. Children with suspected developmental disorders are referred for further assessment by a specialized team in a child psychiatry unit or habilitation service. Diagnostic information is reported to the National Patient Register. The register has nearly complete national coverage[12] of psychiatric diagnoses since 1973. With a rare disease, the sensitivity is a smaller problem than the specificity of the diagnostic codes. We relied on previous validation studies of psychiatric codes generally[12,14] and for autism specifically.[16] Prospective follow-up was conducted until December 31, 2009. Autistic disorder was defined by codes from the *International Classification of Diseases, version 9* code 299.A/B/X and *version 10* (*ICD-10*) code F84.0; ASD also included *ICD-10* codes F84.1 (atypical autism), F84.5 (Asperger syndrome), F84.8 (other pervasive developmental disorders), and F84.9 (pervasive developmental disorder, unspecified).

Covariates

We considered several factors that might confound or modify the familial associations. Parental psychiatric history has been associated with autism in the offspring. Parental psychiatric history was classified as present or not present for each parent separately using any psychiatric diagnosis at any time before the birth of the oldest child in a sibling or cousins pair using *ICD 7th-10th revisions* (eTable 1 in the Supplement). We also obtained information on paternal and maternal age at birth of the child, birth year, and sex.

Statistical Methods

Relative Recurrence Risk

The RRR for siblings is the risk of an autism diagnosis in a sibling of a child with autism compared with a sibling of a child without autism. We calculated RRR in families of different genetic relatedness: full siblings, half siblings, and cousins. Cousin

Original Investigation

Association Between Casino Opening or Expansion and Risk of Childhood Overweight and Obesity

Jessica C. Jones-Smith, PhD; William H. Dow, PhD; Kristal Chichlowska, PhD

IMPORTANCE Economic resources have been inversely associated with risk of childhood overweight/obesity. Few studies have evaluated whether this association is a direct effect of economic resources or is attributable to unmeasured confounding or reverse causation. American Indian–owned casinos have resulted in increased economic resources for some tribes and provide an opportunity to test whether these resources are associated with overweight/obesity.

OBJECTIVE To assess whether openings or expansions of American Indian–owned casinos were associated with childhood overweight/obesity risk.

DESIGN, SETTING, AND PARTICIPANTS We used repeated cross-sectional anthropometric measurements from fitness testing of American Indian children (aged 7-18 years) from 117 school districts that encompassed tribal lands in California between 2001 and 2012. Children in school districts encompassing American Indian tribal lands that either gained or expanded a casino were compared with children in districts with tribal lands that did not gain or expand a casino.

MAIN OUTCOMES AND MEASURES Per capita annual income, median annual household income, percentage of population in poverty, total population, child overweight/obesity (body mass index [BMI] ≥85th age- and sex-specific percentile) and BMI *z* score.

RESULTS Of the 117 school districts, 57 gained or expanded a casino, 24 had a preexisting casino but did not expand, and 36 never had a casino. The mean slots per capita was 7 (SD, 12) and the median was 3 (interquartile range [IQR], 0.3-8). Among districts where a casino opened or expanded, the mean change in slots per capita was 13 (SD, 19) and the median was 3 (IQR, 1-11). Forty-eight percent of the anthropometric measurements were classified as overweight/obese (11 048/22 863). Every casino slot machine per capita gained was associated with an increase in per capita annual income (β = $541; 95% CI, $245-$836) and a decrease in percentage in poverty (β = −0.6%; 95% CI, −1.1% to −0.20%) among American Indians living on tribal lands. Among American Indian children, every slot machine per capita gained was associated with a decreased probability of overweight/obesity by 0.19 percentage points (95% CI, −0.26 to −0.11 percentage points) and a decrease in BMI *z* score (β = −0.003; 95% CI, −0.005 to −0.0002).

CONCLUSIONS AND RELEVANCE In this study, opening or expanding a casino was associated with increased economic resources and decreased risk of childhood overweight/obesity. Given the limitations of an ecological study, further research is needed to better understand the mechanisms behind this association.

Editorial page 915

Supplemental content at jama.com

Author Affiliations: Department of International Health (Human Nutrition), Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland (Jones-Smith); School of Public Health, University of California, Berkeley (Dow); independent consultant, Sacramento, California (Chichlowska).

Corresponding Author: Jessica C. Jones-Smith, PhD, Johns Hopkins Bloomberg School of Public Health, 615 N Wolfe St, Room E2545, Baltimore, MD 21218 (jjoness@jhsph.edu).

JAMA. 2014;311(9):929-936. doi:10.1001/jama.2014.604

Precision Medicine & Big Data

- They are fraternal twins
- Much promised – perhaps too much
- Clinical trials remain important
- Data remain siloed within healthcare (Precision Medicine) and from healthcare (Big Data)

We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run

- Amara’ Law (Roy Amara – Institute for the Future)