Big Data, Family Medicine Research and Missing the Quadruple Aim

An Annotated Bibliography

The Robert Graham Center

March 2015

David Killeen, MD Candidate 2019

Elizabeth Wilkinson, BA

Yalda Jabbarpour, MD

INTRODUCTION

In 2014, Health Affairs magazine devoted its July issues to "Big Data" in healthcare. They defined "Big Data" with three Vs: Volume, Velocity and Variety. Since that time, health care information, EHRs and data sets have rapidly grown. Yet, the big data revolution has been slow to reach Family Medicine research, especially in the United States.

Within Family Medicine, big data should be used to help achieve the quadruple aim while staying patient-focused. There are many uses for big data within Family Medicine, including providing information on health equity and population health to give policymakers and public health officials the data they need to for policy changes, and providing information to doctors, clinics and health systems on better disease management and quality improvement. Big data allows predictive learning to aid clinicians in achieving better diagnosis and treatment for patients.

The United Kingdom (UK) was an early adopter of big data in primary care research. Health Data Research UK is a national database available to investigators with goal of making discoveries to improve patients' lives. This database exemplifies the potential for big data on a national scale. Large health datasets have many uses, including discovering disease trends, improving care quality, improving health equity and identifying unnecessary testing and adverse reactions.

There are still many challenges that need to be addressed so big data can be used to achieve these goals and its full potential. Challenges of applying and using big data are mainly centered on the quality of the data, the administrative burden of collection, and the protection of patient privacy.

Uses:

- Population Health/Health Equity
- Disease Management
- Identify Cost Savings
- Predictive Learning Technologies

Challenges:

- Data Quality
- Data collection, unification and access and complying data variations over time
- Privacy

Future Uses:

- Integrating Patient Data, making it "live" with tools such as a pedometers, scales, and self-reported patient data and experiences.
- Reducing Physician Burnout

Big data is not getting any smaller. The amount of information we collect on patients through visits, labs, admissions, prescriptions and other sources will continue to increase. It is up to Family Medicine researchers to develop the policies and tools need to ensure that big data is used to achieve the Quadruple Aim. There must be continued work to improve the quality of the data we use to lower costs, achieve better outcomes, and improve the patient and clinician experience.

PART 1

USES FOR BIG DATA

1. Health Equity (Disease Management)

Global, Regional, National and Subnational Big Data to Inform Health Equity Research: Perspectives from The Global Burden of Disease Study

Mokdad AH, Mensah GA, Krish V, et al. Global, Regional, National, and Subnational Big Data to Inform Health Equity Research: Perspectives from the Global Burden of Disease Study 2017. *Ethn Dis*. 2019;29(Suppl 1):159-172. doi:10.18865/ed.29.S1.159

Theme: Health Equity (Global - Regional)
Journal, Year: Ethnicity and Disease, 2017
Summary: The Global Burden of Disease Study in 2017 provides an assessment of all-cause mortality and estimates for the causes of death, YLD and DALY in 195 countries using a systemic analysis of published studies and data sources to compare between countries and within countries to examine disparities from 1990 to 2017. There was little improvement in the number of disabilities from 1990 to 2017 compared with the improvements observed for deaths. Females had more Disability Adjusted Life Years from mental health and diabetes while males had more DALYs from injuries and HIV/AIDS. South American, South Africa, Philippines and Russia had more deaths from interpersonal violence than would be expected given their sociodemographic information. This data can help healthcare systems and policy makers identify drivers of cost and better target resources.

Association of Primary Care Physician Supply with Population Mortality in the United States, 2005-2015

Basu S, Berkowitz SA, Phillips RL, Bitton A, Landon BE, Phillips RS. Association of Primary Care Physician Supply With Population Mortality in the United States, 2005-2015. *JAMA Intern Med*. 2019;179(4):506-514. doi:10.1001/jamainternmed.2018.7624

Theme: Health Equity (National)
Journal, Year: JAMA Internal Medicine, 2019
Summary: Basu et al. used multiple data sources – AMA Physician Masterfile, National Center for Health Statistics, US Census Bureau and the Human Mortality Database along with other socioeconomic covariates –to analyze the effect of PCP density on population-level mortality. They found that for every increase in 10 PCPs per 100,000 population there was an associated increase of 51.5 days in life expectancy.

Morbidity, mortality and missed appointments in healthcare: a national retrospective data linkage study

McQueenie R, Ellis DA, McConnachie A, Wilson P, Williamson AE. Morbidity, mortality and missed appointments in healthcare: a national retrospective data linkage study. *BMC Medicine*. 2019;17(1):2. doi:10.1186/s12916-018-1234-0

Theme: Disease management (Clinical)
Journal, Year: BMC Medicine, 2019
Summary: McQueenie et al. demonstrated that as the number of long-term conditions a patient had increased, there was an increased risked for missing general practice appointments and an increased risk of all-cause mortality. There were 824,374 patients in the sample using data from the National Health Service and Scottish death records. Patients with four or more long-term conditions were twice as likely to miss appointments compared to those with none. Patients with only mental health-related long-term conditions who had two or more missed appointments per year had an 8-fold increase in all-cause mortality.

Birth outcomes in Colorado's undocumented immigrant population

Reed MM, Westfall JM, Bublitz C, Battaglia C, Fickenscher A. Birth outcomes in Colorado's undocumented immigrant population. *BMC Public Health*. 2005;5:100. doi:10.1186/1471-2458-5-100

Theme: Health Equity (Immigration Status), Quality Improvement, Challenge
Journal, Year: BMC Public Health, 2005
Summary: This retrospective study found that undocumented mothers were younger, less educated, less likely to gain enough weight during pregnancy, less likely to receive early prenatal care, and had higher rates of anemia. Researchers linked emergency Medicaid data with Colorado Vital Statistics Birth Record to characterize maternal risk factors (before and during pregnancy) and birth outcomes for both undocumented (5,961) and all other women (112,943). One limitation of this study was merging two distinct databases without unique identifiers.

Physical health indicators in major mental illness: data from the Quality and Outcome Framework in the UK

Martin JL, Lowrie R, McConnachie A, et al. Physical health indicators in major mental illness: data from the Quality and Outcome Framework in the UK. *Lancet*. 2015;385 Suppl 1:S61. doi:10.1016/S0140-6736(15)60376-2

Theme: Health Equity, Quality Improvement
Journal, Year: Lancet, 2015
Summary: This comparison analysis found that patients with major mental illness were less likely to have their BMI and BP recorded than patients with chronic kidney disease or diabetes throughout the UK. Data for this study came from the UK's Quality and Outcome Framework (QOF), which is tied to the National Health Service's system for performance management.

Prevalence and Epidemiology of Diabetes in Canadian Primary Care Practices: A Report from the Canadian Primary Care Sentinel Surveillance Network

Greiver M, Williamson T, Barber D, et al. Prevalence and Epidemiology of Diabetes in Canadian Primary Care Practices: A Report from the Canadian Primary Care Sentinel Surveillance Network. *Canadian Journal of Diabetes*. 2014;38(3):179-185. doi:10.1016/j.jcjd.2014.02.030

Theme: Disease Management, Population Health
Journal, Year: Canadian Journal of Diabetes, 2014
Summary: Data was extracted from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN), which includes EMRs of 10 primary care practice-based research networks. There were 272,469 patients in the study. Criteria used to diagnosis diabetes was 2 ICD occurrences for diabetes, presence of Hbg A1C >7% or two elevated fasting glucoses. This study found a diabetes prevalence rate of 7.6%. They did not use medication criteria. This data is helpful in capturing some undiagnosed diabetes by using lab values.

2. Quality Improvement

Control of glycemia and blood pressure in British adults with diabetes mellitus and subsequent therapy choices: a comparison across health states

McAlister FA, Lethebe BC, Lambe C, Williamson T, Lowerison M. Control of glycemia and blood pressure in British adults with diabetes mellitus and subsequent therapy choices: a comparison across health states. *Cardiovascular Diabetology*. 2018;17(1):27. doi:10.1186/s12933-018-0673-4

Theme: Quality Improvement, Obtaining Data Challenge
Journal, Year: Cardiovascular Diabetology, 2018
Summary: This retrospective cohort study in the UK used data from 670 NHS primary care practices with 4.4 million actively registered patients. The study identified and followed patients with diabetes to see if glycemic control, SBP control, and treatment deintensification differed based on health state (fit, mildly frail, or moderately/severely frail). The study found that deintensification was more common in patients with moderate-severe frailty than those who were fit. Almost one-third of all patients with diabetes in this study had treated HbA1C <6.5% or SBP <130 mmHg with little difference when re-measured 6 months later. Many clinical performance measures focus on under-treatment and failure to meet targets, but there is a need to focus on over-treatment and potential harms.

Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: prospective cohort study

Hippisley-Cox J, Coupland C. Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: prospective cohort study. *BMJ Open*. 2015;5(3):e007825. doi:10.1136/bmjopen-2015-007825

Theme: Quality Improvement, Disease Management – Physician-Patient
Journal, Year: BMJ Open 2015
Summary: This prospective cohort study used QResearch databases to create a validated risk prediction algorithm to estimate the future risks of common cancer. Data was used from the EHR of over 750 practices in the UK, along with hospital admission data, mortality statistics and the UK's national cancer registry. Three-quarters of the almost 5 million patients were used to develop the algorithm while the remaining one quarter were used for validation. The resulting algorithm can be used in clinical practice to help target screening programs at a population level or inform discussion between the doctor and patient to identify interventions to reduce future cancer risk.

Development of an algorithm for determining smoking status and behaviour over the life course from UK electronic primary care records

Atkinson MD, Kennedy JI, John A, Lewis KE, Lyons RA, Brophy ST. Development of an algorithm for determining smoking status and behaviour over the life course from UK electronic primary care records. *BMC Med Inform Decis Mak*. 2017;17. doi:10.1186/s12911-016-0400-6

Theme: Quality Improvement, Disease Management in Clinic, Obtaining Data Challenge
Journal, Year: BMC Medical Informatics and Decision Making, 2017
Summary: This studied looked at the differences in smoking status between General Practitioner (GP) reported data and Welsh Health Survey (WHS) data. They found that 6% of GP records had missing smoking status, which was higher than the 1.1% in the WHS. They developed an algorithm using the Secure Anonymized Information Linkage databank to classify a patient as a non-smoker, former smoker or smoker of 6,836 patients. They also used codes such as referral to cessation therapy, nicotine replacement therapy and other smoking related to codes to adequately assess smoking status. The algorithm detects 30% more smokers than the WHS data.

Trends in end digit preference for blood pressure and associations with cardiovascular outcomes in Canadian and UK primary care: a retrospective observational study

Greiver M, Kalia S, Voruganti T, et al. Trends in end digit preference for blood pressure and associations with cardiovascular outcomes in Canadian and UK primary care: a retrospective observational study. *BMJ Open*. 2019;9(1):e024970. doi:10.1136/bmjopen-2018-024970

Theme: Obtaining Data Challenge, Patient Safety, Disease Management in Clinic
Journal, Year: BMJ, 2018
Summary: This cross-sectional observational study looked at end digit preference (EDP) in blood pressure reading in primary care practices in Canada and the UK and its association between uptake of automated office BP machines and cardiovascular outcomes. Datasets included the Canadian Primary Care Sentinel Surveillance Network (181 sites with over 700,000 patients) and the Royal College of General Practitioners Research and Surveillance Centre (164 sites and over 500,000 patients). EDP was greater among female patients and patients without hypertension or diabetes. There was a higher risk of MI, stroke and angina at sites with higher EDP. Family practices that never or rarely use automated BP machines had higher EDP than those that had at least some use. European guidelines recommended automated machines. In the UK, BPs ending in zero went from 71.2% in 1996-1997 to 36.7% in 2005-2006.

National Use and Effectiveness of b-Blockers for the Treatment of Elderly Patients After Acute Myocardial Infarction

Krumholz HM, Radford MJ, Wang Y, Chen J, Heiat A, Marciniak TA. National use and effectiveness of beta-blockers for the treatment of elderly patients after acute myocardial infarction: National Cooperative Cardiovascular Project. JAMA. 1998;280(7):623-629.

Theme: Quality Improvement, Disease Management
Journal, Year: JAMA 1998
Summary: This retrospective cohort study looked at patients over the age of 65 who suffered from an acute MI (115,015 patients) from 1994 to 1995, if they were placed on a beta-blocker at discharge, and their mortality within a year of discharge. Clinical data abstraction originated from hospital records and was then entered directly into the database. The study found that less than half of eligible patients (those with no contraindication for a beta-blocker) were placed on a beta blocker at discharge. The 1-year mortality for patients placed on a beta blocker was 7.7% compared to 12.6% for those not prescribed a beta-blocker ($P<0.001$). This study would be an improved big data case study had it used pharmacy records to see if patients were placed on a beta blocker later or had it looked at compliance during the 1 year after discharge.

Web-scale pharmacovigilance: listening to signals from the crowd

White RW, Tatonetti NP, Shah NH, Altman RB, Horvitz E. Web-scale pharmacovigilance: listening to signals from the crowd. J Am Med Inform Assoc. 2013;20(3):404-408. doi:10.1136/amiajnl-2012-001482

Theme: Quality Improvement, Disease Management (Adverse Drug Reactions)
Journal, Year: Journal of the American Medical Informatics Association, 2013
Summary: This study used search engine data from 6 million web searchers who made 82 million queries for drugs, symptoms and conditions in a 12-month time frame. They found that 0.43% of sample web searchers searched at least one of the top 100 selling drugs. They examined queries specific for pravastatin and paroxetine, finding that hyperglycemia queries were higher in users searching for both drugs than either drug alone. This study indicates a potential public health benefit from using web searches to provide national sensors to discover potential adverse drug reactions or side effects.

Does interhospital transfer improve outcome of acute myocardial infarction? A propensity score analysis from the Cardiovascular Cooperative Project

Westfall JM, Kiefe CI, Weissman NW, et al. Does interhospital transfer improve outcome of acute myocardial infarction? A propensity score analysis from the Cardiovascular Cooperative Project. BMC Cardiovasc Disord. 2008;8:22. doi:10.1186/1471-2261-8-22

Theme: Quality Improvement, Disease Management (Emergency)
Journal, Year: BMC Cardiovascular Disorders, 2008
Summary: This retrospective cohort study used Medicare claims data of 234,769 (184,295 after exclusion) patients who suffered an acute MI from 6,684 (4,765 after exclusion) hospitals to examine outcome of transferred patients. They found that transferred patients were younger, more likely to be male, less likely to be African-American, and less likely to have diabetes or heart failure. 30-day mortality (from the HCFA administrative data) was lower in transferred patients. For hospitals without advanced cardiac services, transfer might be seen more as a treatment option. This study signals an opportunity to use big data for a predictive learning model to identify patients who would most benefit from a transfer.

Tooth Loss and Cardiovascular Disease Mortality Risk – Results from the Scottish Health Survey

Watt RG, Tsakos G, de Oliveira C, Hamer M. Tooth Loss and Cardiovascular Disease Mortality Risk – Results from the Scottish Health Survey. PLoS One. 2012;7(2). doi:10.1371/journal.pone.0030797

Theme: Quality Improvement, Disease Management (Prevention)
Journal, Year: PLoS ONE, 2012
Summary: This prospective cohort study used data from the Scottish Health Survey linked to mortality data and showed an increase risk for all-cause mortality, CVD, and stroke for patients with poor dental status (specifically edentulous). The original sample size was 16,144 participants with a final sample of 12,871 after exclusion.

Reducing complexity: a visualisation of multimorbidity by combining disease clusters and triads

Schäfer I, Kaduszkiewicz H, Wagner H-O, Schön G, Scherer M, van den Bussche H. Reducing complexity: a visualisation of multimorbidity by combining disease clusters and triads. *BMC Public Health*. 2014;14(1):1285. doi:10.1186/1471-2458-14-1285

Theme: Disease Management (Population Level), Quality Improvement
Journal, Year: BMC Public Health 2014
Summary: This study used diagnosis codes from a large German insurer with 1.7 million enrollees. The authors used a variety of common diagnoses (>1% prevalence) to find disease clusters in patients over 65. They found that lower back pain was associated with 16 other conditions. Depression was an important mediator of disease connections in females, and hypertension serves as a bridge between disease. This study is limited by not including all diseases (only >1% prevalence). This study demonstrates the use of data to provide better and more complete services and treatment for whole person and not just individual diagnosis.

Control of Glycemia and Cardiovascular Risk Factors in Patients with Type 2 Diabetes in Primary Care in Catalonia

Vinagre I, Mata-Cases M, Hermosilla E, et al. Control of glycemia and cardiovascular risk factors in patients with type 2 diabetes in primary care in Catalonia (Spain). *Diabetes Care*. 2012;35(4):774-779. doi:10.2337/dc11-1679

Theme: Disease Management (Population Level And Individual), Challenge – Data Input
Journal, Year: Diabetes Care 2012
Summary: This study used data from the EHR of the Catalonian health system which covers 3.7 million patients (around 80% of the population). The study specifically used records from the CatSalut, which is a database of prescriptions, and found a diabetes prevalence rate of 7.6%. The ability to link primary care records to pharmacy database added strength to this study, which was the largest study of its kind in Europe.

3. Predictive Learning for Disease Management

Mucopolysaccharidosis type II detection by Naïve Bayes Classifier: An example of patient classification for a rare disease using electronic medical records from the Canadian Primary Care Sentinel Surveillance Network

Ehsani-Moghaddam B, Queenan JA, MacKenzie J, Birtwhistle RV. Mucopolysaccharidosis type II detection by Naïve Bayes Classifier: An example of patient classification for a rare disease using electronic medical records from the Canadian Primary Care Sentinel Surveillance Network. *PLOS ONE*. 2018;13(12):e0209018. doi:10.1371/journal.pone.0209018

Theme: Disease Management - Predictive Learning, Rare Diseases
Journal, Year: PLOS ONE, 2018
Summary: MPS II is a rare, progressive, X-linked disease caused by deficiency in a lysosomal enzyme. Ehsani-Moghaddam et al. developed a Naïve Bayes classification (NBC) algorithm utilizing clinical diagnosis and symptoms of patients in the Canadian Primary Care Sentinel Surveillance Network. They identified 125 patients of the 506,497 males that had the highest likelihood of disease. The diagnosis was not confirmed with the gold standard test, absence of I2S enzyme, but the model could reserve the expensive test for patients that have a positive prescreening result using the NBC model.

County-level Vulnerability Assessment for Rapid Dissemination of HIV or HCV Infections among Persons who Inject Drugs, United States

Van Handel MM, Rose CE, Hallisey EJ, et al. County-Level Vulnerability Assessment for Rapid Dissemination of HIV or HCV Infections Among Persons Who Inject Drugs, United States. *J Acquir Immune Defic Syndr*. 2016;73(3):323-331. doi:10.1097/QAI.0000000000001098

Theme: Predictive Learning for Disease Prevention, Challenges (Missing Data – Incomplete)
Journal, Year: Journal of Acquired Immune Deficiency Syndromes, 2016
Summary: This study focused on how to identify counties that might be vulnerable for an infectious disease outbreak among IV drug users. They used 48 proxy measures as potential indicators. Data sources included the National Notifiable Disease Surveillance System (NNDSS), DEA records, CMS-NPI, American Community Survey, interstate highway access (ESRI maps), US Census and waiver data from SAMHSA . Some of the data sources were >3 years old and data was missing from 173 counties in 8 states for the time period studied. They identified 220 counties in 26 states that were vulnerable communities. Indicators absent from this study that could be used on other local level-analyses include ER visits/admissions for overdoses.

The multimorbidity cluster analysis tool: identifying combinations and permutations of multiple chronic diseases using a record-level computational analysis.

Nicholson K, Bauer M, Terry A, Fortin M, Williamson T, Thind A. The Multimorbidity Cluster Analysis Tool: Identifying Combinations and Permutations of Multiple Chronic Diseases Using a Record-Level Computational Analysis. *J Innov Health Inform*. 2017;24(4):962. doi:10.14236/jhi.v24i4.962

Theme: Aid to Predictive Learning
Journal, Year: Journal of Health Informatics, 2017
Summary: This article focused on the development of a toolkit created using patients with multiple chronic disease and identifying combinations of those diseases (example – hypertension, obesity, cancer) and distinct permutations, which is the order that the diseases occurred. They used data from the Canadian Primary Care Sentinel Surveillance Network, which contains de-identified EMR data of 1,000,000 primary care patients. The authors entered in 75,000 individual records into the program. They found 6,095 unique combinations and 14,911 unique permutations among female patients, and 4,316 and 9,736 for male patients, respectively. The toolkit can be modified to add the time that elapses between diseases.

Construction of a Multisite DataLink Using Electronic Health Records for the Identification, Surveillance, Prevention, and Management of Diabetes Mellitus: The SUPREME-DM Project

Nichols GA, Desai J, Elston Lafata J, et al. Construction of a multisite DataLink using electronic health records for the identification, surveillance, prevention, and management of diabetes mellitus: the SUPREME-DM project. *Prev Chronic Dis*. 2012;9:E110.

Theme: Disease Management, Predictive Learning – Future Studies to Follow Patterns, Quality Improvement (Patient Identification), Challenges (Not All Same EHR)
Journal, Year: Preventing Chronic Disease 2012
Summary: This study used EHR data from 11 integrated health systems to identify patients with diabetes using real time laboratory testing. This study found a diabetes incident rate of 6.9% out of the 15 million unique members. Of those 6.9%, 39.4% had incident diabetes, which means they had not been identified in at least the last two years as part of the health system. This type of data can be used to build registries for both patient identification and inform future research on the progression and timeline of diseases. A national surveillance system could help lead to primary and secondary prevention of disease over time and lead to a better understanding of complications.

4. Cost-Savings

Medicare Spending after 3 years of the Medicare Shared Savings Program

McWilliams JM, Hatfield LA, Landon BE, Hamed P, Chernew ME. Medicare Spending after 3 Years of the Medicare Shared Savings Program. *New England Journal of Medicine*. 2018;379(12):1139-1149. doi:10.1056/NEJMsa1803388

Theme: Cost Savings
Journal, Year: New England Journal of Medicine, 2018
Summary: This retrospective cohort study examines the cost savings difference between physician-group ACOs and hospital-integrated ACOs who were part of the Medicare Shared Savings Program. Using Medicare Claims data, they found that physician-group ACOs produced increasingly greater cost savings over the six-year period from 2009 to 2015. The net savings to Medicare in 2015 from physician-group ACOs was $256.4 million. They also found that on average, patients' health declined. The study had several limitations.

Practice variation in the use of tests in UK primary care: a retrospective analysis of 16 million tests performed over 3.3 million patient years in 2015/16

O'Sullivan JW, Stevens S, Oke J, et al. Practice variation in the use of tests in UK primary care: a retrospective analysis of 16 million tests performed over 3.3 million patient years in 2015/16. *BMC Medicine*. 2018;16(1):229. doi:10.1186/s12916-018-1217-1

Theme: Cost Savings, Quality Improvement
Journal, Year: BMC Medicine 2018
Summary: The purpose of this study was to identify practice site variation in 44 specific tests. This study used the Clinical Practice Research Datalink, which is a large database of EHRs from UK primary care, accounting for about 7% of the UK population. There were over 16 million tests ordered on 4 million patients in the sample. The tests with the largest variation in use were non-illicit drug monitoring tests (digoxin, lithium, etc.), urine microalbumin, pelvic CT and pap smear. The tests with the lowest variation were testosterone tests.

PART 2:
CHALLENGES OF USING BIG DATA

Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research

Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc*. 2013;20(1):144-151. doi:10.1136/amiajnl-2011-000681

Theme: Challenges in Data Quality
Journal, Year: Journal of the American Medical Informatics Association, 2013
Summary: This review articles highlights the research on the quality of EHR data. Most papers focus on structured data, with only 22% looking at unstructured data. Completeness, correctness, concordance, plausibility and currency are five dimensions of data quality that were assessed. There is a lack of consistent terminology and taxonomy for data quality which creates a barrier to reviewing research. There is an increasing need for a standardized, systemic data quality assessment method to ensure sufficient quality of reused EHR data for clinical research.

Finding and using routine clinical datasets for observational research and quality improvement

McDonnell L, Delaney BC, Sullivan F. Finding and using routine clinical datasets for observational research and quality improvement. *Br J Gen Pract*. 2018;68(668):147-148. doi:10.3399/bjgp18X695237

Theme: Challenges in Data Quality
Journal, Year: British Journal of General Practice, 2018
Summary: This review discusses the increasing important role of data in primary care. The Farr Institute was created from 2012-2018 as a resource to build health informatics capacity and has since merged to form the Health Data Research UK, which unites the UK's health data to make it available to researchers. This review highlights using EMRs, audit data, prescribing data, and health surveys to create special datasets and cohort studies for future research.

Defining and measuring completeness of electronic health records for secondary use

Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. *J Biomed Inform*. 2013;46(5):830-836. doi:10.1016/j.jbi.2013.06.010

Theme: Challenges in Data Quality
Journal, Year: Journal of Biomedical Informatics, 2013
Summary: This article focuses on the completeness of the Electronic Health Record and its impacts for researchers and clinicians that want to secondarily use EHR data. They used four different definitions for completeness: documentation, breadth, density and predictability. Data from Allscripts, Cerner and Eagle Registration were used with approximately 3. 9 million patients. The study found that 26.9% of records met one definition of completeness, while only 0.6% met criteria for all four definitions. Researchers should determine what their data needs will be and select an appropriate definition for completeness for data reuse.

What's holding up the big data revolution in healthcare?

Dhindsa K, Bhandari M, Sonnadara RR. What's holding up the big data revolution in healthcare? *BMJ*. 2018;363:k5357. doi:10.1136/bmj.k5357

Theme: General Challenges of Big Data
Journal, Year: BMJ, 2018
Summary: This editorial discusses the current limits of big data. These include the limitations of algorithms and machine learning, which are trained to recognize patterns, but sometimes can be misdirected or find patterns that fail to have real-world applications. Also, data is distributed among different health systems and there is not unity in documentation or standards. This editorial calls for more training in data collection.

Digital Pharmacovigilance and Disease Surveillance: Combining Traditional and Big-Data Systems for Better Public Health

Salathé M. Digital Pharmacovigilance and Disease Surveillance: Combining Traditional and Big-Data Systems for Better Public Health. *J Infect Dis*. 2016;214(suppl_4):S399-S403. doi:10.1093/infdis/jiw281

Theme: Ethical Challenges, Public Health, Disease Management and Prevention
Journal, Year: Journal of Infectious Diseases, 2016
Summary: This article discusses the ability to harness patient (user) data to predict flu trends and adverse drug reactions.

Are big data analytics helpful in caring for multimorbid patients in general practice? - A scoping review

Waschkau A, Wilfling D, Steinhäuser J. Are big data analytics helpful in caring for multimorbid patients in general practice? - A scoping review. BMC Family Practice. 2019;20(1):37. doi:10.1186/s12875-019-0928-5

Theme: Big Data Analytics
Journal, Year: BMC Family Practice, 2019
Summary: This piece is an extensive literature search for articles that used big data. The authors' conclusion is that most articles are using big data for patients with multiple conditions, however one article was disease-specific.

Impact of Double Counting and Transfer Bias on Estimated Rates and Outcomes of Acute Myocardial Infarction

Westfall JM, McGloin J. Impact of double counting and transfer bias on estimated rates and outcomes of acute myocardial infarction. Med Care. 2001;39(5):459-468.

Theme: Big Data Challenges, Analytics
Journal, Year: Journal of Medical Care, 2001
Summary: This article discusses patient double-counting in reported outcomes and rates of AMI. The authors used state hospital discharge data from 8 states over a two-year period. Double counting ranged from 10-15%, higher in rural counties. Older patients and female patients were less likely to be double counted.