MEDICAL BIG DATA AND BIG DATA QUALITY PROBLEMS

SHARONA HOFFMAN*

***

Medical big data has generated much excitement in recent years and for good reason. It can be an invaluable resource for researchers in general and insurers in particular. This Article, however, argues that users of medical big data must proceed with caution and recognize the data's considerable limitations and shortcomings. These include data errors, missing information, lack of standardization, record fragmentation, software problems, and other flaws. This Article analyzes a variety of data quality problems and then formulates recommendations to address these deficiencies, including data audits, workforce and technical solutions, and regulatory approaches.

***

I. INTRODUCTION

The term "big data" is suddenly pervasive. The *New York Times* deemed this the "Age of Big Data" in a 2012 article,[1] and a Google search for the term yields over 15 million hits. "Big data" is difficult to define precisely, but it is characterized by three attributes known as "the three Vs": its large volume, its variety, and its velocity, that is, the frequency with which it is generated.[2] A particularly rich, but sensitive, type of big data is medical big data, which holds great promise as a resource for researchers and analysts in general, and insurers in particular. Public and private enterprises are launching numerous medical big data initiatives. One of the largest is scheduled to become operational in September 2015 and to link information from hospitals, academic centers, community clinics, insurers, and others sources. This data repository, funded by the

---

* Edgar A. Hahn Professor of Law and Professor of Bioethics, Co-Director of Law-Medicine Center, Case Western Reserve University School of Law; B.A., Wellesley College; J.D., Harvard Law School; LL.M. in Health Law, University of Houston. The author would like to thank Professors Peter Swire and Andy Podgurski for their thoughtful comments regarding the subject of this Article and Tracy (Yeheng) Li for her dedicated research assistance.

[1] Steve Lohr, *The Age of Big Data*, N.Y. TIMES (Feb. 11, 2012), http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=all&_r=0.

[2] Philip Russom, *Big Data Analytics*, TDWI BEST PRACTICES REPORT 1, 6 (4th Quarter 2011).

federal government, will contain information pertaining to twenty-six to thirty million Americans.[3]

Medical big data may consist of patient electronic health records (EHR), insurance claims, and pharmacy prescription drug information. It is of interest to a broad range of insurers, including those issuing health, life, disability, and long-term care policies, who may use it for purposes of underwriting, evaluating physicians, assessing benefits coverage, and detecting fraud. Medical big data is also invaluable for purposes of biomedical research, public health practice, institutions' quality assessment and improvement efforts, and post-marketing surveillance of drugs and devices, among other initiatives.[4] Such data uses are known as "secondary uses" of medical information, to be distinguished from the data's primary use for clinical and billing purposes.[5]

This Article's primary argument is that as valuable as medical big data can be, it must be approached cautiously. Clinicians collect data for treatment and billing purposes, and thus, it may not always be a good fit for secondary uses.[6]

Anyone employing large collections of complex medical data must recognize the data's considerable limitations and shortcomings.[7] Data quality problems are particularly relevant to insurers because they affect not only secondary use but also their primary work of processing benefit claims. Furthermore, because public programs, including Medicare, Medicaid, and the Children's Health Insurance Program, cover over thirty

---

[3] Ariana Eunjung Cha, *Scientists Embark on Unprecedented Effort to Connect Millions of Patient Medical Records*, WASH. POST (Apr. 15, 2014), http://www.washingtonpost.com/national/health-science/scientists-embark-on-unprecedented-effort-to-connect-millions-of-patient-medical-records/2014/04/15/ea7c966a-b12e-11e3-9627-c65021d6d572_story.html.

[4] Sharona Hoffman & Andy Podgurski, *The Use and Misuse of Biomedical Data: Is Bigger Really Better?* 39 AM. J.L. & MED. 497, 506–15 (2013).

[5] Taxiarchis Botsis et al., *Secondary Use of EHR: Data Quality Issues and Informatics Opportunities*, AMIA JOINT SUMMITS TRANSLATIONAL SCI. PROC. 1, 1 (2010); Jessica S. Ancker et al., *Root Causes Underlying Challenges to Secondary Use of Data*, AMIA ANN. SYMP. PROC. 57, 57 (2011).

[6] Brian J. Wells et al., *Strategies for Handling Missing Data in Electronic Health Record Derived Data*, 1 EGEMS 1, 1 (2013), *available at* http://repository.academyhealth.org/egems/vol1/iss3/7/.

[7] *See* Hoffman & Podgurski, *supra* note 4 (for an additional discussion of data quality and analysis problems).

percent of the population,[8] claims accuracy is of great importance to the government and taxpayers alike. While this Article will be illuminating for insurers, it has much broader applicability as well. All researchers and analysts using medical data for secondary purposes should be familiar with the data flaws analyzed here and may benefit from the recommendations that are developed.

This Article will proceed as follows. Part II of this Article details the purposes for which insurers may use big data. Part III analyzes a large number of data quality problems that may affect EHRs. These can be generally characterized as: 1) deficiencies in data veracity, 2) data voids, and 3) software problems. Part IV formulates recommendations to address data quality problems, including data audits, workforce and technical solutions, and regulatory approaches.

II. INSURERS' USE OF BIG DATA

Insurers have much to gain from using medical big data. Insurers' own claims databases constitute a rich resource for analysis. With medical releases from patients, insurers can also gain access to pharmacies' prescription drug databases and patients' full EHRs, including medical histories, diagnoses, treatments, and other details. Insurers may seek to analyze medical information for a variety of purposes, including underwriting, physician tiering, decisions about coverage scope, and fraud and abuse investigations.

A. UNDERWRITING

Underwriting is the process by which insurers choose whom they will insure and under what terms.[9] To that end, insurers issuing policies for life, long-term care, and disability insurance generally require applicants to

---

[8] *Health Insurance Coverage of the Total Population*, THE HENRY J. KAISER FAMILY FOUND. (2012), http://kff.org/other/state-indicator/total-population/.

[9] 42 U.S.C. § 1395ss(x)(3)(E) (2012). The provision defines underwriting as including: "(i) rules for, or determination of, eligibility (including enrollment and continued eligibility) for benefits under the policy; (ii) the computation of premium or contribution amounts under the policy; (iii) the application of any pre-existing condition exclusion under the policy; and (iv) other activities related to the creation, renewal, or replacement of a contract of health insurance or health benefits."

sign medical releases that allow insurers to review their health records.[10] Based on health information, insurers may reject applicants who are perceived to be at high risk for costly medical problems (or, in the case of life insurers, early death) or charge them high premiums. Some insurers purchase applicants' prescription drug histories from companies such as ScriptCheck and IntelliScript that obtain prescription information from pharmacy benefit management companies.[11] ScriptCheck, for example, advertises that it helps insurers "uncover crucial application omissions or assess the veracity of the application."[12] Specifically, ScriptCheck provides

> Profiles [that] include the results of a five-year history search with detailed drug and insurance eligibility information, treating physicians, drug indications and pharmacy information. In addition, the likelihood that the applicant has a particular condition is included, which is derived from the predictive modeling that is performed by Optum MedPoint.[13]

Health insurers constitute a special case. Unlike life, disability, and long-term care insurers, they are subject to considerable regulatory restrictions and anti-discrimination mandates that govern underwriting. Under the Genetic Information Nondiscrimination Act, health insurers may not obtain or use genetic information for underwriting purposes.[14] Furthermore, the Health Insurance Portability and Accountability Act (HIPAA) has long prohibited health insurers that issue group policies from charging particular group members different premiums or from denying policies to particular members of the group because of their health status. Thus, for example, if Blue Cross offers a group policy to an employer, it

---

[10] *Fact Sheet 8: Introduction to Medical and Health Information Privacy,* PRIVACY RIGHTS CLEARINGHOUSE (Aug. 2014), https://www.privacyrights.org/medical-records-privacy.

[11] Chad Terhune, *They Know What's in Your Medicine Cabinet*, BLOOMBERG BUSINESSWEEK (July 22, 2008), http://www.businessweek.com/stories/2008-07-22/they-know-whats-in-your-medicine-cabinet; David Lazarus, *Your Prescription History Is Their Business*, L.A. TIMES (Oct. 21, 2013), http://articles.latimes.com/2013/oct/21/business/la-fi-lazarus-20131022.

[12] *ScriptCheck*®, EXAMONE, http://wwsw.examone.com/our-solutions/scriptcheck (last visited Oct. 12, 2014).

[13] *Id.*

[14] 42 U.S.C. § 300gg-53 (2012); 26 U.S.C. § 9802(b)(3)–(c) (2012).

cannot decline to cover employees with a cancer history or charge them higher premiums than others.[15] By contrast, traditionally, insurers offering individual policies were not subject to the same underwriting restrictions.[16] The Patient Protection and Affordable Care Act (PPACA), however, now severely limits the discretion of health insurers operating in the individual market. The law establishes requirements for "fair health insurance premiums"[17] and bans all preexisting condition exclusions.[18] Nevertheless, the PPACA applies only to health insurers and does not extend to life, long-term care, or disability insurers.[19]

B. PHYSICIAN TIERING

Some insurers analyze claims data in order to rank or tier physicians within the same specialty type and geographic market.[20] Insurers frequently categorize doctors into tiers based on their cost and quality of performance. They then offer consumers financial incentives, such as lower co-payments, in order to encourage them to visit higher-tiered doctors.[21]

For purposes of tiering, insurers assess two factors: cost efficiency and performance quality. To evaluate the cost of physicians' care, insurers divide each patient's claim records into specific "episodes of care" by employing data-mining algorithms. Insurers attribute each episode of care (e.g. a patient's pneumonia) to a treating physician and calculate an actual cost figure.[22] This, in turn, is compared to an expected cost figure,

---

[15] 42 U.S.C. §§ 300gg-1(b), -11 (2012).

[16] 42 U.S.C. § 300gg-41 (2012); Sharona Hoffman, *Unmanaged Care: Towards Moral Fairness in Health Care Coverage*, 78 IND. L.J. 659, 678 (2003).

[17] 42 U.S.C. § 300gg (2012).

[18] *Id*.

[19] 42 U.S.C. §§300gg, -4 (2012).

[20] CIGNA, *Cigna Care Designation & Physician Quality & Cost-Efficiency Displays 2013 Methodologies Whitepaper* (Feb. 2013), *available at* http://www.cigna.com/pdf/2013-cigna-care-designation-methodology.pdf.

[21] *See* Anna D. Sinaiko & Meredith B. Rosenthal, *The Impact of Tiered Physician Networks on Patient Choice*, HEALTH SERVS. RES. 1348, 1357 (2014), *available at* http://onlinelibrary.wiley.com/doi/10.1111/1475-6773.12165/pdf.

[22] Episodes are attributed to particular physicians based on attribution rules, as seen in the rule that dictates "responsibility is assigned to a physician who accounts for 30% or more of professional and prescribing costs included in the episode."

determined by averaging the actual cost of all similar episodes managed by physicians in the same specialty. Each doctor's cost efficiency measure is the ratio of her total actual costs to total expected costs, and doctors are tiered based on their comparative ratios.[23]

The quality of care figure is developed by analyzing information about the degree to which physicians comply with clinical guidelines relating to various conditions.[24] For example, analysts might assess whether patients with type II diabetes were given all the recommended tests and medications. Performance is scored either in terms of the physician's compliance rate compared to the average adherence rate for the specialty or in terms of a fixed compliance standard.[25]

C. RESEARCH REGARDING BENEFITS COVERAGE AND FRAUD

Health insurers may also conduct research to determine if certain patients should be covered for and encouraged to obtain additional services in order to save costs in the long-run. For example, elderly patients may benefit from home visits by a nurse after a hospitalization in order to prevent medical problems that could result in a second hospitalization. Likewise, individuals with chronic diseases such as diabetes may benefit from care management programs.[26]

Insurers can also mine medical data resources in order to detect health care fraud and abuse. They can establish claim norms and then identify anomalous claims patterns that might signify fraudulent conduct.[27]

See Lewis G. Sandy et al., *Episode-Based Physician Profiling: A Guide to the Perplexing*, 23 J. GEN. INTERNAL MED. 1521, 1522 (2008).

[23] *Id.*

[24] *Id.*

[25] *Id.*

[26] *Care Management Analytics*, KNOWLEDGENT, http://knowledgent.com/whitepaper/care-management-analytics (last visited Oct. 12, 2014); Jennifer Valentino-DeVries, *May the Best Algorithm Win . . .*, WALL ST. J. (Mar. 16, 2011), http://online.wsj.com/news/articles/SB10001424052748704662604576202392747278936.

[27] *See* Hian Chye Koh & Gerald Tan, *Data Mining Applications in Healthcare*, 19 J. HEALTHCARE INFO MGMT. 64 (2005).

III. DATA QUALITY PROBLEMS

The validity of researchers' and analysts' findings will often depend on the accuracy and completeness of the information upon which they are based. Unfortunately, patient EHRs and the insurance claims and prescriptions orders that flow from them are often deeply flawed. They suffer from data veracity defects and data voids. In addition, software or programming problems may generate errors in the data itself, may limit researchers' ability to extract data, or may obstruct data analysis.[28] Researchers must understand and consider these many potential shortcomings and pitfalls as they proceed with their analysis.

A. DATA VERACITY

EHRs are created by very busy clinicians. On average, doctors spend only thirteen to eighteen minutes with each patient.[29] Whether they attempt to enter data during the patient encounter or attend to documentation afterwards, they are likely to work quickly and to make mistakes.

---

[28] K. Bruce Bayley et al., *Challenges in Using Electronic Health Record Data for CER Experience of 4 Learning Organizations and Solutions Applied*, 51 MED. CARE S80, S81 (2013); George Hripcsak & David J. Albers, *Next-Generation Phenotyping of Electronic Health Records*, 20 J. AM. MED. INFORMATICS ASS'N 117, 117–18 (2013).

[29] *See* Andrew Gottschalk & Susan A. Flocke, *Time Spent in Face-to-Face Patient Care and Work Outside the Examination Room*, 3 ANNALS FAM. MED. 488, 491 (2005) (finding that the average time per patient was 13.3 minutes); Kimberly S. H. Yarnall et al., *Family Physicians as Team Leaders: See "Time" to Share the Care*, PREVENTING CHRONIC DISEASE: PUB. HEALTH RES. PRAC. & POL'Y 1, 6, Apr. 2009, http://www.cdc.gov/pcd/issues/2009/apr/08_0023.htm (finding that the mean length for an acute care visit is 17.3 minutes, the mean for a chronic disease care visit is 19.3 minutes, and the average for a preventive care visit is 21.4 minutes, and that of total clinical time spent by physicians, these comprise 45.8%, 37.4%, and 16.8% respectively); Kevin Fiscella & Ronald M. Epstein, *So Much to Do, So Little Time: Care for the Socially Disadvantaged and the 15-Minute Visit*, 168 ARCHIVES INTERNAL MED. 1843, 1843 (2008) ("The average office visit in the United States lasts for about 16 minutes.").

1. Input Errors

Clinicians entering data into EHRs often mistype words, invert numbers, or select wrong menu items from drop-down menus. They may also choose erroneous diagnosis codes, check boxes incorrectly, or uncheck boxes inappropriately if the default setting has all boxes checked.[30]

Presumably, such errors are made innocently. However, there are also some perverse incentives at play. If a clinician checks a few too many boxes, for example, she can make it look like she did more during the clinical encounter than she actually did, and consequently, she can bill a higher amount. Similarly, selecting a code for a slightly more serious condition than the patient has may justify increased charges. Such billing manipulations are known as "upcoding."[31] According to one study, upcoding services provided to Medicare patients is so common that it may account for as much as fifteen percent of Medicare's expenditures for general office visits, or $2.13 billion annually.[32]

2. Data Entered Into Wrong Patient Charts

Data can be entered into the wrong patient chart if multiple patient charts are open at the same time or if a prior user did not log off properly after viewing another patient's EHR.[33] Such errors are particularly likely in hospitals. During a typical hospitalization, approximately 150 individuals view each patient's chart, and multiple records may be handled at once in nursing stations.[34]

---

[30] Farah Magrabi et al., *An Analysis of Computer-Related Patient Safety Incidents to Inform the Development of A Classification*, 17 J. AM. MED. INFORMATICS ASS'N. 663, 665, 669 (2010); Sharona Hoffman & Andy Podgurski, *E-Health Hazards: Provider Liability and Electronic Health Record Systems*, 24 BERKELEY TECH. L. J. 1523, 1544–45 (2009) (discussing input errors); Botsis et al., *supra* note 5, at 1; Ancker et al., *supra* note 5, at 57.

[31] Christopher S. Brunt, *CPT Fee Differentials and Visit Upcoding Under Medicare Part B*, 20 HEALTH ECON. 831, 840 (2011).

[32] *Id.* (the $2.13 billion figure is in 2007 dollars).

[33] Elizabeth Borycki, *Trends in Health Information Technology Safety: From Technology-Induced Errors to Current Approaches for Ensuring Technology Safety*, 19 HEALTH INF. RES. 69, 70 (2013).

[34] Judy Foreman, *At Risk of Exposure: In the Push for Electronic Medical Records, Concern is Growing about How Well Privacy Can Be Safeguarded*, L.A.

3. Copy and Paste Problems

The EHR copy and paste feature is notorious as a source of errors.[35] It is designed to save time, allowing physicians to copy narrative from a prior visit and paste it into new visit notes. However, if the copied information is not carefully edited and updated, the physician will inadvertently introduce errors into the record.[36] For example, in one reported case, the record of a patient hospitalized for many weeks because of complications from surgery indicated each day that this was "post-op day No. 2" because the note was never edited.[37] In another case, the statement "Patient needs drainage, may need OR" appeared in notes for several consecutive days, even after the patient successfully underwent a procedure to drain his abscess.[38] In yet another instance, a patient's EHR indicated erroneously that he had a below-the-knee amputation (BKA) because a voice recognition dictation system entered "BKA" into the record instead of the real problem - diabetic ketoacidosis, whose acronym is DKA.[39]

Copy and paste is very commonly used. In a study of 100 randomly selected hospital admissions, copied text was found in seventy–eight percent of medical residents' sign-out notes (written when their shift ended) and fifty-four percent of patient progress notes.[40]

TIMES (June 26, 2006), http://articles.latimes.com/2006/jun/26/health/he-privacy26.

[35] Eugenia L. Siegler & Ronald Adelman, *Copy and Paste: A Remediable Hazard of Electronic Health Records*, 122 AM. J. MED. 495, 495–96 (2009) (cautioning that cut and paste functions can lead to patient problem lists never changing, notes and errors being copied by multiple staff members, and loss of accurate narrative).

[36] Lena Mamykina et al., *Clinical Documentation: Composition or Synthesis?*, 19 J. AM. MED. INFORMATICS ASS'N. 1025, 1027 (2012).

[37] Kevin B. O'Reilly, *EHRs: "Sloppy and Paste" Endures Despite Patient Safety Risk*, AM. MED. NEWS (Feb. 4, 2013), http://www.amednews.com/article/20130204/profession/130209993/2/.

[38] *Id.*

[39] Paul Hsieh, *Can You Trust What's In Your Electronic Medical Record?*, FORBES (Feb. 24, 2014), http://www.forbes.com/sites/paulhsieh/2014/02/24/electronic-medical-record/.

[40] Jesse O. Wrenn et al., *Quantifying Clinical Narrative Redundancy in an Electronic Health Record*, 17 J. AM. MED. INFORMATICS ASS'N 49, 52 (2010).

The data quality problems that copy and paste generates have been widely recognized. In 2014, the American Health Information Management Association issued a statement calling for copy/paste functionality to be "permitted only in the presence of strong technical and administrative controls which include organizational policies and procedures, requirements for participation in user training and education, and ongoing monitoring."[41] In the absence of such measures, the errors caused by copying and pasting EHR text can confuse treating physicians and claims administrators, harm patients, and taint records that will later be employed for secondary use by insurers and other researchers.

4. Estimating Error Rates

A variety of studies have focused on error rates in EHRs. One study involved oncology patients at an academic medical center and, in part, examined duplicate data that was entered into two research databases.[42] It showed that the rate of discrepancies between the two databases ranged between 2.3 and 26.9 percent, depending on the type of data, with demographic data having fewer inconsistencies and treatment data having many more discrepancies.[43] Another publication found an average error rate of 9.76 percent.[44] Australian researchers who audited 629 admissions at two Sydney hospitals identified 1,164 prescribing errors in

---

[41] AM. HEALTH INFO. MGMT. ASS'N, *Appropriate Use of the Copy and Paste Functionality in Electronic Health Records* (Mar. 17, 2014), http://library.ahima.org/xpedio/groups/public/documents/ahima/bok1_050621.pdf.

[42] Saveli I. Goldberg et al., *Analysis of Data Errors in Clinical Research Databases*, AMIA 2008 ANN. SYMP. PROC. 242, 242–43 (2008) (attributing errors to data entry mistakes, misinterpretation of hard-copy documents when information was typed into the database, and perpetuation of errors that were contained in the original paper documents and were not corrected during the transition to EHRs).

[43] *Id.* at 243–44.

[44] Meredith L. Nahm, *Quantifying Data Quality for Clinical Trials Using Electronic Data Capture*, PLOS ONE, AUG. 2008, at 1 (discussing a literature review of "42 articles that provided source-to-database error rates, primarily from registries" and finding that the "average error rate across these publications was 976 errors per 10,000 fields"); *see also* James J. Cimino et al., *Use of Clinical Alerting to Improve the Collection of Clinical Research Data*, AMIA 2009 SYMP. PROC. 218, 218 (2009) (discussing data error rates pertaining to research databases).

those patients' records, equivalent to 185 errors per 100 admissions.[45] They noted, however, that error rates had decreased significantly since the hospitals transitioned from paper medical records to EHRs, dropping from 625 inaccuracies per 100 admissions to 212 at one hospital and from 362 to 185 errors per 100 admissions at the other.[46]

B. DATA VOIDS

EHR data is often incomplete, lacking elements that would be valuable for secondary uses.[47] Data voids may arise because available data is not recorded or important information is not gathered. They may also occur because of billing code limitations, lack of data standardization, and record fragmentation.

1. Missing Data

In some instances physicians do not carefully record all the data that is available to them. For example, they may neglect to indicate clearly that a patient does not have particular symptoms or conditions and instead leave blank data fields. Analysts who see these empty fields will not know how to interpret them: did the patient not suffer the symptom at issue or did the physician overlook the question?[48]

In addition, data about treatment outcomes is often missing.[49] Patients who are given medications such as antibiotics often are not asked to return to the doctor and report on their progress. Therefore, the patient's EHR will detail the diagnosis and prescription, but will not indicate whether she recovered or failed to improve and sought treatment from a different physician or specialist.

---

[45] Johanna I. Westbrook et al., *The Safety of Electronic Prescribing: Manifestations, Mechanisms, and Rates of System-Related Errors Associated with Two Commercial Systems in Hospitals*, 20 J. AM. MED. INFORMATICS ASS'N 1159, 1161 (2013).

[46] *Id.* at 1164–65.

[47] Wells et al., *supra* note 6, at 1–3.

[48] *Id.* at 2.

[49] Craig Newgard et al., *Electronic Versus Manual Data Processing: Evaluating the Use of Electronic Health Records in Out-of-Hospital Clinical Research*, 19 ACAD. EMERGENCY MED. 217, 225 (2012).

Graphical representations are another element that may be useful to analysts but missing from EHRs. In the era of paper records, some doctors were accustomed to drawing anatomical pictures to depict the patient's medical condition, specifying by way of illustration exactly where the problem was and what it looked like. EHR systems' graphical representation tools are cumbersome and inadequate at best.[50] The inability to draw on paper is frustrating for some clinicians who feel that the absence of depictions compromises the quality of their documentation.

Studies that have evaluated data completeness have found diverse results.[51] Several studies focusing on patients' medication lists in EHRs found the following: 1) 27% of drugs were missing from ambulatory oncology patients' drug lists; 2) 53% of patient-reported medications were not recorded by primary care providers; and 3) an average of 3.1 medications were missing from the drug lists of Veterans Affairs (VA) patients who were 65 and older with five or more prescriptions.[52] A study of EHRs at eight VA clinical sites found that the following percentage of patients had missing data: 24% to 38% had incomplete LDL (low-density lipoprotein) measurements; 3% to 31% had incomplete blood pressure measurements, and 5% to 23% were missing HbA1c (blood sugar) results.[53]

---

[50] David S. Sanders et al., *Electronic Health Record Systems in Ophthalmology: Impact on Clinical Documentation*, 120 AM. ACAD. OPHTHALMOLOGY 1745, 1751–53 (2013).

[51] Kitty S. Chan et al., *Review: Electronic Health Records and the Reliability and Validity of Quality Measures: A Review of the Literature*, 67 MED. CARE RES. & REV. 503, 506 (2010).

[52] *Id.* at 515 (citing Saul N. Weingart et al., *Medication Reconciliation in Ambulatory Oncology*, 33 JOINT COMM'N J. QUALITY PATIENT SAFETY 750, 752 (2007)); Prathibha Varkey et al., *Improving Medication Reconciliation in the Outpatient Setting*, 33 JOINT COMM'N J. QUALITY PATIENT SAFETY. 286, 290 (2007); Peter J. Kaboli et al., *Assessing the Accuracy of Computerized Medication Histories*, 10 AM. J. MANAGED CARE 872, 872 (2004).

[53] Joseph L. Goulet et al., *Measuring Performance Directly Using the Veterans Health Administration Electronic Medical Record: A Comparison with External Peer Review*, 45 MED. CARE 73, 81 (2007).

2. Records of Sicker Patients Are More Complete

Experts have noted that the records of sick patients contain much more information than those of healthy patients.[54] Sick patients have more clinical visits, testing, and procedures than do individuals who are well and rarely if ever seek medical care. This information disparity may be problematic for researchers who want to know as much about healthy individuals and their health habits as they do about those who are less robust. It can also lead to selection bias, which is an error in choosing the individuals that will take part in a scientific study that occurs when the participants are not representative of the population as a whole.[55] If selection bias is present, the study's results may be valid for the group that was studied (e.g. very sick people), but cannot be generalized as applicable to others (e.g. healthier patients).[56]

3. Limitations of Billing Information

Billing information may be particularly vulnerable to data voids and insufficient specificity.[57] Diagnostic codes for billing may be too general to indicate the particulars of the patient's condition. For example, a billing code may indicate "myelodysplastic syndromes," which include a

---

[54] *See, e.g.*, Susan Rea et al., *Bias in Recording of Body Mass Index Data in the Electronic Health Record*, AMIA SUMMITS ON TRANSLATIONAL SCI. PROC. 214, 217 (2013) ("[T]he BMI on higher disease status patients was also demonstrated when comparing the frequencies of patients having particular diagnoses between subgroups having versus not having a BMI recorded."); Nicole G. Weiskopf, *Sick Patients Have More Data: The Non-Random Completeness of Electronic Health Records*, AMIA SUMMITS ON TRANSLATIONAL SCI. PROC., 1472, 1476 (2013) ("Sicker patients tend to have more complete records and healthier patients tend to have records that are less complete.").

[55] For an example of selection bias, see generally KENNETH J. ROTHMAN ET AL., MODERN EPIDEMIOLOGY 135–36 (3d ed. 2008) (explaining selection bias in the context of epidemiologic studies).

[56] Hoffman & Podgurski, *supra* note 4, at 522.

[57] *See generally* William R. Hersh et al., *Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research*, 51 MED. CARE S30, S33 (2013) ("The most commonly known problematic transformation of data occurs when data are coded, often for billing purposes").

broad range of conditions.[58] Moreover, insurance claims may not contain important information, such as detailed medical histories or treatments that are not covered by insurance.[59] Insurers who rely on billing information alone for purposes of research and analysis may thus be relying on very incomplete information.[60]

4. Lack of Data Standardization

Another data void arises from lack of data standardization and harmonization. Different EHR systems and different doctors use medical terms, phrases, acronyms, and abbreviations differently. They may use the same term to mean different things or different terms to mean the same thing. To illustrate, the abbreviation "MS" can mean "mitral stenosis," "multiple sclerosis," "morphine sulfate," or "magnesium sulfate."[61] Such inconsistencies can lead to grave difficulties in data interpretation.[62]

---

[58] See *id.* for a discussion of certain codes that indicate too broad a range of conditions.

[59] *Id.* at S32 (citing the example of hospital-acquired urinary tract infections from catheters for which Medicare will not provide reimbursement).

[60] *Id.*; Elmer V. Bernstam et al., Abstract, *Oncology Research Using Electronic Medical Record Data*, 28 J. CLINICAL ONCOLOGY e16501 (2010), *available at* http://meeting.ascopubs.org/cgi/content/abstract/28/15_suppl/e16501 ("Machine learning natural language processing techniques are more accurate than either billing data or text-word searches at identifying patients with malignancies within large data sets.").

[61] Christopher G. Chute, *Medical Concept Representation, in* MEDICAL INFORMATICS: KNOWLEDGE MANAGEMENT AND DATA MINING IN BIOMEDICINE 170 tbl.6-1 (Hsinchun Chen et al. eds., 2005).

[62] Wells, *supra* note 6, at 2 ("[T]he free text areas of the patient chart . . . are difficult to analyze quantitatively due to the breadth of human expression, grammatical errors, "the use of acronyms and abbreviations, and the potential for different interpretations of the same phrase depending on context."); Nicole Gray Weiskopf & Chunhua Weng, *Methods and Dimensions of Electronic Health Record Data Quality Assessment: Enabling Reuse for Clinical Research*, 20 J. AM. MED. INFORMATICS ASS'N 144, 147–48 (2013) (discussing terminology and dimensions of data quality).

5. Record Fragmentation

Further data inadequacies are attributable to record fragmentation. Patients see different doctors in different health care facilities that have different EHR systems.[63] If the separate EHR systems are not interoperable,[64] pieces of the patient's record will be housed in different locations and analysts may not be able to put it together into a comprehensive record that reflects the patient's full medical history.[65] In the alternative, if researchers collect information from multiple facilities and do not realize that different segments of the record belong to the same patient, they might count the same individual multiple times in their study, thus skewing their results. This is particularly likely to occur if the data that is analyzed by secondary users is de-identified in order to protect patient privacy.[66] In a February 2014 speech, Dr. Karen DeSalvo, National Coordinator for Health Information Technology, acknowledged that the health care community has "not reached . . . [its] shared vision of having . . . [a nationally] interoperable system where data can be exchanged and meaningfully used to improve care."[67]

---

[63] Hersh et al., *supra* note 57, at S31-S32.

[64] Interoperable systems can communicate with each other, exchange data, and operate seamlessly and in a coordinated fashion across organizations. BIOMEDICAL INFORMATICS: COMPUTER APPLICATIONS IN HEALTH CARE AND BIOMEDICINE 952 (Edward H. Shortliffe & James J. Cimino eds., 3d ed. 2006).

[65] Botsis et al., *supra* note 5, at 4 (stating that the EHR system that was mined for purposes of the study did not contain records of patients who were transferred to dedicated cancer centers because of the severity of their disease or who had initially been treated elsewhere).

[66] For a discussion of data de-identification, see Sharona Hoffman & Andy Podgurski, *Balancing Privacy, Autonomy and Scientific Needs in Electronic Health Records Research*, 65 SMU L. REV. 85, 104–05, 128–33 (2012).

[67] Daniel R. Verdon, *ONC's DeSalvo Issues Next Health IT Challenge: Build Interoperable EHR Systems*, MED. ECON. (Mar. 4, 2014), http://medicaleconomics.modernmedicine.com/medical-economics/news/oncs-desalvo-issues-next-health-it-challenge-build-interoperable-ehr-systems. The Office of the National Coordinator for Health Information Technology is part of the U.S. Department of Health and Human Services and is charged with promoting and facilitating the country's transition to widespread use of health information technology.

C. SOFTWARE PROBLEMS

Analysis of medical data may further be hampered by software problems. Limitations in the software's capabilities may make it difficult or impossible to extract the narrative text portions of EHRs. Software or programming flaws may also generate errors in the data contained in EHRs or in their analysis.

1. Narrative Text

EHRs are composed of structured, coded data and narrative text (also called "free-text") consisting of clinicians' notes concerning patients.[68] The narrative text often includes very important information that is not recorded elsewhere, such as the date of the condition's onset, notes concerning medication use, care summaries, and more.[69] To illustrate, coded data may indicate that the patient's asthma has worsened, but the narrative may explain that she is smoking more frequently. Unstructured narrative is often difficult to extract from EHRs because contemporary natural language processing technology is imperfect.[70]

In addition, at times, information in the free-text comments directly contradicts structured data in the EHR because of input errors.[71] For example, the structured data may indicate that one dosage was prescribed, whereas the notes state that the patient was instructed to take a different dose.[72] In such cases, analysts may not be able to determine whether the structured data or notes are correct.

---

[68] Hersh et al., *supra* note 57, at S33; Andrea L. Benin et al., *Validity of Using an Electronic Medical Record for Assessing Quality of Care in an Outpatient Setting*, 43 MED. CARE 691, 696 (2005).

[69] Hersh et al., *supra* note 57, at S33; Bayley et al., *supra* note 28, at S83.

[70] Bayley et al., *supra* note 28, at S83; Hersh et al., *supra* note 57, at S33.

[71] Dean F. Sittig & Hardeep Singh, *Defining Health Information Technology–Related Errors: New Developments since To Err is Human*, 171 ARCHIVES INTERNAL MED. 1281, 1283 (2011), *available at* http://archinte.jamanetwork.com/article.aspx?articleid=1105855.

[72] *Id.*

2. Software and Programming Defects

Software defects arising from errors in a computer program's source code or design can adversely affect both data analysis and the quality of the original data contained in EHRs. To ensure software integrity, highly skilled software professionals must carefully design and then thoroughly test their products.[73]

Software bugs can cause computer programs to produce incorrect or unexpected results or to behave in unintended ways. While subtle errors are often difficult to detect, insurance analysts and other researchers should be vigilant and examine unanticipated or egregious results to determine whether they were generated by flawed software. To illustrate, when calculating the appropriate drug dosage for a patient, the weight-based dosing algorithm may fail to convert a weight measure that was entered in pounds to a weight measure in kilograms, the unit upon which the calculation is based. In such a case, the patient would receive approximately double the correct dose.[74]

Software failures impact not only data analysis, but also the accuracy of the EHR data itself. Numerous instances of dangerous software problems have been reported. In one case, a woman's cervical cancer was not detected for four years because an EHR system's default setting displayed a prior, normal Pap smear result rather than her more recent abnormal test results. The patient, a young woman who had not yet had children, ended up needing a full hysterectomy.[75] In another case, a doctor ordered "daily" blood draws for a hospitalized patient, which conventionally means that they are performed at 6:00 a.m. Instead, however, the EHR system had been programmed to interpret the term

---

[73] Rebecca Sanders & Diane Kelly, DEALING WITH RISK IN SCIENTIFIC SOFTWARE DEVELOPMENT, 25 IEEE SOFTWARE 21, 25, 27 (2008), *available at* http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4548404; Diane F. Kelly, *A Software Chasm: Software Engineering and Scientific Computing*, 24 IEEE SOFTWARE 120, 118 (2007), *available at* http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4375255; Les Hatton, *The Chimera of Software Quality*, 40 COMPUTER 104, 104 (2007), *available at* http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4292028.

[74] Sittig & Singh, *supra* note 71, at 1283.

[75] Stacy Singer, *Electronic Medical Records May Cause Patient Care Errors, Florida Medical Board Says*, PALM BEACH POST (June 5, 2010), http://www.palmbeachpost.com/news/news/electronic-medical-records-may-cause-patient-care-/nL7Yc/.

"daily" to mean 4:00 p.m., so blood was taken in the afternoon. Because of the absence of updated blood work, the patient was given an excessive amount of the anticoagulant warfarin, which caused a serious bleeding risk, though no harm was ultimately suffered.[76] Such errors are not only potentially catastrophic for patient care, but also problematic for secondary use, because analysts may not realize that they are considering a prior year's test results or medication dosages that were prescribed in the absence of updated blood chemistry values.

IV. RECOMMENDATIONS

While contemporary medical big data suffers from many shortcomings, it remains an extremely promising resource for insurers and other researchers. Improving data quality should be a priority goal not only for doctors and patients, but also for anyone interested in secondary use. A number of measures can be implemented to enhance data accuracy and usability. First, both analysts and patients can contribute to quality assessment and improvement efforts through data audits. Second, the public and private sectors can work together to support the health care workforce, to enhance EHR automation and data extraction capabilities, and to develop best practices and training materials. Finally, a variety of federal regulations can bolster oversight efforts. These include the Meaningful Use regulations that govern EHR systems, the HIPAA Privacy and Security Rules, and the Common Rule that governs medical research.

A. DATA AUDITS

Both clinicians and secondary users of EHR data should routinely conduct data audits to assess the records' accuracy and error rates.[77]

---

[76] Megan E. Sawchuk, CTR. FOR DISEASE CONTROL WHITE PAPER, THE ESSENTIAL ROLE OF LABORATORY PROFESSIONALS: ENSURING THE SAFETY AND EFFECTIVENESS OF LABORATORY DATA IN ELECTRONIC HEALTH RECORD SYSTEMS (on file with author).

[77] Stephany N. Duda et al., *Measuring the Quality of Observational Study Data in an International HIV Research Network*, 7 PLoS ONE 1, 1 (2012), *available at* http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0033908.

Insurers already conduct data audits in order to detect fraud.[78] Data audits should also focus on general data quality because even innocent mistakes can impact insurance claims. For example, physicians' entry of incorrect dosage amounts into prescription orders can cause patients to suffer costly complications, and inadvertent selection of wrong menu items or boxes regarding the services provided can cause insurers to pay excessive reimbursement amounts.

Insurance claims data can be verified by requesting further information from providers or patients or by examining source material such as laboratory reports and pharmacy records. Other types of data in EHRs, such as diagnoses or treatment plans, may also be substantiated by inspecting source documentation from laboratories or pharmacies, or they can be cross-checked against insurance claims.[79] Experts advise that data audits focus on the following five questions:

1) Are the data complete?

2) Are the data correct?

3) Are there data inconsistencies or contradictions between different elements of the EHR or between the EHR and other source material (e.g. insurance claims)?

4) Does information seem implausible in light of other data about the patient or general scientific knowledge?

5) Is information current (e.g. was it copied and pasted without proper editing)?[80]

Auditors, who find that data is incomplete, clearly erroneous, inconsistent, implausible, or outdated, can follow up with physicians and require explanations and, where appropriate, corrections. An additional benefit of audits is their deterrent effect: clinicians who believe they are likely to be audited may be more cautious about EHR data entry.

Patients themselves can become active partners in efforts to enhance data quality. The HIPAA Privacy Rule furnishes patients with a right to inspect or obtain copies of their records and to request amendments if they detect mistakes.[81] In order to balance patients' rights and providers' needs, the Rule allows healthcare providers to charge "reasonable, cost-

---

[78] Tammy Worth, *Spike in Retrospective Audits: But Industry Insiders Dispute Any Abnormalities*, HEALTHCARE FIN. NEWS (June 1, 2013), http://www.healthcarefinancenews.com/news/spike-retrospective-audits.

[79] Duda et al., *supra* note 77, at 2.

[80] Weiskopf & Weng, *supra* note 62, at 145.

[81] 45 C.F.R. §§ 164.524–.526 (2013).

based" fees for copies of records[82] and to deny requests for amendment on valid grounds, such as a determination that no mistake exists.[83] In addition, providers need only note the amendment once and then supply a link to the amendment's location in other parts of the record that are affected by the change.[84] If patients more regularly scrutinize their records and ask for corrections, they could add an important layer of data quality oversight without over-burdening their physicians.

B. WORKFORCE AND TECHNICAL SOLUTIONS

Changes in workforce practices and technology can go far to alleviate the problem of inadequate data quality. Among these potential tools are the use of scribes, enhanced automation, improved natural language processing, and the creation of best practices guidelines and training programs.

1. Scribes

One approach that is favored by some clinicians is the use of scribes.[85] Scribes shadow physicians and do the work of entering data into the EHR while the doctor examines the patient. Thus, documentation is accomplished by a professional who is devoting all of her attention to the data-entry task.[86] Scribes, who reportedly numbered approximately 10,000 in early 2014, can be hired through companies such as PhysAssist and ScribeAmerica, which provide them with pre-employment training.[87] While some worry about patient privacy and the cost of hiring scribes, other

---

[82] 45 C.F.R. § 164.524(c)(4) (2013).

[83] 45 C.F.R. § 164.526(a)(2) (2013).

[84] § 164.526(c)(1)).

[85] Katie Hafner, *A Busy Doctor's Right Hand, Ever Ready to Type*, N.Y. TIMES (Jan. 12, 2014), http://www.nytimes.com/2014/01/14/health/a-busy-doctors-right-hand-ever-ready-to-type.html?_r=0; Scott A. Shipman & Christine A. Sinsky, *Expanding Primary Care Capacity by Reducing Waste and Improving the Efficiency of Care*, 32 HEALTH AFF. 1990, 1993 (2013).

[86] Hafner, *supra* note 85.

[87] *See* PhysAssist Scribes, http://www.iamscribe.com/index.php (last visited Oct. 15, 2014); ScribeAmerica, https://www.scribeamerica.com/ (last visited Oct. 15, 2014).

physicians have found that scribes significantly improve their work quality and, consequently, job satisfaction.[88]

2. Automation

Advances in technology are also likely to enhance data accuracy and completeness. Some medical devices that collect patient data could automatically transmit measurements to EHRs without requiring human intermediaries who might mistype information or make other mistakes. Examples are devices that measure vital signs, such as blood pressure, pulse, oxygen rates, and temperature.[89] In addition, voice recognition software that is of high quality could reduce the risk of typos and promote the inclusion of more details in EHRs because documentation by dictation rather than by typing would take less time.[90]

EHRs could further be programmed to generate alerts if implausible or clearly erroneous data is entered.[91] In one study focusing on height and weight measures, researchers had the EHR alert clinicians if they entered figures that deviated by ten percent or more from height and weight measurements that were previously recorded.[92] Thus, for example, if a patient's weight was recorded as being 150 pounds in one visit and 190 pounds three months later, a message would ask the clinician to check the two entries because it is unlikely that the patient gained forty pounds in

---

[88] Hafner, *supra* note 85.

[89] ECRI Institute, *Making Connections*, HEALTH DEVICES 102, 104 (2012), *available at* https://www.ecri.org/Documents/HIT/Making_Connections_Integrating_Medical_Devices_with_Electronic_Medical_Records(Health_Devices_Journal).pdf; *Partners HealthCare and Center for Connected Health Launch Personal Health Technology Platform to Improve Care Delivery*, PARTNERS HEALTHCARE (June 20, 2013), http://www.partners.org/About/MediaCenter/Articles/Partners-Center-for-Connected-Health-Technology-Platform.aspx.

[90] Robert Hoyt & Ann Yoshihashi, *Lessons Learned from Implementation of Voice Recognition for Documentation in the Military Electronic Health Record System*, 7 PERSP. HEALTH INFO. MGMT. 1, 1 (2010), *available at* http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2805557/.

[91] Krystl Haerian et al., *Use of Clinical Alerting to Improve the Collection of Clinical Research Data*, 2009 AMIA SYMP. PROC. 218, 219–20, *available at* http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2815392/pdf/amia-f2009-218.pdf.

[92] *Id.* at 219.

such a short period of time. The researchers observed that after the alerts were implemented, EHR error rates fell from 2.4% to .9%.[93]

3. Natural Language Processing

For purposes of secondary use of medical data, improved natural language processing (NLP) tools would be particularly useful. NLP tools would enable analysts to extract more comprehensive data from EHRs, including information such as medical history and progress notes contained only in the narrative text portion of the record.[94] While applications such as the Electronic Medical Record Search Engine (EMERSE)[95] have long been available, experts note that NLP capabilities are "still far from perfect"[96] and leave much room for improvement.

4. Best Practices Standards and Training Programs

EHR users would benefit greatly from best practices standards and training programs concerning appropriate and efficient data entry practices. Best practices guidelines and training programs could be developed cooperatively by vendors, government experts, and health care providers' professional organizations.[97] These resources should help users formulate strategies to enhance EHR accuracy and completeness, with special attention paid to the most pervasive challenges, such as copy and paste features.

C. FEDERAL REGULATIONS

Another critical component of efforts to improve EHR data quality is federal regulation. While many in today's political climate are loath to impose regulatory constraints upon the free market, regulatory interventions have long been customary in the very complex and critically

---

[93] *Id.* at 220.

[94] Bayley et al., *supra* note 28, at S83.

[95] David A. Hanauer, *EMERSE: The Electronic Medical Record Search Engine*, 2006 AMIA ANNU. SYMP. PROC., 941, 941, *available at* http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1839699/pdf/AMIA2006_0941.pdf.

[96] Hersh et al., *supra* note 57, at S33.

[97] AM. HEALTH INFO MGMT. ASS'N, *supra* note 41, at 2–3.

important realm of health care. Good data quality can be considered a "positive externality" because those responsible for it, namely vendors and clinicians, do not reap all the benefits of high EHR quality.[98] Rather, third parties such as patients, insurers, researchers, and others have much to gain from data accuracy and comprehensiveness as well. Because the public's interest is at stake, the government is justified in intervening to induce those who produce and use EHR systems to meet high quality standards. In addition, because the federal government covers over thirty percent of American patients through Medicare, Medicaid, and the Children's Health Insurance Program,[99] it has a direct interest in ensuring that providers do not submit erroneous claims. The federal government could pursue at least three well-established regulatory avenues to address data quality problems: the Meaningful Use Regulations, the HIPAA Security Rule, and the Common Rule.

1. Meaningful Use Regulations

The Meaningful Use regulations, issued by the Centers for Medicare and Medicaid Services (CMS), govern providers' use of EHR systems.[100] The regulations, which are being rolled out in three phases, establish what health care providers need to do in order to demonstrate that they are meaningful users of EHR systems and thus are eligible for government incentive payments for adoption of the systems.[101] The Meaningful Use regulations could be harnessed to promote interoperability, data harmonization, and routine data audits.

---

[98] Abigail McWilliams et al., *Guest Editors' Introduction Corporate Social Responsibility: Strategic Implications*, 43 J. MGMT. STUD. 1, 9 (2006) (defining "externality" as "the impact of an economic agent's actions on the well-being of a bystander" and citing innovation as an example of a positive externality because of its general social benefits).

[99] THE HENRY J. KAISER FAMILY FOUND., *supra* note 8.

[100] Sharona Hoffman & Andy Podgurski, *Meaningful Use and Certification of Health Information Technology: What about Safety?*, 39 J. L. MED. & ETHICS 77, 78 (2011); 42 C.F.R. §§ 495.2–495.370 (2013).

[101] Hoffman & Podgurski, *supra* note 100, at 78. President Obama's stimulus legislation, the American Recovery and Reinvestment Act of 2009, "provides for payments of up to $44,000 per clinician under the Medicare incentive program and $63,750 per clinician under the Medicaid program." *Id.* at 77.

The current stage of Meaningful Use regulations, stage 2, begins to address interoperability and data standardization. The regulations require health care providers who transition patients to different care settings (e.g. from a hospital to a rehabilitation center) or refer them to other doctors to transmit electronically to the next provider a certain percentage of their summary of care documents. In addition, providers must submit data to immunization registries and furnish syndromic surveillance information to public health authorities.[102] At the same time, EHR certification regulations require vendors to build data portability capabilities into EHR systems that will enable clinicians to meet these Meaningful Use standards.[103] Such data exchanges necessitate some degree of interoperability and data standardization so that the recipients can receive and understand the submitted health information.

Stage 3 regulations are under development and will take effect in 2017.[104] These regulations should focus to a greater extent on interoperability and data harmonization so that documentation can always be exchanged among healthcare providers with different EHR systems and understood by them.[105] Patient records should not be irreparably fragmented among different physician practices and hospitals, and terms or acronyms such as "MS" should not mean different things in different EHRs. Just as drivers can look at most car dashboards and have little difficulty reading all of the instruments and displays, clinicians who have

---

[102] 42 C.F.R. §§ 495.6(e)(8)–(10) (2013); *see also Stage 2 Eligible Professional (EP) Meaningful Use Core and Menu Measures Table of Contents*, CTR. FOR MEDICARE & MEDICAID SERV. (Oct. 2012), http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Downloads/Stage2_MeaningfulUseSpecSheet_TableContents_EPs.pdf.

[103] *See* 45 C.F.R. §§ 170.314(b), (f) (2014) (addressing care coordination and public health).

[104] Robert Tagalicod & Jacob Reider, *Progress on Adoption of Electronic Health Records*, CTR. FOR MEDICARE & MEDICAID SERV. (Dec. 13, 2013, 12:41 PM), http://www.cms.gov/eHealth/ListServ_Stage3Implementation.html.

[105] Anthony Brino, *Senators Press for EHR Interoperability*, HEALTHCARE IT NEWS (Jan. 6, 2014), http://www.healthcareitnews.com/news/senators-press-ehr-interoperability (reporting that House and Senate bills call upon the Department of Health and Human Services "to adopt a common interoperability standard by 2017, as part of the rules for meaningful use Stage 3"); Verdon, *supra* note 67 (reporting that Dr. Karen DeSalvo, National Coordinator for Health Information Technology, has declared that interoperability will be a national priority).

been trained on one EHR system should be able to navigate and operate other EHRs.

Furthermore, CMS would be wise to consider incorporating requirements for periodic data audits into future Meaningful Use regulations. Providers could be instructed to conduct audits in order to verify that they do not have an unacceptably high error rate and to assess mechanisms to improve data accuracy and completeness.

2. The HIPAA Privacy and Security Rules

Several provisions of the HIPAA Privacy and Security Rules could serve as additional tools to improve data quality. As already noted, the HIPAA Privacy Rule empowers patients to review their EHRs and to request corrections if they detect errors.[106] In addition, the HIPAA Security Rule's General Requirements section states that covered entities bear responsibility for ensuring "the confidentiality, integrity, and availability" of electronic health information that they create, receive, maintain, or transmit.[107] The term "integrity" should be interpreted broadly to include data quality.

The regulations detail a variety of enforcement mechanisms, including investigation, corrective action mandates, and penalties.[108] The Department of Health and Human Services' Office of Civil Rights ("OCR") is authorized to investigate complaints of HIPAA violations filed by complaining parties and to initiate its own investigations as well.[109] To that end, OCR has launched an audit program.[110] The issue of data quality

---

[106] 45 C.F.R. §§ 164.524–.526 (2013).

[107] 45 C.F.R. § 164.306(a)(1) (2013). The HIPAA Security Rule covers health plans, health care clearinghouses, and health care providers who transmit health information electronically, and their business associates. 45 C.F.R. § 164.104(a)(1)–(3) (2013).

[108] 45 C.F.R. §§ 160.300–.426 (2013).

[109] 45 C.F.R. §§ 160.306–.308 (2013); *How OCR Enforces the HIPAA Privacy and Security Rules*, U.S. DEPARTMENT OF HEALTH & HUM. SERVICES, http://www.hhs.gov/ocr/privacy/hipaa/enforcement/process/howocrenforces.html (last visited Oct. 6, 2014).

[110] *Audit Program Protocol*, U.S. DEPARTMENT OF HEALTH & HUM. SERVICES, http://www.hhs.gov/ocr/privacy/hipaa/enforcement/audit/protocol.html (last visited Oct. 6, 2014); Patrick Ouellette, *OCR Readies Pre-Audit Survey for HIPAA Covered Entities, BAs*, HEALTHITSECURITY.COM (Feb. 25, 2014),

should be among OCR's areas of focus during audits, and the agency should require covered entities to demonstrate that they have implemented measures to verify and improve data quality.

Furthermore, ensuring that patients have access to their records and that patients can have mistakes corrected in their EHRs should be enforcement priorities for OCR. In a March 31, 2014 report, OCR indicated that patients' lack of access to their health information was the third most frequently investigated complaint.[111] Failure to amend records in response to legitimate requests for correction is not listed among the top five complaints, but it is not clear if this is because providers generally comply with the requests or because patients do not submit such requests frequently.[112] OCR has been criticized for not being aggressive enough in its enforcement activities.[113] Experts, however, note that the agency's oversight efforts have been intensifying recently.[114] One hopes that this trend will continue and that government enforcement will be an important component of the data quality enhancement toolkit.

3. The Common Rule

The federal research regulations, known as the Common Rule,[115] can further incentivize physicians to be vigilant about the accuracy and completeness of their EHRs. Many physicians are also researchers,[116] and

---

http://healthitsecurity.com/2014/02/25/ocr-readies-pre-audit-survey-for-hipaa-covered-entities-bas/.

[111] *Enforcement Highlights*, U.S. DEPARTMENT OF HEALTH & HUM. SERVICES (Mar. 31, 2014), http://www.hhs.gov/ocr/privacy/hipaa/ enforcement/highlights/. The report covers the period of April 2003 (the HIPAA Privacy Rule's effective date) through March 2014. *Id*.

[112] *Id.*

[113] *See* Alaap B. Shah & Ali Lakhani, *OCR Lacks Insight into HIPAA Security Rule Compliance*, BLOOMBERG BNA (Feb. 21, 2014), http://www.bna.com/ocr-lacks-insight-into-hipaa-security-rule-compliance/. ("[O]CR's report card, although somewhat changed, is not materially improved since the OIG's 2011 report wherein a 'need for greater OCR oversight and enforcement' was recommended.").

[114] *Id.*

[115] 45 C.F.R. §§ 46.101–.505 (2013).

[116] *See generally* Acad. of Physicians in Clinical Research, *About APCR*, APCRNET.ORG, http://www.apcrnet.org/FunctionalMenuCategory/AboutAPCR.aspx (last visited Oct. 6, 2014).

some of the research projects that they conduct are observational studies that involve review of medical records.[117]

Research involving identifiable patient information[118] is subject to oversight by institutional review boards (IRB) pursuant to detailed Common Rule guidance.[119] The regulations specify the criteria for IRB approval of studies that are governed by the regulations.[120] Several provisions address data collection, requiring IRBs to consider how researchers plan to monitor data to ensure the safety of participants and to protect their privacy.[121] An additional approval criterion should be added to the regulations: a requirement that investigators who will collect data from EHRs indicate in their research protocols what steps they will take to monitor data quality. A mandate that researchers conduct regular data audits or otherwise double-check information contained in EHRs could enhance the reliability of research findings. In addition, it may induce clinicians who are themselves researchers or are sensitive to the needs of researchers to be more careful about EHR data input.

---

[117] 45 C.F.R. § 46.102(f) (2013) (explaining that research covered by the Common Rule can be conducted in two ways: (1) intervention or interaction with individuals or (2) study of "identifiable private information.")

[118] *Id.* (indicating that the regulations cover "[p]rivate information … that is individually identifiable (i.e., the identity of the subject is or may readily be ascertained by the investigator or associated with the information.") Thus, by contrast, record-based studies that use only de-identified information are exempt from the federal research regulations and IRB approval.)

[119] 45 C.F.R. §§ 46.107–.109 (2013) (addressing IRB membership, functions and operations, and review of research. According to the U.S. Food and Drug Administration, an IRB is "an appropriately constituted group that has been formally designated to review and monitor biomedical research involving human subjects" with "authority to approve, require modifications in (to secure approval), or disapprove research." *Institutional Review Boards Frequently Asked Questions — Information Sheet*, U.S. FOOD & DRUG ADMIN., http://www.fda.gov/regulatoryinformation/guidances/ucm126420.htm (last updated June 25, 2014). IRB review is conducted in order to protect "the rights and welfare of human research subjects"). *Id.*

[120] 45 C.F.R. § 46.111 (2013).

[121] 45 C.F.R. § 46.111(a)(6), (7) (2013).

V. CONCLUSION

Medical big data is a growing resource for insurance analysts and other researchers. Yet, EHR data is often significantly flawed and deficient. EHR data quality inadequacies are particularly troubling in the insurance realm because they can cause insurers to pay excessive or inappropriate claims reimbursement amounts. This, in turn, can generate premium increases for consumers or a squandering of taxpayer money in the case of public programs such as Medicare. Moreover, incorrect EHR data that is put to secondary uses can lead to erroneous inferences and poor insurance coverage or other health-related policies. Consequently, it is critical that vendors, health care providers, and government authorities aggressively attack the challenges of data quality. Solutions must be formulated by all stakeholders, not least of which is the government. It is only with significant improvements that the great potential of medical big data can be realized.