Diabetes and Asthma Case Identification, Validation, and Representativeness When Using Electronic Health Data to Construct Registries for Comparative Effectiveness and Epidemiologic Research

Jay R. Desai, MPH, Pingsheng Wu, PhD,† Greg A. Nichols, PhD,‡ Tracy A. Lieu, MD, MPH,§|| and Patrick J. O'Connor, MD, MPH**

Background: Advances in health information technology and widespread use of electronic health data offer new opportunities for development of large scale multisite disease-specific patient registries. Such registries use existing data, can be constructed at relatively low cost, include large numbers of patients, and once created can be used to address many issues with a short time between posing a question and obtaining an answer. Potential applications include comparative effectiveness research, public health surveillance, mapping and improving quality of clinical care, and others.

Objective and Discussion: This paper describes selected conceptual and practical challenges related to development of multisite diabetes and asthma registries, including development of case definitions, validation of case identification methods, variation in electronic health data sources; representativeness of registry populations, including the impact of attrition. Specific challenges are illustrated with data from actual registries.

Key Words: registries, comparative effectiveness research, surveillance, diabetes, asthma

(*Med Care* 2012;50: S30–S35)

Patient registries are defined as "an organized system that uses observational study methods to collect uniform data (clinical and other) to evaluate specified outcomes for a population defined by a particular disease, condition, or exposure, and that serves a predetermined scientific, clinical, or policy purpose."[1] Large administrative datasets and regional and national patient registries have been used for decades to support clinical research, public health surveillance, quality improvement, and accountability activities in the United States and other countries.[2–9] Advances in health information technology and widespread implementation of electronic health data (EHD), including increased access to clinical data, offer new opportunities to establish large disease-specific registries, cohorts, and databases, especially for managed care systems, large integrated health systems, large medical groups with sophisticated EHD systems, and large public and private insurers.[10–13]

Because such registries use existing data, they can be constructed at relatively low cost, include large numbers of patients, and once constructed, be used to address many issues with a short time between posing a question and obtaining an answer. Potential applications include: (a) comparative effectiveness research (CER) that assesses the relative effectiveness of various treatments and systems of care in defined patient populations; (b) filling knowledge gaps that are ethically and/or logistically not feasible for randomized controlled studies, or that require large samples like studies among specific subpopulations or studies of rare conditions; (c) public health surveillance of diseases, risk factors, patterns of care, undesirable side effects of care, and quality of care; and (d) for identifying patient and care factors associated with important clinical outcomes or with resource use.[14–18]

Essential tasks in developing and applying results from any health registry or research study include identifying cases and understanding the population from which the cases are derived. In this article, we focus on EHD case identification and validation, the influence of different EHD data sources, and population representativeness, including issues of attrition. Our discussion is drawn from ongoing experiences in developing 2 large multisite disease-specific registries: (a) The SUPREME Diabetes DataLink (DataLink) includes 12 sites with 1.2 million persons with diabetes; and (b) the Population-based Effectiveness in Asthma and Lung Diseases (PEAL) Network, which has identified patients with lung diseases, including almost 1.4 million patients with asthma at 4 Health Maintenance Organization Research

From the *HealthPartners Research Foundation, Minneapolis, MN; †Department of Medicine, Vanderbilt University School of Medicine, TN; ‡Kaiser Permanente Northwest, OR; §Department of Population Medicine, Harvard Medical School, Harvard Pilgrim Health Care Institute; and ||Division of General Pediatrics, Children's Hospital Boston, Boston, MA.

Funding was provided by a contract from AcademyHealth. Additional funding was provided by AHRQ 1R01HS19859 [SUrveillance, PREvention, and ManagEment of Diabetes Mellitus (SUPREME-DM)] and AHRQ 1R01HS19669 (The Population-Based Effectiveness in Asthma and Lung Disease).

The authors declare no conflict of interest.

Reprints: Jay R. Desai, MPH, HealthPartners Research Foundation, 8170 33rd Avenue South, Bloomington, MN 55425. E-mail: jay.r.desai@healthpartners.com.

Copyright © 2012 by Lippincott Williams & Wilkins

ISSN: 0025-7079/12/5007-0S30

Network (HMORN) sites plus Tennessee's TennCare (Medicaid) population. The DataLink and PEAL Network are 2 initiatives that build on the data resources of the HMORN, a consortium of 16 private integrated health care systems across the United States and in Israel comprising of over 11 million enrollees.[13] Both projects receive funding from the Agency for Healthcare Research and Quality with an emphasis on constructing multisite EHD-linked registries for CER. Issues of governance, confidentiality, protection of human subjects, and technical issues like data architecture, integration, consolidation, cleaning, storage, and management are clearly key considerations in the creation and use of registries for CER but they are beyond the scope of this discussion.[1,19–21]

CASE DEFINITIONS FOR DIABETES AND ASTHMA

Establishing valid case definitions to identify those with diabetes or asthma is a fundamental task in registry construction. However, the task is complicated by the fact that there is often no feasible gold standard for assessing validity of case definitions derived from EHD. Numerous epidemiological studies have tried to assess and improve on case definitions for diabetes and asthma with self-report or medical record review as criterion measures.[22–30] Diabetes validation studies have found relatively good sensitivity (76%–97%), specificity (95%–100%), and predictive positive values (PPV) up to 94% using various EHD algorithms compared with the medical record.[24,25] Accuracy of pediatric asthma (sensitivity: 44%–92%; specificity: 80%–94%) and adult asthma (sensitivity: 68%–95%; specificity: 59%–90%) are more variable, complicated by variation in severity and the remission and relapse of asthma in patients.[23,31] Verification of asthma at a single point in time based on biological measures is difficult because those with asthma may have normal pulmonary function off medications at many points in time and spirometry or other validated pulmonary function testing is often not done.[32] Although diagnostic standards for diabetes are clearly linked to glucose and glycated hemoglobin (A1c) thresholds, diabetes patients receiving active lifestyle or pharmacological treatment may have normal glucose and A1c levels when tested. Unlike university-based research registries that often only include patients with advanced disease states who meet rigorous criteria for diagnosis of a specific condition, EHD-based case definitions often rely heavily upon diagnosis codes, routinely done laboratory tests, and pharmacy data.

Lack of a rigorous biometric gold standard for classification of diabetes or asthma status using routinely collected clinical data makes it difficult or impossible to precisely, determine the sensitivity, specificity, and PPV of proposed case identification methods. However, we can approach this conceptually and then practically. Ideally, we want to maximize both sensitivity and specificity, with the latter leading to improved PPV by reducing false positives. Unfortunately, there is typically a trade-off between sensitivity and specificity and we have to decide which is more important in chronic disease case identification. Furthermore, PPV is driven by the prevalence of a condition, the lower the prevalence, such as with diabetes and asthma, the lower the PPV even with sensitivity and specificity above 90%.[33] Registries designed to support clinical management, patient outreach, or practical randomized interventions require high certainty that identified cases truly have the condition of interest and, therefore, would be designed to maximize PPV. This can be done by using a restrictive case definition but often at the expense of lowering sensitivity (the likelihood that all those with the condition of interest will actually be identified). For example, requiring ≥ 3 diabetes diagnosis codes would usually assure relatively high PPV but likely miss many cases. In contrast, requiring only 1 diabetes diagnostic code for diabetes will increase sensitivity but will very likely include many false positives and yield lower PPV.[25] Although registries to guide clinical interventions want to maximize PPV, registries to support public health surveillance, population-level monitoring of care quality, accountability, and health services research, or to support CER studies may often be better designed to maximize sensitivity from the outset with refinement as needed. In many cases, such as public health surveillance, it is important to capture the vast majority of cases consistently over time, despite including more false positives.

The approach taken by both the Diabetes DataLink and the PEAL Network was to initially maximize sensitivity by using broad case definitions to capture all persons with any indication of diabetes or asthma. Such an approach allows flexibility in tailoring case definitions to particular studies by subsequently applying more stringent criteria. For example, if the high sensitivity registry includes all those with ≥ 1 diabetes diagnosis codes, PPV could be increased by identifying a patient subset based on ≥ 2 codes. Note that if a registry attempts to maximize PPV initially, for example, by requiring ≥ 2 diagnosis codes, it is not possible to extract from this registry all those with only ≥ 1 diabetes codes.

Therefore, the DataLink initially captured patients with any indication of diabetes on the basis of diagnosis codes, diabetes-specific pharmaceutical use, or elevated glucose, or glycated hemoglobin (A1c) values. This was further refined to identify prevalent diabetes cases as patients who met at least 1 of the following criteria within an 18-month period: 1+ diabetes-specific medication, 1+ inpatient diabetes diagnosis, 2+ face-to-face outpatient diabetes diagnoses on separate days, or 2+ elevated blood glucose values performed on separate days or 1 elevated oral glucose tolerance test. This method is based on prior studies and is compatible with the then-current diabetes definition used by the National Committee on Quality Assurance (ie, Healthcare Effectiveness Data and Information Set).[3,24,25] The current DataLink case definition is unable to distinguish type 1 and type 2 diabetes. However, it does capture persons with undiagnosed diabetes who meet blood glucose laboratory criteria but have no diagnosis or treatment. As expected, patients who may have clinical diabetes but do not have a diagnosis, take no glucose-lowering pharmacotherapy, and have no laboratory glucose or A1c tests would be missed.

The PEAL Network has identified an initial set of patients with lung diseases, including asthma, based on International Classification of Diseases, 9th revision (ICD-9) codes. At the start of each study, more specific case definitions will be employed. For example, a few studies with TennCare data identify pediatric asthma cases using an

© 2012 Lippincott Williams & Wilkins

established modified Healthcare Effectiveness Data and Information Set algorithm incorporating diagnostic coding and asthma-specific medication use.[23] Criteria include: (a) at least 1 inpatient, outpatient, or emergency department claim listing asthma (ICD-9 493.xx) in any discharge diagnosis fields during an 18-month period, a time frame intended to capture at least 1 well-visit for minimally symptomatic children presenting once a year for routine preventive care to their physician[23,34,35]; or (b) 2+ prescriptions for any short-acting β-agonist or 1+ prescriptions for any other asthma medication (inhaled anti-inflammatory, long-acting β-agonist, leukotriene modifying drug) during an 18-month period will be considered to have asthma. To further increase specificity of the definition, asthma diagnosis is limited to children of school age because "transient early wheezing" instead of asthma is often diagnosed at younger ages.

Both the diabetes and asthma definitions represent base case definitions that may be further refined or restricted based on the desired balance between sensitivity, specificity, and PPV. Most CER studies start with an inclusive base case definition and then apply additional inclusion and exclusion criteria. For example, a CER study comparing 2 treatments head-to-head in real-world settings, may exclude from analysis those with any contraindication to either one of the comparators. If metformin and sulfonylurea are to be compared, those for whom either of the treatments is contraindicated (sulfa allergy, renal dysfunction, congestive heart failure, others) would be excluded before analysis to enable a fair comparison of the 2 treatments on clinical outcomes such as cardiovascular events or mortality.

To address concerns about validity, there are several ways to build confidence in EHD-based case definitions. One approach used within both the Diabetes DataLink and the PEAL Network is to extend the time frame used to identify or confirm cases. In developing a diabetes case definition, the DataLink relies on previous EHD work that found 18-month–2-year time frames were reasonable to allow for at least 2 diabetes outpatient visits or 2 elevated blood glucose labs, the latter being current diagnostic practice.[6,24,25] A similar rationale is applied in the TennCare with an 18-month window for case identification.[34,35] However, it may be worthwhile for investigators to vary these time periods and assess impact on case capture rates, recognizing that shorter time periods often identify fewer cases but they are more likely to be "true" cases, but may also be more severe cases.

Another use of time is to periodically reapply the case definition to a patient population every few months or else to use a rolling identification process. This may increase sensitivity by capturing those with mild asthma that only relapses periodically, or picking up initially euglycemic diet-controlled diabetes patients as they drift into higher A1c ranges or require pharmacotherapy over a longer period of time.[23,27] In the DataLink, a rolling case identification process is applied over a multiyear period, resulting in a dynamic cumulative prevalence cohort as more and more years of EHD become available. This increases case capture relative to the traditional static approach where a defined calendar time period is used to identify cases resulting in a unique cohort followed over time without adding new cases. The time frame and process used for establishing the cohort are important because they can yield different results. For example, in 1 DataLink health system, the prevalence from identifying cases within only a 2-year period (2008–2009) is 7.4%. If the rolling cumulative approach is used from 2000 through 2009, the prevalence is 8.2%. With the rolling approach, static cohorts or serial cross-sectional samples of persons with diabetes can still be created.

Another important strategy to improve capture of "true" cases is to adopt more restrictive case definitions. For example, instead of requiring at least 2 outpatient visits with a diabetes diagnosis in 2 years, one might require 3 such visits during that same time frame. Applying more stringent case criteria increases PPV, whereas using less stringent criteria maximizes sensitivity, as discussed earlier.

Weiner et al[36] suggest a third strategy—using a Bayesian inference engine to assign confidence probabilities to case identification. A conceptually similar but computationally simpler approach is to determine how many of a set of diabetes or asthma case definition criteria or data sources are met within specified time intervals. The more met, the greater the confidence that a "true" case has been identified. For example, in the DataLink the primary data sources for case identification are diagnosis codes, pharmacy use, and laboratory values. In 1 DataLink system, over a 2-year period (2008–2009) 18.1% of cases were identified based on meeting criteria from only 1 source, 36.2% were captured by independently meeting criteria from 2 sources, whereas 45.6% independently met criteria from all 3 data sources. This might suggest more case confidence for those meeting multiple qualifying criteria compared with only 1 qualifying criteria. Alternatively, a higher confidence probability could be placed on the 65.1% of cases qualifying through laboratory values.

Such methods can be used to tailor the performance characteristics of a registry to the purpose for which it is used, and thus increases confidence in specific case definition criteria, which may vary across registry applications. These strategies have been used in developing the DataLink and PEAL Networks, but there is much that remains to be done to better understand and quantify the impact of case definition strategies on relative sensitivity, specificity and PPV. Furthermore, applying a range of case identification strategies from broad to more restrictive may help assess the robustness of study results.

VARIATION IN REGISTRY MEMBERS RELATED TO SOURCES OF DATA

Data sources available to those who are constructing a registry may influence the number and the characteristics of registry members, and have far-reaching consequences when registries are used for certain purposes.[37] When registries based on different sources of data are used to compare quality of care or assess patterns of care across care systems, bias may be introduced into those comparisons. Claims-based diagnoses (ie, outpatient and inpatient) are the most commonly available EHD, followed by pharmacy claims and laboratory data. In the TennCare population, 49% of child asthma cases were identified by asthma-specific medication claims only, 12% by claims-based diagnoses codes only, and

© 2012 Lippincott Williams & Wilkins

the remainder (39%) by both data sources. Table 1 examines the capture of prevalent diabetes cases at 1 Datalink Health System. During a 2-year period (2008–2009), 91.4% of cases were identified based on ≥2 outpatient claim-based diagnosis codes. The sequential addition of cases based on different data sources yielded 0.8% from ≥1 inpatient claims diagnosis codes, 4.8% from ≥1 diabetes-specific medication claim, and 3.0% from at least 2 elevated blood glucose labs. This suggests that available electronic outpatient and pharmacy claims information identify virtually all diagnosed diabetes cases, with few new cases added through inpatient claims. Table 1 also suggests that cases identified solely through 1 source may have different characteristics. Those captured only through inpatient diagnoses tend to be younger women with low A1c levels suggesting pregnancy-related identification, even with the deliberate exclusion of pregnancy-related diabetes diagnoses codes. Pharmacy-only cases also tend to be younger but in poorer health (some may be on diabetes medications for polycystic ovarian syndrome), whereas it is likely that those identified only by laboratory data are those with undiagnosed and untreated diabetes—a group of special interest in some studies—as their A1c levels are lower. Interestingly, the percent meeting diabetes recommended that low density lipoprotein (LDL-c) and blood pressure level thresholds are lower as might be expected as they are not being managed to these levels.

Variation in diagnosis coding or in data entry practices can result in missing information and impact case identification, thus biasing comparisons across clinics, medical groups, or care systems. Currently, there is considerable variation in diagnostic coding practices within and across care systems.[38] Some experts believe that broader use of electronic medical records (EMR) will likely improve standardization and comparability of coding practices because primary care provider coding patterns can be easily compared and outliers identified and corrected.[38] However, issues of omission and commission may persist. Diagnosis and medications may not be captured in structured formats but only in text notes. The increased use of natural language processing technologies may improve this by identifying key information (ie, diagnoses and medications) from notes and translating it into a structured format.[39] Another concern is technology or databases not effectively interfacing with EMRs such as with point-of-care desktop analyzers not being integrated into EMRs or laboratory databases. Dual or supplemental insurance coverage may also vary across demographic subgroups, care systems, or time, and may result in incomplete outpatient, inpatient, or pharmacy claims data. EMR-derived data may also be incomplete if a subject receives care from different systems. External factors such as changes in diabetes or asthma diagnostic criteria and care recommendations, reimbursement practices, or the increasing attention to accountability measures may all influence data accuracy and completeness in unanticipated ways and may vary locally, nationally, and over time. There is currently no systematic approach to monitoring such variation or its impact on EHD quality. The creation of large nationally representative registries provides the opportunity to examine these issues and variation in greater depth.

EXTERNAL VALIDITY OR POPULATION REPRESENTATIVENESS

The enrolled populations of health systems in the DataLink and PEAL Network are defined by insurance coverage, geographical scope, and, for diabetes, receiving care from medical groups that use EMRs. Although both registries include robust samples of those with Medicaid, Medicare, and commercial health insurance, subjects with no health insurance may be underrepresented. In recent years the uninsured 0 to 64 years of age were 6% in Minnesota and 21% in Tennessee.[40] However, for CER studies, the important question is whether populations receiving health care services are adequately represented. A CER study comparing 2 different diabetes education strategies in a clinic serving predominantly higher income patients is likely to have different results than the same study done at a federally qualified community health center. Large registries such as DataLink and PEAL provide excellent opportunities for a wide range of CER studies. They have substantial age, race, ethnic, and socioeconomic diversity, and the members of these registries receive care from many providers and health systems in the region. Furthermore, if case identification is based on EMR data many uninsured patients may also be included.

Population representativeness is also affected by the inevitable attrition with time in any registry or traditional study cohort. Potential consequences, such as loss of study power, bias, or generalizability, may occur but are also influenced by specific study designs and aims. Fewtrell et al[41] have suggested minimum reporting requirements to address attrition. These include providing clarity on the flow of participants at each stage of the cohort, the ability of the final sample size to detect the hypothesized outcome effect, the potential for bias, generalizability, and appropriate sensitivity analyses.[41]

In the Diabetes DataLink and PEAL Network the main source of attrition is disenrollment because of changes in insurance coverage. This may be moderated in integrated health systems or Accountable Care Organizations, and even with changes in coverage individuals may remain in the same provider system, particularly older patients with chronic conditions. Therefore, their EHD may often remain available. Once again using a single DataLink system as an example, among the 2004 prevalent diabetes cohort, there was 47.2% attrition by 2010. Of this, an absolute 9.7% died and an absolute 37.5% disenrolled. In a 2006 incident cohort, those lost because of disenrollment tended to be younger (18–44 y of age) and have worse HbA1c and LDL-c control at baseline. Various analytic strategies can be used to compensate for bias introduced by attrition, but a full discussion of these is beyond the scope of this paper.[41,42]

SUMMARY AND CONCLUSIONS

Registries based on administrative data available in Medicaid and Medicare are useful for tracking prevalence, health care utilization, and costs in some of our most vulnerable populations. However, many registries now combine administrative data with more detailed clinical data including vital signs, laboratory tests, and even the verbal data from office notes. The potential applications of such registries for CER, surveillance of

© 2012 Lippincott Williams & Wilkins

TABLE 1. Sequential Identification and Select Characteristics of Prevalent Diabetes Cases Based on Qualifying Criteria From January 1, 2008 Through December 31, 2009 at 1 Participating DataLink Site

	Case Definitions[†]						
	A	B	C	D	E	F	G
Population Characteristics*	Outpatient Diagnosis Codes	Inpatient Diagnosis Codes Only	Outpatient and Inpatient Diagnosis Code	Additional Pharmacy Claims Only	Diagnosis Codes + Pharmacy Claims	Laboratory Only	All Claims + Laboratory Values (All Criteria)
Prevalence							
N (%)	12,916 (7.0)	111 (0.1)	13,027 (7.1)	140 (0.1)	13,167 (7.1)	422 (0.2)	13,589 (7.4)
Female (y) (%)	49.1	59.5	49.1	54.3	49.2	50.2	49.2
0–17	1.1	0.9	1.1	2.1	1.1	0.5	1.0
18–44	11.9	26.1	12.1	31.4	12.3	8.5	12.1
45–64	48.5	39.6	48.4	33.6	48.3	44.1	48.1
65–74	18.7	9.9	18.6	7.9	18.5	22.5	18.7
75+	19.8	23.4	19.8	25.0	19.9	24.4	20.0
Body mass index							
Mean (kg/m^2)	32.58	33.63	32.59	31.48	32.57	30.53	32.50
≥ 30 kg/m^2 (%)	40.5	36.0	40.5	47.1	40.5	54.7	41.0
Current smoker (%)	11.2	11.4	11.2	27.6	11.3	15.2	11.5
HbA1c							
Mean	7.18	6.34	7.18	7.40	7.18	6.35	7.16
< 8 (%)	80.8	90.9	80.9	75.9	80.8	98.0	81.3
LDL-c (mg/dL)							
Mean	85.91	85.27	85.91	102.03	86.01	97.58	86.35
< 100 (%)	73.5	70.6	73.5	50.8	73.3	55.8	72.8
Systolic blood pressure (mm Hg)							
Mean	124.40	122.22	124.38	124.98	124.39	125.51	124.43
< 130 (%)	71.1	72.0	71.1	65.8	71.0	59.2	70.6
Diastolic blood pressure (mm Hg)							
Mean	70.78	70.99	70.78	72.17	70.79	72.11	70.84
< 80 (%)	81.4	71.0	81.3	75.0	81.3	70.9	80.9

*Most recent values by December 31, 2009.

[†] Qualifying case identification criteria: between January 1, 2008 and December 31, 2009, (a) at least 2 outpatient claims with diabetes diagnosis claims (ICD-9-CM codes 250.xx, 357.2, 366.41, 362.01–362.07) on separate days; (b) at least 1 inpatient claim with diabetes diagnosis (same ICD-9-CM codes); (c) at least 1 pharmacy claim for an antihyperglycemic medication except for metformin, thiazolidinedione, or exenatide; or (d) at least 2 elevated tests of HbA1c, fasting plasma glucose, random plasma glucose, or some combination of these and measured on separate days.

patterns of care, and other purposes are much greater than what has been previously available in the United States. There are many strategies available to establish the accuracy of case definitions when clinical and administrative data are available. A guiding principle in chronic disease registry design is to initially be as inclusive as possible of cases, and then subsequently apply more restrictive case definitions for specific research purposes. Despite recent advances, challenges still remain in accurately identifying cases of diabetes or asthma, constructing dynamic population cohorts, assessing data quality, and understanding how biases related to attrition and other factors may influence cohort characteristics and, ultimately, study results.

REFERENCES

1. Gliklich RE, Dreyer NA. Registries for Evaluating Patient Outcomes: A User's Guide. 2nd ed. (Prepared by Outcome DEcIDE Center [Outcome Sciences, Inc. d/b/a Outcome] under Contract No. HHSA29020050035I TO3.) AHRQ Publication No.10-EHC049. Rockville, MD: Agency for Healthcare Research and Quality. September 2010.
2. Kern E, Beischel S, Stalnaker R, et al. Building a diabetes registry from the Veterans Health Administration's computerized patient record system. *J Diabetes Sci Technol*. 2008;2:7–14.
3. NQF. *National Quality Forum Diabetes Measures*. 2010; Available at: http://www.qualityforum.org/Measuring_Performance/Measure_Maintenance/Diabetes.aspx. Accessed January 18, 2012.
4. Zhu VJ, Tu W, Rosenman MB, Overhage JM. Facilitating clinical research through the Health Information Exchange: lipid control as an example. *AMIA Annu Symp Proc*. 2010;2010:947–951.
5. Harjutsalo V, Podar T, Tuomilehto J. Cumulative incidence of type 1 diabetes in 10,168 siblings of Finnish young-onset type 1 diabetic patients. *Diabetes*. 2005;54:563–569.
6. McBean AM, Li S, Gilbertson DT, Collins A. Differences in diabetes prevalence, incidence, and mortality among the elderly of four racial/ethnic groups: whites, blacks, hispanics, and asians. *Diabetes Care*. 2004;27:2317–2324.
7. Sorensen HT, Lash TL. Use of administrative hospital registry data and a civil registry to measure survival and other outcomes after cancer. *Clin Epidemiol*. 2011;3:1.
8. Banta JE, Morrato EH, Lee SW, Haviland MG. Retrospective analysis of diabetes care in California Medicaid patients with mental illness. *J Gen Intern Med*. 2009;24:802–808.
9. Miller DR, Pogach L. Longitudinal approaches to evaluate health care quality and outcomes: the Veterans Health Administration diabetes epidemiology cohorts. *J Diabetes Sci Technol*. 2008;2:24–32.
10. Magid DJ, Shetterly SM, Margolis KL, et al. Comparative effectiveness of angiotensin-converting enzyme inhibitors versus beta-blockers as second-line therapy for hypertension. *Circ Cardiovasc Qual Outcomes*. 2010;3:453–458.
11. Pace WD, Cifuentes M, Valuck RJ, et al. An electronic practice-based network for observational comparative effectiveness research. *Ann Intern Med*. 2009;151:338–340.
12. Hsiao C, Hing E, Socey TC, et al. *Electronic Medical Record/Electronic Health Record Systems of Office-based Physicians: United States 2009 and Preliminary 2010 State Estimates*. 2010; NCHS Health E-Stats. Available at: http://www.cdc.gov/nchs/data/hestat/emr_ehr_09/emr_ehr_09.htm. Accessed January 18, 2012.
13. HMORN. *HMO Research Network*. Available at: http://www.hmoresearchnetwork.org/. Accessed January 18, 2012.
14. FCC. *Federal Coordinating Council for Comparative Effectiveness Research*. 2011. Available at: http://www.hhs.gov/recovery/programs/cer/cerannualrpt.pdf. Accessed January 18, 2012.
15. IOM. *Initial National Priorities for Comparative Effectiveness Research*. 2011. Available at: http://www.nap.edu/catalog.php?record_id=12648#toc. Accessed January 18, 2012.
16. Dreyer NA, Garner S. Registries for robust evidence. *JAMA*. 2009;302:790–791.
17. Glasgow RE, Magid DJ, Beck A, et al. Practical clinical trials for translating research to practice: design and measurement recommendations. *Med Care*. 2005;43:551–557.
18. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. *JAMA*. 2003;290:1624–1632.
19. Kremers HM. Methods to analyze real-world databases and registries. *Bull NYU Hosp Jt Dis*. 2009;67:193–197.
20. Richesson RL. Data standards in diabetes patient registries. *J Diabetes Sci Technol*. 2011;5:476–485.
21. Szklo M. Population-based cohort studies. *Epidemiol Rev*. 1998;20:81–90.
22. Hartert TV, Togias A, Mellen BG, et al. Underutilization of controller and rescue medications among older adults with asthma requiring hospital care. *J Am Geriatr Soc*. 2000;48:651–657.
23. Wakefield DB, Cloutier MM. Modifications to HEDIS and CSTE algorithms improve case recognition of pediatric asthma. *Pediatr Pulmonol*. 2006;41:962–971.
24. Saydah SH, Geiss LS, Tierney E, et al. Review of the performance of methods to identify diabetes cases among vital statistics, administrative, and survey data. *Ann Epidemiol*. 2004;14:507–516.
25. O'Connor PJ, Rush WA, Pronk NP, et al. Identifying diabetes mellitus or heart disease among health maintenance organization members: sensitivity, specificity, predictive value, and cost of survey and database methods. *Am J Manag Care*. 1998;4:335–342.
26. Harris SB, Glazier RH, Tompkins JW, et al. Investigating concordance in diabetes diagnosis between primary care charts (electronic medical records) and health administrative data: a retrospective cohort study. *BMC Health Serv Res*. 2010;10:347. Accessed January 18, 2012.
27. Mosen DM, Macy E, Schatz M, et al. How well do the HEDIS asthma inclusion criteria identify persistent asthma? *Am J Manag Care*. 2005;11:650–654.
28. Fuhlbrigge AL, Carey VJ, Finkelstein JA, et al. Validity of the HEDIS criteria to identify children with persistent asthma and sustained high utilization. *Am J Manag Care*. 2005;11:325–330.
29. Hartert TV, Neuzil KM, Shintani AK, et al. Maternal morbidity and perinatal outcomes among pregnant women with respiratory hospitalizations during influenza season. *Am J Obstet Gynecol*. 2003;189:1705–1712.
30. Hartert TV, Speroff T, Togias A, et al. Risk factors for recurrent asthma hospital visits and death among a population of indigent older adults with asthma. *Ann Allergy Asthma Immunol*. 2002;89:467–473.
31. Gershon AS, Wang C, Guan J, et al. Identifying patients with physician-diagnosed asthma in health administrative databases. *Can Respir J*. 2009;16:183–188.
32. Standards of medical care in diabetes—2011. *Diabetes Care*. 2011;34(suppl 1):S11–S61.
33. Hennekens CH, Buring JE. In: Sherry P, Mayrent L, eds. *Epidemiology in Medicine*. Philadelphia, PA: Lippincott Williams & Wilkins; 1987.
34. Wu P, Dupont WD, Griffin MR, et al. Evidence of a causal role of winter virus infection during infancy in early childhood asthma. *Am J Respir Crit Care Med*. 2008;178:1123–1129.
35. Valet RS, Gebretsadik T, Carroll KN, et al. High asthma prevalence and increased morbidity among rural children in a Medicaid cohort. *Ann Allergy Asthma Immunol*. 2011;106:467–473.
36. Weiner MG, Lyman JA, Murphy S, et al. Electronic health records: high-quality electronic data for higher-quality clinical research. *Inform Prim Care*. 2007;15:121–127.
37. Kahn MG, Ranade D. The impact of electronic medical records data sources on an adverse drug event quality measure. *J Am Med Inform Assoc*. 2010;17:185–191.
38. Chan KS, Fowles JB, Weiner JP. Review: electronic health records and the reliability and validity of quality measures: a review of the literature. *Med Care Res Rev*. 2010;67:503–527.
39. Meystre S, Haug P. Automation of a problem list using natural language processing. *BMC Med Inform Decis Mak*. 2005;5:30. Accessed January 18, 2012.
40. *State Health Facts*. 2011. Available at: http://www.statehealthfacts.org/ (Accessed January 18, 2012).
41. Fewtrell MS, Kennedy K, Singhai A, et al. How much loss to follow-up is acceptable in long-term randomised trials and prospective studies? *Arch Dis Child*. 2008;93:458–461.
42. Dumville JC, Torgerson DJ, Hewitt CE. Reporting attrition in randomised controlled trials. *BMJ*. 2006;332:969–971.

© 2012 Lippincott Williams & Wilkins