Enhancing Clinical Content and Race/Ethnicity Data in Statewide Hospital Administrative Databases: Obstacles Encountered, Strategies Adopted, and Lessons Learned

Michael Pine, Niranjana M. Kowlessar, Jason L. Salemi, Jill Miyamura, David S. Zingmond, Nicole E. Katz, and Joe Schindler

Objectives. Eight grant teams used Agency for Healthcare Research and Quality infrastructure development research grants to enhance the clinical content of and improve race/ethnicity identifiers in statewide all-payer hospital administrative databases.

Principal Findings. Grantees faced common challenges, including recruiting data partners and ensuring their continued effective participation, acquiring and validating the accuracy and utility of new data elements, and linking data from multiple sources to create internally consistent enhanced administrative databases. Successful strategies to overcome these challenges included aggressively engaging with providers of critical sources of data, emphasizing potential benefits to participants, revising requirements to lessen burdens associated with participation, maintaining continuous communication with participants, being flexible when responding to participants' difficulties in meeting program requirements, and paying scrupulous attention to preparing data specifications and creating and implementing protocols for data auditing, validation, cleaning, editing, and linking. In addition to common challenges, grantees also had to contend with unique challenges from local environmental factors that shaped the strategies they adopted.

Conclusions. The creation of enhanced administrative databases to support comparative effectiveness research is difficult, particularly in the face of numerous challenges with recruiting data partners such as competing demands on information technology resources. Excellent communication, flexibility, and attention to detail are essential ingredients in accomplishing this task. Additional research is needed to develop strategies for maintaining these databases when initial funding is exhausted.

Key Words. Hospital administrative data, data enhancement, information technology in health

There is increasing demand for comprehensive databases for understanding population health and assessing the benefits and harms of alternative health care interventions (Marko and Weil 2010; Shah, Drozda, and Peterson 2010). By enhancing hospitals' administrative data with carefully selected clinical, demographic, and resource utilization data elements, existing datasets can provide invaluable information for health care stakeholders and decision makers (Concato et al. 2010; Fleurence, Naci, and Jansen 2010).

Prior research has demonstrated the value of enhancing administrative claims databases with additional elements including numerical laboratory values, other test results, and present-on-admission (POA) modifiers. Specifically, the incorporation of POA modifiers of ICD-9-CM diagnosis codes into claims databases has improved the validity of risk-adjustment models (Pine et al. 2007) and resulted in improved quality measurement (Dalton et al. 2013). Similarly, previous studies have shown improvements in predictions of inpatient mortality and complications as a result of linking claims data to hospital numerical laboratory data (Jordan et al. 2007; Pine et al. 2007).

Improvements in the processes and protocols to collect patient race/ethnicity/primary language (R/E/L) information are needed to address disparities in health care and health outcomes attributable to differences in race, ethnicity, and primary language, which are well documented and persistent even after adjusting for differences in related characteristics such as education, income, insurance, access to care, and health status. It has been estimated that racial and ethnic disparities in health and health care cost the United States \$1.24 trillion between 2003 and 2006: over \$200 billion for direct medical expenses, and another \$1 trillion for indirect costs such as lost quality of life years and lost productivity (LaVeist, Gaskin, and Richard 2011). In recent years, there has been increased focus on eliminating disparities in access to care and health outcomes for racially, ethnically, and linguistically diverse populations in the United States.

Address correspondence to Michael Pine, M.D., M.B.A, Michael Pine and Associates, 1 East Upper Wacker Drive #1210, Chicago, IL 60601; e-mail: mpine@consultmpa.com. Niranjana M. Kowlessar, Ph.D., is with the Social & Scientific Systems, Inc., Silver Spring, MD. Jason L. Salemi, Ph.D., M.P.H., is with the Baylor College of Medicine, 3701 Kirby Drive, Room LMPL-600, Mail Stop BCM700, Houston, TX 77098. Jill Miyamura, Ph.D., is with the Hawaii Health Information Corporation, Honolulu, HI. David S. Zingmond, M.D., Ph.D., is with the UCLA Division of General Internal Medicine and Health Services Research, Los Angeles, CA. Nicole E. Katz, M.P.H., is with the University of New Mexico, Albuquerque, NM. Joe Schindler, B.A., is with the Minnesota Hospital Association, St. Paul, MN.

In 2010, in an effort to build on prior work initiated by state data organizations and ultimately create comprehensive, validated, and sustainable data infrastructures for comparative effectiveness research (CER), AHRQ, under the American Recovery and Reinvestment Act, funded eight 3-year infrastructure development research grants. Grantees from Florida, Hawaii, Minnesota, New Jersey, and New York collaborated with state and health care provider organizations to enhance existing hospital claims and discharge databases with clinical information and to demonstrate the utility of the new data infrastructure by using the new databases in a CER study. Enhancements included linking existing hospital administrative data to hospital laboratory data; hospital inpatient, ambulatory, and emergency room claims and discharge data; birth and death certificate data; inpatient pharmacy order data; and preadmission emergency medical services data. Grantees from California, New Mexico, and the Improving Data and Enhancing Access-Northwest (IDEA-NW) project—which included partners from Idaho, Oregon, and Washington—improved the quality of their race and ethnicity data so that their enhanced databases would improve analyses of disparities in health care services and resulting clinical outcomes. All eight grantees recognized a deficiency in the availability and use of robust, longitudinal data systems to support CER and disparities research in their states (Salihi, Salinas, and Mogos 2013; Salinas-Miranda et al. 2013) and disciplines, and felt that enhancement of their databases with detailed, individual-level sociodemographic, health, and geospatial information would be valuable in improving epidemiological, clinical, and health services research.

While widespread adoption of these enhancements would likely increase the value of both state data and nationally integrated data for tracking population health, monitoring risk-adjusted clinical performance, and studying the comparative effectiveness of various health care interventions, a variety of administrative, technical, and logistical problems confronted grantees. Hurdles faced by all grantees included obtaining approval for use and linkage of new databases, and garnering active participation from key stakeholders (e.g., hospitals and state health departments). Clinical data grantees also struggled with designing effective protocols for linking diverse data elements with minimal identifying information from disparate sources. Grantees working to improve race, ethnicity, and tribal data faced unique challenges in ensuring the data collected would conform to Office of Management and Budget standards.

This paper details obstacles encountered, lessons learned, and both successful and unsuccessful strategies adopted by grantees as they aimed to

improve the clinical content and race/ethnicity/language data in their states' hospital-based, encounter-level databases.

RECRUITING DATA PARTNERS

Grantees' experiences acquiring the data necessary to enhance their existing databases varied greatly, depending on the nature and planned use of the data, available resources, existing relationships with data owners and partners, and internal operational decisions. In selecting data sources, grantees faced trade-offs between the quality and usability of data, and the ease of securing, acquiring, and transmitting the data. Some data elements—such as race and ethnicity—required acquisition of primary data with the associated start-up and continuing costs; others—such as numerical laboratory data—could be obtained from electronic data repositories, though often with idiosyncratic local data formats and information systems. Each team reached its own compromise between immediate access to imperfect existing data and more resource-intensive on-site transformation of data into appropriate formats.

Gaining Approval from Decision Makers

Successful recruitment of data partners required aggressive multilevel communication with senior hospital management, heads of departments (e.g., information technology [IT]) whose cooperation was essential to the project, and key personnel from areas most likely to benefit directly from participation (e.g., quality improvement). Communication often was facilitated through collaboration with groups such as hospital associations and state data organizations that had strong ties to targeted data suppliers. Assistance included those groups' publicizing and endorsing projects, advising grantees about how best to secure participation, writing strong letters of recommendation to key decision makers, and providing venues for educational orientation sessions.

When changes in a provider's data collection procedures or the acquisition of new data (e.g., obtaining information about race, ethnicity, and language) were required, strong commitment from senior management at the start of the project was essential. Letters to hospital CEOs and selected senior managers with follow-up direct contact, when appropriate, appeared to be most effective. However, when data could be obtained from existing data repositories—once senior management approval was obtained—successful

data acquisition generally was achieved through tailored approaches to identifying affected data managers and securing their individual buy-in.

Well-structured, informative orientation sessions held at central locations with webinar transmission proved to be an effective method of securing the participation of potential data suppliers. Presentations included background material about the project and its potential significance, details about what would be expected of participants and what assistance they could expect from project staff, and potential direct benefits. Attendees generally included midlevel operational staff and physicians who were interested in clinical performance assessment and improvement. Questions from attendees helped identify barriers to participation and possible methods of overcoming these obstacles.

Benefiting from Legal Requirements

Environmental factors—including regulatory or reimbursement requirements—and prior organizational relationships were important determinants of grantees' experiences with recruiting data partners. Grantees in New Mexico and California both benefitted from their states' legal mandates that required hospitals to collect and more precisely report race and ethnicity data. Data acquisition by IDEA-NW was facilitated by tapping secondary data sources that already had obtained and formatted required data from multiple sites. Similarly, the University of South Florida capitalized on a long track record of existing data use agreements with Florida's Agency for Health Care Administration, which had a legal mandate to collect and store detailed hospital data. This working relationship eliminated the need for the grantee team to secure cooperation from individual hospitals, which in Florida would have constituted more than 300 separate facilities.

Another approach for encouraging participation involved linking data submission to the receipt of substantial monetary rewards from external sources such as those available from the Medicare and Medicaid Electronic Health Record (EHR) Incentive Program for the meaningful use of certified EHR technology (Centers for Medicare & Medicaid Services n.d.). However, unexpected changes in requirements and timetables for the programs undermined attempts by some grantees to benefit from the initiatives.

Demonstrating Benefits to Participation

From the perspective of a data supplier, direct benefits of participation may include the following: (1) more accurate information about comparative

clinical performance that can be used to support internal incentive programs and performance improvement initiatives, (2) data that can be used to meet regulatory or reimbursement requirements, and (3) data that can help secure direct payment for successfully meeting program requirements.

Despite the benefits associated with participation, many providers or suppliers of data were reluctant to do so because developing databases that are important to the health care community can be costly (though through data use fees, contracts, etc., the cost of ongoing acquisition/maintenance can be covered) and, unfortunately, the return on investment is not immediate. A substantial number of hospitals were unwilling to participate without hard data about the required level of effort. Even when hospitals expressed interest in working with grantees, numerous competing high-priority IT initiatives (such as implementation of electronic medical records and adoption of meaningful use standards) often exhausted the very hospital resources that would be required to compile and transmit the projects' requested data (Blumenthal and Tavenner 2010).

Whether changes in data reporting are required or voluntary, individual participating hospitals often must devote time and scarce resources to implementing new data collection systems or modifying current ones to comply. Some participating hospitals incurred costs converting local codes into a standardized data format and extracting required data from existing data repositories. Others incurred costs training staff in data collection and validation, while others had to alter IT systems to obtain and transmit required data. In particular, projects that depended upon transfer of data from hospitals' electronic data repositories found that hospitals with weak IT infrastructures were unable to access these data cost-effectively and therefore were unexpectedly difficult to recruit. On the other hand, shared responsibilities and divided authority at larger, more sophisticated facilities often made it difficult for grantees to locate and communicate with individuals who were empowered to approve collaborative arrangements. Private for-profit hospitals were especially reluctant to collaborate in projects that did not promise to benefit their bottom line directly.

To overcome these challenges, grantees expended a great deal of effort creating and communicating tangible benefits and incentives to targeted data partners, describing potential direct benefits in marketing materials and presentations. Based on new capabilities of their enhanced claims databases, many grantees developed standard and customized tools and reports that met current needs of their data partners. Participating California hospitals were offered forms, protocols, and instruction materials to

improve their collection of patient demographic data and elicit race, ethnicity, and language abilities and preferences from their patients. These materials included definitions of standardized categories for race and ethnicity, and information about new initiatives and policies pertaining to acquisition and reporting of these data. Individualized hospital report cards demonstrating how enhanced data elements might be incorporated into performance monitoring and quality improvement activities were developed to enhance the value of participation. Minnesota shared results of screens to identify deficiencies in hospitals' coding of POA modifiers to ICD-9-CM diagnosis codes; the New Mexico grantees produced feedback reports for participating facilities on each facility's data quality and completeness; and New York, Minnesota, and Hawaii all prepared demonstrations of how the addition of numerical laboratory results to already collected claims data improved assessments of risk-adjusted clinical outcomes. (And one group of hospitals in Hawaii anticipated enough potential direct benefit from participation to request expansion of data collection beyond that requested by the grantee.)

Sometimes nonmonetary inducements proved inadequate to raise project participation to a high enough priority to ensure timely submission of usable data. Direct payment of up to \$5,000 per facility to cover the costs of required systems modifications in New York, and provision of customized expert external IT support in Minnesota were extremely effective in turning unresponsive volunteers into active participants. In general, these investments were carefully tailored to cover only nonrecurring set-up expenses.

Maintaining Flexibility in Project Requirements

When recruitment fell short of expectations, grantees used a variety of approaches to increase participation. These included altering requirements for participation by creating new data collection tools, altering data acquisition and transmission protocols, adjusting timelines, and transferring responsibility for some aspects of data preparation and quality control from hospitals to centralized grantee-administered sites. For example, grantees collecting laboratory data initially required the use of standard Logical Observation Identifiers Names and Codes (LOINC) to identify tests (Regenstrief Institute 2014) but discovered that some hospitals were reluctant to participate because they had no experience using these codes. In

response to this feedback, the grantees altered their data submission requirements, reprogrammed their interfaces to accept the formats preferred by potential participating hospitals, and developed, tested, and implemented methods to map transmitted data into operationally acceptable standardized formats.

Even when data were relatively easy to acquire and transmit, administrative and legal obstacles often resulted in lengthy delays. In some instances,

Table 1: Recruiting Data Partners

<i>Challenges Encountered</i>	<i>Strategies Employed to Address Challenges</i>
Gaining approval from decision makers	<ul style="list-style-type: none"> • Multilevel communication with senior hospital management, heads of departments at partner organizations, and other key personnel. • Collaborations with state data organizations and/or hospital associations that had strong ties to and could communicate the potential benefits of participation to targeted data partners. • Well-structured, informative orientation sessions targeted at potential data suppliers.
Demonstrating benefits to participation	<ul style="list-style-type: none"> • Grantees created and communicated tangible benefits and incentives to targeted data partners using marketing materials and presentations. • Based on new capabilities of their enhanced claims databases, many grantees developed standard and customized tools and reports that met current needs of their data partners. • Individualized hospital report cards demonstrating how enhanced data elements might be incorporated into performance monitoring and quality improvement activities demonstrated the value of participation. • Incentives such as direct payment of up to \$5,000 per facility for system modifications and expert external IT support were extremely effective in turning unresponsive volunteers into active participants.
Recruitment fell short of expectations	<ul style="list-style-type: none"> • Grantees reduced provider burdens by reducing scope of their requests, and shifting some of the technical and operational work to themselves. <ul style="list-style-type: none"> ○ Grantees altered requirements for participation by creating new data collection tools and data acquisition and transmission protocols and also by adjusting timelines for data submission. ○ Grantees transferred responsibility for some aspects of data preparation and quality control from hospitals to centralized grantee-administered sites.

hospitals were restricted explicitly from participation in research activities. More often, complex local regulations governing the release, sharing, and use of identifiable personal health data, restrictions on access to specific clinical information such as results of newborn drug screening, the reluctance of hospital attorneys to assume any risks inherent in nonmandated data-sharing arrangements, and hospital administrators' particular concerns about the sensitivity of race, ethnicity, and language information presented serious challenges to grantees as they attempted to acquire the data needed to achieve their analytic objectives (Table 1).

ENSURING EFFECTIVE PARTICIPATION BY DATA PARTNERS

After facilities were recruited, continuous communication was essential to maintain engagement and obtain timely submission of usable data. Flexibility was particularly important; every grantee was forced to narrow its menu of required data elements in response to the limited availability of easily retrievable, consistently recorded electronic data.

Customized Approaches to Communication

Grantees differed in their approaches to communicating with their data providers based on project requirements and environmental factors. Most grantees, including Hawaii, New York, Minnesota, California, and New Mexico, maintained direct ongoing communication with primary data providers. Others, including IDEA-NW and the Florida grantee, communicated primarily through reports and intermediaries. Team members from IDEA-NW routinely reported misclassifications discovered during data reconciliation to responsible facilities, produced statewide reports disseminated to health organizations throughout the region, and shared experiences and recommendations with regional organizations such as State Hospital Associations and State Departments of Health. The University of South Florida was able to leverage its collaborative partnership with Florida's Agency for Health Care Administration to assist hospitals in standardizing and accurately transmitting required infant and maternal data, as that agency routinely audits Florida hospitals' inpatient and ambulatory discharge data, prepares and distributes numerous data quality reports to these hospitals, and works directly with hospitals to identify and correct data deficiencies.

cies. Providing hospitals with clear, detailed data specifications, operational protocols, and training materials greatly facilitated internal communication and increased the probability of receiving complete, accurate, and timely data.

Proactive engagement of targeted data providers often spelled the difference between high-quality and suboptimal submissions. Websites, informational mailings, e-mail communications, phone calls, on-site training sessions, individualized expert consultation, and face-to-face meetings with critical personnel all were employed to keep data providers up to date and engaged, to share knowledge and experiences, to learn of potential problems and delays, and to facilitate data acquisition and transfer. For example, grantees in New Mexico found that improvements in the quality of race/ethnicity data were achieved when critical IT personnel were included in their educational efforts.

Continuous feedback between grantee organizations and participating facilities was of paramount importance in identifying gaps in data accuracy and completeness, in increasing sensitivity to the importance of creating and adhering to standardized data specifications, and in implementing best practice protocols for data collection and transmission. For example, the California grantees provided their participating hospitals with scripts, sample questionnaires, answers to frequently asked questions, and educational materials they could adapt and use to improve the quality of their race, ethnicity, and language data collection. Site-specific feedback and dissemination of facility-specific reports about comparative data quality and about proven methods to improve data quality were particularly useful in developing and reinforcing participating organizations' and agency stakeholders' commitment to securing and supplying complete, accurate, and valid administrative and clinical information.

Lessening the Burden on Data Providers

The rapid pace of health care reform has put tremendous pressure on hospitals and other potential data suppliers, particularly in the area of IT. Competing priorities often force many important projects to be abandoned or delayed; thus, these initiatives have a better chance of coming to fruition if financial and technical support can be provided to reduce dependence on overextended internal resources. Also essential is the availability of reliable information from past experience setting up operational systems for obtaining and transmitting data from comparable data sources.

To address data providers' limited IT resources, grantees at times loosened their requirements for how submitted data were formatted. Once received, the grantees reformatted the data to satisfy their usability requirements.

The desired type and frequency of data transmission were important determinants of the type of transmission tools selected for each project. For projects collecting data in discrete batches at infrequent intervals, importing simple ASCII-formatted files into SAS, SQL Server, or Oracle databases was relatively simple and straightforward. However, when concurrent or frequent transmission of recently acquired data was important, data transfer using Health Level 7 (i.e., HL7) interfaces was preferred by grantees and by sites already versed in this method of transferring data (Health Level Seven International 2014). Although HL7 is regarded as an industry standard for exchanging, managing, and integrating health care data, many data providers were not well equipped to format their data to satisfy HL7 transmission requirements. Rather than abandon HL7 data transmission, some grantees provided participating facilities with HL7 "interface engines" that greatly facilitated their utilization of this superior technology. These software programs could be used by the grantees to interact with numerous health care providers without the need for major customization.

Insurmountable Barriers

Even when hospital leadership and staff were sincerely committed to the projects' aims and objectives, other managerial and clinical priorities often resulted in delays in data acquisition and transmission and occasionally resulted in hospitals withdrawing from a project entirely. In Hawaii, staffing problems resulted in one hospital's failure to complete the transfer of its live and historical data. In Minnesota, three hospitals withdrew because of competing demands on resources caused by requirements for computer conversions and ICD-10 implementation. Staff turnover within hospitals sometimes resulted in unanticipated gaps in knowledge and understanding that affected both the quality and timeliness of data transmission and increased orientation and training costs. Lack of communication within participating facilities was a recurring problem that often necessitated careful tracking and tactful interventions by grantees. Personal relationships between grantees and hospital staff that were developed or strengthened during hospital recruitment sometimes proved invaluable in overcoming hospitals' internal barriers to effective participation (Table 2).

Table 2: Ensuring Effective Participation by Data Partners

<i>Challenges Encountered</i>	<i>Strategies Employed to Address Challenges</i>
Maintaining engagement and obtaining timely data submission	<ul style="list-style-type: none"> • Grantees engaged in continuous communication with their data providers, reported their findings from their data explorations back to providers • Providing hospitals with detailed data specifications, protocols, and training materials greatly improved communication and increased the probability of receiving complete, accurate, and timely data submissions
Ensuring high-quality data submissions	<ul style="list-style-type: none"> • Continuous customized feedback to critical personnel was employed to keep data providers engaged, learn of potential problems and delays, and facilitate data acquisition and transfer. • Site-specific feedback and facility-specific reports on comparative data quality and proven methods for improving it were useful in reinforcing participating organizations' commitments to providing high-quality data.

PRODUCING HIGH-QUALITY DATA

Meticulous auditing, validation, cleaning, editing, and reformatting were required to produce the high-quality integrated datasets required to conduct credible studies of comparative effectiveness. While many approaches to database preparation were applicable for both demographic and clinical enhancements, individualized methods were also required.

All grantees used predetermined data specifications to assess the completeness and conformity of transmitted data. When data were submitted in nonstandard formats, translational programs were often employed to transform them to meet the standards. If missing or uncorrectable nonconforming data were judged to be critical, some grantees requested updated submissions from data providers, but all were careful to limit such requests so as not to impose unduly on their data partners. Aggregated data then were checked for internal consistency and face validity, generally using absolute standards (e.g., clinically credible bounds for specific laboratory test results), standards of internal consistency (e.g., consistency of longitudinal reporting of demographic information), or external standards of plausibility (e.g., observed hospital rates compared to comparable rates from similar institutions). Identified errors and inconsistencies were corrected using internally derived protocols or feedback from data providers.

The integrity of clinical data often could be assessed readily based on interrelationships of qualitative and quantitative measures (e.g., diagnoses and laboratory test results). Median values of numerical data for specific measures performed on specific patient populations were extremely consistent from facility to facility. When audits of numerical test results from individual hospitals revealed median test results that differed from expected values, communication with facilities invariably revealed errors in data acquisition or transmission. When correlations between objective clinical data and coded diagnoses revealed inconsistencies, they generally appeared to be the result of poor documentation in medical records rather than errors in newly acquired data. Because enhancements using clinical data often required obtaining new data directly from hospitals' data repositories, it was essential to establish secure, reliable, cost-effective methods of transmitting accurate data to a centralized facility. Test datasets were used to ensure that each step from data collection to creation of an integrated clinically enhanced claims database was clearly specified and fully operational.

In contrast to clinical data, which may vary in unpredictable ways, demographic data such as race and ethnicity should remain constant for an individual across time. For race and ethnicity data, application of internal and external reference standards to assess the accuracy of the data was of limited use in identifying problematic data. In California, inconsistent reporting of the race and ethnicity of a patient readmitted to the same hospital was helpful in identifying a data integrity issue. Fixed thresholds based on absolute prevalence rates or changes in prevalence were not seen as useful; rather, self-reported data were the preferred reference standard to assess the accuracy of race and ethnicity data. This sometimes could be accomplished by linking states' hospital discharge databases to routinely collected, patient-identifiable, self-reported race and ethnicity data. However, available reference databases were of limited use because they generally focused on narrowly defined patient populations such as patients enrolled in cancer trials. To obtain a high-quality reference database, New Mexico's Department of Health elected to perform a resource-intensive telephone survey of patients hospitalized in the prior year.

For all grantees, the goal of data auditing was not only to create high-quality research databases but also to improve the scope and quality of future data available for research and decision making. To the extent possible, grantees attempted to share results of data audits with their data providers' decision makers and frontline data collectors and to effect changes that would improve the completeness and quality of their states' hospital discharge databases.

Table 3: Producing High-Quality Data

<i>Challenges Encountered</i>	<i>Strategies Employed to Address Challenges</i>
Data submission in nonstandard formats	<ul style="list-style-type: none"> • Translational programs were employed whenever possible to transform submitted nonconforming data to meet the standards to minimize requests to providers for updated submissions.
Discrepancies in audited data	<ul style="list-style-type: none"> • When audits of individual hospitals' data revealed discrepancies, communication with the facility invariably revealed errors in data acquisition or transmission, not outlier data values. • To ensure that each step from data collection to creation of an integrated clinically enhanced claims database was clearly specified and fully operational, test datasets were employed.
Identifying a reference standard for race and ethnicity data	<ul style="list-style-type: none"> • While self-reported data were the preferred reference standard for accuracy of race and ethnicity data, in one case, a telephone survey of prior hospital patients yielded a high-quality reference database. • A roster of AI/AN individuals who had registered at tribal, Indian Health Service, and urban Indian clinics in the northwest—the most complete listing of AI/AN available at the time—was available from Northwest Portland Area Indian Health Board to use as a reference standard.

However, in some cases, grantees were limited to communicating findings to individuals who uploaded data and not to organizational decision makers (Table 3).

LINKING DATA FROM MULTIPLE SOURCES

In seeking to link data from diverse sources to create valuable new resources for health care researchers and decision makers, all grantees faced challenges selecting and standardizing linking variables, developing and implementing linking protocols that met stakeholder-guided objectives, and evaluating the quality of the linkages they achieved.

Selecting and Standardizing Linking Variables

Available datasets contained different combinations of personal identifiers and other data elements that could be used to link individual patients'

records. The most frequently used linking variables included unique identifiers such as Social Security numbers (SSNs), medical record numbers and patient control numbers, and other less specific identifiers such as patients' names, addresses, gender, and date of birth, and dates and sites of care. Even when unique identifiers were available, organizations with primary responsibility for these data—such as state governments, private health plans, and health care providers—often were bound by diverse, complex sets of laws and policies regarding data privacy, confidentiality, and sharing (Bradley et al. 2010), and had service and research agendas that differed from those of grantees that sought permission to use their data. When access to unique identifiers was denied, grantees modified initial data linkage protocols or relied on linkages created by their data suppliers, often with guidance and consultation from the grantees' research teams. The quality and timeliness of these linkages varied based on available time and personnel resources including skill in the use of data linkage programs. In addition to the availability of specific linking variables, other factors influencing grantees' choices of linkage protocols included (1) the nature of the data being linked, (2) relationships between research teams and data sources, (3) stakeholder concerns, (4) grantees' knowledge of and experience with the databases being linked, and (5) grantees' expertise using various linking software programs.

Every grantee utilized deterministic matching, either alone or as a precursor to probabilistic matching (Clark and Hahn 1995; Jaro 1995; Blakely and Salmond 2002; Méray et al. 2007; Mason and Tu 2008). All performed deterministic matches with internally developed SAS code, which, in contrast to less labor-intensive point-and-click packages, readily handled extremely large datasets (sometimes exceeding 5 million records) and provided users with high degrees of customizability and control. California matched death certificates and cancer registry data to hospital discharge data using SSNs, dates of birth and death, gender, and ZIP codes. New York linked patient-level administrative hospital discharge data across facilities with inpatient numerical laboratory results at a single facility, and with postdischarge death certificate data. The New York team used their own iterative matching algorithms that included exact and partial matching on medical record numbers, SSNs, first and last names, gender, dates of birth, dates of admission and discharge, street addresses, and ZIP codes. Matching algorithms employed by grantees often incorporated string comparators for names and allowed for common typographical errors and transpositions in SSNs and dates.

Florida developed a specialized multistage stepwise deterministic record linkage protocol on birth, death, and hospital encounter-level data to solve the

unique problem of linking mother–infant dyads (Salemi et al. 2013). To link maternal and infant longitudinal data, Florida used parental SSNs when infants did not have their own unique SSNs, permitted “crossover” linkages when fathers’ SSNs appeared on hospital discharge records, and devised and implemented algorithms to assign records to individual infants who shared paternal information in cases of multiple births.

Developing Linking Protocols That Met Stakeholder Objectives

When unique patient identifiers were unavailable, grantees often created data linkages using less specific identifiers and a variety of probabilistic matching techniques supported by a diverse set of off-the-shelf tools. Grantees generally selected specific tools based primarily on familiarity, cost, ease of use, and the degree to which users could control the linkage process. The following tools were used by grantees to link core hospital administrative records to data enhancements: Hawaii used LinkageWiz (Anon n.d.b) for laboratory data; Minnesota used SQL for laboratory data, pharmacy orders, and death certificates (Anon n.d.c); New Jersey used Link King for emergency medical service and mortality records (Anon n.d.d); California used Auto-Match for birth and death certificates (Anon n.d.a); and New Mexico and IDEA-NW used Link Plus for vital statistics and disease registry data (Centers for Disease Control and Prevention n.d.).

Evaluating the Quality of Linkages

Some grantees conducted focused studies to assess the quality of the linkage they achieved and make recommendations about how to improve data matches in the future. Florida compared maternal and infant sociodemographic, health, and geospatial characteristics among linked and unlinked records to detect potential biases that should be factored into analyses of comparative effectiveness and public health (Salemi et al. 2013). Minnesota assessed the sensitivity and specificity of alternative matching algorithms applied to its heavily censored administrative database, using matches based on high-quality unique identifiers as a reference standard. The IDEA/NW project used a roster of individuals who had registered at tribal, Indian Health Service, and urban Indian clinics in the northwest (the most complete listing of AI/AN available at the time) to perform record linkages with health data systems in Idaho, Oregon, and Washington (Table 4).

Table 4: Linking Data from Multiple Sources

<i>Challenges Encountered</i>	<i>Strategies Employed to Address Challenges</i>
Selection of linkage variable(s)	<ul style="list-style-type: none">● When access to unique identifiers such as Social Security numbers or medical record numbers was unavailable or denied:<ul style="list-style-type: none">○ Grantees modified initial data linkage protocols or relied on linkages created by their data suppliers, often with guidance and consultation from the grantees' research teams.
Developing and implementing linking protocols	<ul style="list-style-type: none">● When unique identifiers were unavailable, grantees often created data linkages using a variety of probabilistic matching techniques supported by a diverse set of off-the-shelf tools.<ul style="list-style-type: none">○ Grantees generally selected specific tools based primarily on familiarity, cost, ease of use, and the degree to which users could control the linkage process.
Evaluating linkage quality	<ul style="list-style-type: none">● Grantees conducted focused studies to assess the quality of the linkage they achieved and make recommendations about how to improve data matches in the future.

CREATING SELF-SUSTAINING ENHANCED CLAIMS DATABASES

Grantees began their projects understanding that developing a sound infrastructure and achieving operational objectives would not be sufficient for survival.

The ultimate sustainability of clinically enhanced database initiatives depend to a large extent upon the ease with which participants can continue to supply required data once the database is established. Once operational, data acquisition and transmission protocols that utilize existing data repositories and minimize the need for on-site resources for processing data prior to transmission greatly improve the prospect of continued participation. Grantees were able to support data partners in achieving sustainability either by establishing electronic data standardization and integration protocols on-site or by creating easily implemented data formatting and transmission specifications and performing most data standardization and integration operations themselves. Upfront investments in the development of flexible infrastructure and easily disseminated tools that lower future costs of participation can be instrumental in the retention of initial participants and the recruitment of new data partners. Conversely, while rigid operational requirements and high levels of dependence upon the expertise and resources of data partners may have

short-run advantages, they will seriously impair efforts to sustain and expand enhanced databases when start-up funding is exhausted.

Ultimately, the successful transition from research and development to operational maturity depends upon development of a viable plan that ensures a revenue stream sufficient to support continued operations.

Looking Ahead

The clinical data grantees plan to sustain the accomplishments they have achieved through the AHRQ grants by (1) carefully tailoring data requirements to ensure that only well-specified, readily available clinical data of proven utility are incorporated in its clinically enhanced claims database, (2) providing individualized technical support for facilities that wish to participate, (3) linking its database initiative with other high-priority programs such as the federal government's "meaningful use" initiative or mandated statewide comparative clinical performance reporting programs, and (4) developing technologies that link enhanced databases to reimbursement arrangements (e.g., prospective risk-adjusted bundled payments for episodes of care), thereby providing an opportunity for participating facilities to secure measurable financial returns on their investment in the program.

The race/ethnicity data grantees plan to sustain the accomplishments they have achieved through the AHRQ grant by pursuing other grants and funding sources, and by building and leveraging their existing relationships with partner organizations. In addition, they plan to engage in coordinated efforts and build new partnerships to reduce health disparities by disseminating best practices and tools to help initiate or expand activities to improve R/E/L data. All grantee organizations promised to share research findings with their data partners and to explore how lessons learned could benefit both past and future participants. Preliminary findings have been shared with Hawaii area hospitals at the annual statewide conference organized by the Hawaii Health Information Corporation, and the California grantee has prepared a follow-up survey to assess how participation in the study has affected hospitals' attitudes and practices regarding the acquisition and use of race and ethnicity data.

Future research and development will be required to develop operationally sound strategies that will enable health data organizations to create enhanced claims databases to support CER while providing measurable financial and competitive advantages for the organizations contributing the data.

DISCUSSION

In general, all grantees were successful in creating useful enhanced administrative databases, but in virtually every case, initial plans had to be modified to overcome unanticipated difficulties. Some grantees encountered challenges with recruitment of data partners that had to do with concerns about the financial burden involved in database development. Grantees had to expend considerable effort to create and communicate tangible benefits and incentives to providers. These included monetary incentives for participation as well as nonmonetary inducements such as customized tools and reports for the data partners. Other challenges around recruitment stemmed from the additional burden providers would face as a result of data collection requirements.

Grantees also encountered technical hurdles such as data providers not being well equipped to transmit their data using commonly accepted data transmission standards such as HL7 or the submitted data not being formatted to specifications. To address recruitment issues as well as technical hurdles, grantees reduced the scope of their requests and shifted some of the technical and operational work on to themselves, thereby reducing the burden on providers.

While the grantees were generally successful in navigating the unexpected hurdles related to data collection, many problems transitioning from initial database development to a self-sustaining program remained unresolved when this research was completed. Understanding the challenges as well as the different approaches used to overcome them may aid other researchers contemplating or undertaking similar data enhancement projects and increase the likelihood that the resulting data infrastructures will be comprehensive, valid, usable, and sustainable.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: Drs. Salemi, Miyamura, and Zingmond, and Ms. Katz and Mr. Schindler were part of project teams that received grant funding from the Agency for Healthcare Research & Quality (AHRQ) to enhance the clinical content of and improve race/ethnicity identifiers in statewide all-payer hospital claims databases that are part of AHRQ's Healthcare Cost and Utilization Project (HCUP). Dr. Pine performed work on

this project as a subcontractor to the Minnesota and New York state grantees. Dr. Kowlessar performed work on this project as a contractor to AHRQ.

Disclosures: None.

Disclosures: None.

REFERENCES

- Anon. n.d.a “Auto-Match Software” [accessed on March 11, 2014]. Available at <http://www.innovadex.com/Inks/Detail/5770/203382/Auto-Match-Software>
- Anon. n.d.b “LinkageWiz Data Matching Software” [accessed on March 11, 2014]. Available at <http://www.linkagewiz.net/index.htm>
- Anon. n.d.c “Structured Query Language (SQL)” [accessed on March 11, 2014]. Available at <http://www.mysql.com/>
- Anon. n.d.d “The Link King: Record Linkage and Consolidation Software” [accessed on March 11, 2014]. Available at <http://the-link-king.com/>
- Blakely, T., and C. Salmond. 2002. “Probabilistic Record Linkage and a Method to Calculate the Positive Predictive Value.” *International Journal of Epidemiology* 31 (6): 1246–52.
- Blumenthal, D., and M. Tavenner. 2010. “The ‘Meaningful Use’ Regulation for Electronic Health Records.” *New England Journal of Medicine* 363 (6): 501–4. doi:10.1056/NEJMp1006114.
- Bradley, C. J., L. Penberthy, K. J. Devers, and D. J. Holden. 2010. “Health Services Research and Data Linkages: Issues, Methods, and Directions for the Future.” *Health Services Research* 45 (5 Pt 2): 1468–88. doi:10.1111/j.1475-6773.2010.01142.x.
- Centers for Disease Control and Prevention. n.d. “Registry Plus™ Link Plus Technical Information and Installation” [accessed on March 11, 2014]. Available at http://www.cdc.gov/cancer/npcr/tools/registryplus/lp_tech_info.htm
- Centers for Medicare & Medicaid Services. n.d. “EHR Incentive Programs” [accessed on March 11, 2014]. Available at <http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/index.html?redirect=/ehrincentive-programs>
- Clark, D. E., and D. R. Hahn. 1995. “Comparison of Probabilistic and Deterministic Record Linkage in the Development of a Statewide Trauma Registry.” *Proceedings/the . . . Annual Symposium on Computer Application [sic] in Medical Care. Symposium on Computer Applications in Medical Care*: 397–401.
- Concato, J., E. V. Lawler, R. A. Lew, J. M. Gaziano, M. Aslan, and G. D. Huang. 2010. “Observational Methods in Comparative Effectiveness Research.” *The American Journal of Medicine* 123 (12): e16–23. doi:10.1016/j.amjmed.2010.10.004.
- Dalton, J. E., L. G. Glance, E. J. Mascha, J. Ehrlinger, N. Chamoun, and D. I. Sessler. 2013. “Impact of Present-on-Admission Indicators on Risk-Adjusted Hospital Mortality Measurement.” *Anesthesiology* 118 (6): 1298–306. doi:10.1097/ALN.0-b013e31828e12b3.

- Fleurence, R. L., H. Naci, and J. P. Jansen. 2010. "The Critical Role of Observational Evidence in Comparative Effectiveness Research." *Health Affairs* 29 (10): 1826–33. doi:10.1377/hlthaff.2010.0630.
- Health Level Seven International. 2014. About HL7. [accessed on March 11, 2014]. Available at <http://www.hl7.org/index.cfm>
- Jaro, M. A. 1995. "Probabilistic Linkage of Large Public Health Data Files." *Statistics in Medicine* 14 (5–7): 491–8.
- Jordan, H. S., M. Pine, A. Elixhauser, D. C. Hoaglin, D. Fry, K. Coleman, D. Deitz, D. Warner, J. Gonzales, and Z. Friedman. 2007. "Cost-Effective Enhancement of Claims Data to Improve Comparisons of Patient Safety." *Journal of Patient Safety* 3 (2): 82–90. doi:10.1097/01.jps.0000242988.01413.fb.
- LaVeist, T. A., D. Gaskin, and P. Richard. 2011. "Estimating the Economic Burden of Racial Health Inequalities in the United States." *International Journal of Health Services: Planning, Administration, Evaluation* 41 (2): 231–8.
- Marko, N. F., and R. J. Weil. 2010. "The Role of Observational Investigations in Comparative Effectiveness Research." *Value in Health* 13 (8): 989–97. doi:10.1111/j.1524-4733.2010.00786.x.
- Mason, C. A., and S. Tu. 2008. "Data Linkage Using Probabilistic Decision Rules: A Primer." *Birth Defects Research. Part A, Clinical and Molecular Teratology* 82 (11): 812–21. doi:10.1002/bdra.20510.
- Méray, N., J. B. Reitsma, A. C. J. Ravelli, and G. J.onsel. 2007. "Probabilistic Record Linkage Is a Valid and Transparent Tool to Combine Databases without a Patient Identification Number." *Journal of Clinical Epidemiology* 60 (9): 883–91. doi:10.1016/j.jclinepi.2006.11.021.
- Pine, M., H. S. Jordan, A. Elixhauser, D. E. Fry, D. C. Hoaglin, B. Jones, R. Meimban, D. Warner, and J. Gonzales. 2007. "Enhancement of Claims Data to Improve Risk Adjustment of Hospital Mortality." *Journal of the American Medical Association* 297 (1): 71. doi:10.1001/jama.297.1.71.
- Regenstrief Institute. 2014. "Logical Observation Identifiers Names and Codes (LOINC)" [accessed on March 11, 2014]. Available at <https://loinc.org/>
- Salemi, J. L., J. P. Tanner, M. Bailey, A. K. Mbah, and H. M. Salihu. 2013. "Creation and Evaluation of a Multi-Layered Maternal and Child Health Database for Comparative Effectiveness Research." *Journal of Registry Management* 40 (1): 14–28.
- Salihu, H. M., A. Salinas, and M. Mogos. 2013. "The Missing Link in Preconceptional Care: The Role of Comparative Effectiveness Research." *Maternal and Child Health Journal* 17 (5): 776–82. doi:10.1007/s10995-012-1056-1.
- Salinas-Miranda, A. A., M. C. Nash, J. L. Salemi, A. K. Mbah, and H. M. Salihu. 2013. "Cutting-Edge Technology for Public Health Workforce Training in Comparative Effectiveness Research." *Health Informatics Journal* 19 (2): 101–15. doi:10.1177/1460458212461366.
- Shah, B. R., J. Drozda, and E. D. Peterson. 2010. "Leveraging Observational Registries to Inform Comparative Effectiveness Research." *American Heart Journal* 160 (1): 8–15. doi:10.1016/j.ahj.2010.04.012.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.