Current and Innovative Practices in Data Quality Assurance and Improvement

First in a Series under the project "Improving the Utility and Comparability of Health Care Data for Health Services Research, Policy Decisions and Transparency Reports"

National Association of Health Data Organizations

This project was supported by grant number R13HS026663 from the Agency for Healthcare Research and Quality. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality."

Background

Data quality assurance and improvement is a foundational component of the state health data organization (SHDO). While Data quality is important to all data systems, it is especially central to credibility of the SHDO and the data it produces. Today SHDOs maintain 48 inpatient hospital discharge data reporting programs. Over 20 states are in various stages of All-Payer Claims Database (APCD) implementation. These SHDOs face unique challenges related to statewide data collection and public dissemination of the data and are seeking innovative solutions to continuously improve the quality of the data they maintain.

In a three-year Small Conference Grant project[1], funded by the Agency for Healthcare Research and Quality (AHRQ), NAHDO drills down into three topical areas that represent key functional domains of SHDO practice, beginning with data quality assurance and improvement in the first year. Each of these domains are integral to producing high-quality data and information for policy, research, and industry use. The topical areas and NAHDO's focus[2] for multi-state collaboration are as follows:

1) **data quality assurance and improvement**

   NAHDO focus: Identify innovative approaches to assessing and improving the quality and credibility of the underlying data SHDOs maintain.

2) **data enhancement and linkage**

   NAHDO focus**:** Enhancing data and filling gaps by linking and integrating multiple data sets pose technical and legal challenges. Guidance around statistical and governance solutions is needed.

3) **analytics and public reporting**

   NAHDO focus: Guidance to develop actionable reports, innovating reporting tools, and communicate the value of information to policy makers, employers, consumers, and others.

Data quality is the topic for this first year and is the topic of this white paper titled "Current and Innovative Practices in Data Quality Assurance and Improvement". The content of this report draws on almost two years of work by NAHDO's Data Quality Forum and the First Data Quality Workshop, held in Deer Valley, Utah in October 2018, and follow-up working webinars. The white paper highlights challenges and notable practices in data quality improvement and highlights the results and status of the first-ever Data Quality Forum Benchmarking Pilot Project and the dissemination of the APCD Common Data Layout (APCD-CDL ™). A summary of the 2018 Data Quality Workshop is available on the NAHDO website.

[1] This report, the first in a series of three topics, was written by the National Association of Health Data Organizations (NAHDO) as an activity under a three-year conference grant project award titled *"Improving the Utility and Comparability of Health Care Data for Health Services Research, Policy Decisions and Transparency Reports" (R13HS026663)* in the AHRQ category of "Dissemination and Implementation."

[2] Data Quality Improvement Workshop (https://www.nahdo.org/sites/default/files/2019-08/workshop%20summary%20and%20recommendations%202018_0.pdf) was held October 12, 2018, in Deer Valley, Utah, as a part of NAHDOs 33rd Annual Meeting, Health Data Summit 2018. The day-long October 12 Data Quality Workshop included approximately 60 individuals who were mostly data officials from state agencies that manage facility discharge data and/or All-Payer Claims Database (APCD) reporting systems as well as data management vendors and private sector database managers

Data Quality Assurance and Improvement

States have their own data quality protocols (tools, methods, processes) and have also adopted fairly common practices such as using data dictionaries in data submission guides for collection purposes. The Data Quality Forum provides a platform to further understand these existing practices and facilitates a structure for ongoing cross-state exchange and shared learning to advance data quality approaches across states.

Challenges to high-quality data are numerous. Because statewide data systems collect data across disparate data sources and different platforms, it is essential to have standard definitions, technical specifications, and solid compliance processes in place to manage the submissions. Standardized data dictionaries, three-step audit processes, and submitter feedback in the form of dashboards are tools many states use to improve data quality and timeliness. Technical solutions are essential, but the human factor cannot be under-estimated. States emphasize the importance of good working relationships with data submitters and inclusion of submitters in validation actions and using data in a way that matters to both users and submitters.

Unique Challenges of the SHDO: An Example

Florida's Agency for Health Care Administration hospital discharge data system collects data from 1000 facilities, 330 hospitals, and 600 AD free standing ambulatory surgery centers. AHCA process all data internally in xml format, applying 700 edits for outpatient records, applying auditor software. The AHCA data management, vendor calculates a norm report of data elements selected, containing average and standard deviations for a year's worth of data. A second threshold report is generated showing the distribution of each key data element and informing AHCA and submitters which fields may be problematic and which are reliable. AHCA asks each facility to review and validate their submittal norm report. For example race fields for one hospital identified a large amount of mapping errors. Hospitals frequently change their vendors and even hospitals using the same vendor have variations in their submittals to AHCA. *2018 Data Quality Workshop*

Shared Understanding of Data Quality

A shared understanding of data quality is the first step for states seeking common approaches and solutions. As a "straw-man" case, workshop participants agreed that the Maine Health Data Organizations (MHDO) definition and characteristics of data quality provides a framework/context for understanding and improving data quality.

> *Data quality is an "assessment of a data's ability to serve its purpose in a given context. If you apply valid statistical techniques, the user will be able to conduct accurate/correct analysis.*[3]

Characteristics of data quality can help SHDOs assess the quality of their data system:

- Accuracy: degree to which fields are close to their true value
- Completeness: degree to which expected information is received/fields populated

[3] HSRI slides from Data Quality Workshop: https://www.nahdo.org/sites/default/files/Candura_NAHDOPresentation141008revised.pdf

- Integrity: degree to which information is valid, consistent, reliable
- Relevance: if the information important to the users
- Timeliness: data currency after data are cleaned, new data integrated, and extracts developed/released

While accuracy, integrity, and timeliness has been improving in most hospital discharge and claims data sets, SHDOs must be realistic in their data quality approaches related to completeness and relevance. Because the originating data are based on claims transactions designed to administer and pay claims, there are missing data elements[4] and missing populations[5], limiting the usefulness of the data for some use cases. For example, APCD source data is from payers who use claims data to adjudicate a claim and may not retain the entire claim in their data warehouses. Capitated health plan arrangements will not have financial fields at the encounter level (paid, allowed amounts). These gaps are expected due to the nature of the data sources, not due to data quality. Therefore, the scope of SHDO data collection and validation policies and practices related to data quality should be guided by these realities.

Components of Data Validation

Data validation and post-production data validation are comprised of specific technical protocols related to the intake and processing of submitted data. A third component is how the SHDO communicates the validation process with stakeholders. An open and transparent process that is communicated to all stakeholders is important to instilling trust in the data. This communication also serves to help manage user expectations by being be clear about data quality issues and data set limitations.

Incoming Data Validation:

> *Incoming data validation includes activities related to processing at the time of data submission, prior to importing data into the data warehouse*.

Key Considerations to Incoming Data Validation

- Technical documentation is foundational to a transparent and open data system process. A common set of data elements, based on provider and payer reporting capacity contributes to data quality. The use of state-specific or "home-grown" codes adds to provider and payer reporting burden and likely results in poor compliance to SHDO reporting requirements. .
- Data completeness is difficult for the SHDO to determine. Assessment of key fields for overall and key data element validity by comparing the submitter's historic and current submissions for unusual patterns (spikes or dips) in total records, total charges/dollars, per member per month (PMPM) metrics, and for other anomalies.
- Submittal specifications for data collection coupled with timely audit feedback reports to hospitals/payers with key benchmark metrics they can validate and compare are key to the data quality process
- Compliance to reporting requirements requires a balance of enforcement and incentives. Provisions for penalties for non-compliance to reporting requirements as well as demonstrating to submitters that sending "good" clean data through comparative benchmark feedback reports, portals, and return of usable data for their own uses.

[4] Missing data may include financial information from Alternate Payment Models (capitated payments, pay for performance, etc.) and demographic and socio-economic data elements not collected or retained by submitting organizations because those fields are not needed to adjudicated a claim or not available to the reporting entity

[5] Missing populations may include data exempt from state reporting requirements: Veterans Administration, Indian Health Service, TriCare/FEHBP, ERISA Self-funded plans, and 42 CFR Part 2 Substance Use Disorder

- An exemption/extension process is important to provide for flexibility in the compliance process. The SHDO should have in place a process for submitters to request extended time to comply with reporting requirements or request exemptions if their systems don't collect the data they are expected to report. The process permits a mechanism to establish a dialogue between the submitting organization and the agency to resolve reporting issues and work towards improvements.
- Need for national standardization. We discuss national standard formats later in this paper, but there is general agreement that aligning submission specifications with national standards across states reduces submitter reporting burden, improves data quality, and facilitates analytic operability across states.

Notable Best Practices in Incoming Data Validation

Automated data quality checks and edits applied at submission and timely submitter feedback. The longer the time is between reporting and feedback, the more difficult it is for the submitter to pull the records in question and correct them.

Data quality dashboard for submitters: Timely feedback to submitters not only reduces the data processing timeline, but alerts submitters to potential data issues which are easier to correct right way.

Maintaining a "live spreadsheet" of known data quality issues: When a data issue is identified or resolved the log is updated. The log contains information like which database was the problem originally identified, which data submitter was impacted, what was the issue, if the data was resubmitted, time frames, if the issue was validated, if it is an open or closed issue, date closed, and closing notes.

Fines/Penalties: One state fines for non-compliant submitters based on how many times the entity has been delinquent before. Another state moved from a $500 fine and started to include this in provider global budgets so that facilities are aware of consequences of not submitting.

Maintain raw files: Occasionally problems with a data submission is identified after the data editing process. The SHDO should maintain the raw files, storing them securely, to reconcile any downstream issues later.

Post-Production Validation

> *The activities in this stage of data quality are related to improving data quality after the initial data intake validation occur and consists of final checks prior to loading the data into the data warehouse.*

It was clear that much work is yet to be done to clarify and document post-production processes and practices. States (and their vendors) vary in how they approach post-production validation as well as timelines. For example, time between when data intake and post processing validations range from 45 days to three months; though apparently, vendor processing times have been decreasing in past years.

Value-added enhancements and data transformation are also a part of the post-production process also vary across states (and their vendors) in how they address the addition of groupers, Master Patient Index, and recoding/data element transformations.

Key considerations for developing consistent post-production validation across states:

As states examine improvement activities around unique and common practices of post-production validation, the following will guide efforts to identify business rules and common practices:

- What is post versus pre-validation?
- What validations are most effective for post-processing
- What is an analytic data set?

Because of the difficulties related to delving into post-production validation processes, the workshop participants divided post-production practices into data submitter and data user considerations as a way to design and improve a data quality toolkit and process.

Data submitter practices identified that improve data quality:

- Communication/strong feedback loop with submitters. Submitters provide explanations of 'anomalies' frequency
- Standard error reports to submitters
- Increase data use by submitters so that they see value in submitting accurate, complete data
- Fixed schedules, that is clear timelines for internal data review and time to provide data submitters with feedback

Data user components identified that have help to improve data quality:

- Produce data relevant reports, such as provider analyses to employers and other end users
- Define a process to receive feedback from internal data users
- Regular meetings with data users
- Use focus areas/hot topics-Opioid use researchers helped raise data redaction issues in 42 CFR records
- Create data user specific validations
- Analyst cafes to get feedback
- Quality sharing with data users
- Lunch and Learns: share research and data concerns/limitations
- Standardized variables in public use files for public use files, different use cases
- Standardized extract variables by type (hospital/APCD/etc.)
- Standard data products

Notable Practices in Post-production validation

- When reviewing submitted data, agencies should assess for reasonableness against key metrics such as PMPM changes, number of records/claims/members.
- Check for ungroupable records which should be close to zero
- Comparison of post-production results to unprocessed/raw submitted data
- Assess stability in groupers by major conditions (DRGs, APR-DRGs, etc)
- Comparing SHDO data with other data sources (vitals, provider directories, cancer registries, hospital, APCD) for reference checks for the completeness of data is recommended, if possible.
- Trend and longitudinal edits (quarter/annual frequency comparisons) to identify unusual patterns (spikes or dips) indicating reporting issues to discern data submitter errors or changes explained by other factors.
- Know the most important data elements users need/use. One state logs the elements that people look at the most in their system, focusing data quality efforts on those fields that get used all the time, spending less effort on elements used rarely.

There is a need to more fully develop consensus and guidelines on key post-production processes and identify best practices for all states, all payers, with some language around exceptions and expectations. The following recommendations will be addressed as resources become available:

- More discussion among states about national carriers to identify common issues
- Standardized validations for common variables that are contained in states' public use files
- Patient identifier practices, especially related to Social Security Number guidance, especially with Medicare's policy to use a beneficiary identifier for claims

Communicating data quality with end-users

As stated earlier in this report, SHDOs need to be transparent and open with their stakeholders to help them understand reasons for technical approaches and release decisions. Also, communicating the limitations of the data, recognizing that some use cases are not appropriate for the data. A range of strategies are needed to keep stakeholders informed.

Key considerations for stakeholder communication about data quality

- Data agencies must make difficult decisions about how much and what to communicate to users and when. Different stakeholders need various types of information.
- Submitters need robust technical documentation and timely feedback about their data.
- Users need user manuals and data quality notes about issues that may affect their analyses. Technical documentation should be made available to data users that communicates steps taken to ensure the quality of the data such as data submission guide, data book with basic information (#claims, #submitters, #members), user manuals for limited use files, include payer history in submissions.
- Power/sophisticated users may require close personal communication and this option should be made available.

State Data Agencies are leaders in managing and distributing data for public and restricted uses. The lessons learned across SHDOs is extensive, so notable practices in release and user relations are numerous, highlighted in the boxes below.

Notable Practices in Release Documentation

- One state posts a data book with detailed information about the files they make publicly available (total submitters, total claims, etc.) to help users interested in requesting data.
- Obtain vendor QC reports, with basic statistics, running list of enrolled
- For identified data issues, define the scope of the problem(s) and work with users to suggest work arounds and fixes
- Include release notes about the data, including versioning information, at the time the extract was generated.
- A data cover sheet includes a data discovery log, documentation of known data quality issues, a description of the issue and extensiveness of that issue (payers/claims affected) and options for work-arounds
- Quarterly updates are provided for users with extracts
- SAS code, developed by the SHDO, calculates member months for eligibility file for users
- SHDOs have data re-release policies for errored or problematic data files. A re-release is indicated when it's the SHDO's "fault" and a new file is issued to data users. If the error is due to an individual data submitter mistake, it may not warrant a re-release issue (a notice to data users may be warranted) without re-issue unless there is a significant error warranting re-issue.

Notable Practices in User Engagement/Education/Outreach:

- Data User Groups
- Agency presentations/guided learning materials
- Push and Pull communications:
  - Pull: Systematic way for users to submit questions. How and who to contact.
  - Pull: Direct email and/or a service desk to triage/vet questions
  - Push: Follow-up communication from User Group meetings posted on website and email
  - Push: Youtube and Data Academies—classes to potential users of APCD
  - Push: Data Users Groups and quarterly updates
- Agency internal training for those working on and with data set

Call for Multi-state Action:

Workshop discussions identified unmet needs that NAHDO' could coordinate to help states assess and improve their data quality methods. Two of these priorities are being actively implemented and will be discussed in the sections below:

- Develop a process for generating data quality benchmarks for APCDs

- Establish a process for disseminating the APCD Common Data Layout (APCD-CDL^{TM}) for reporting

Activities, such as formation of a national APCD Users Group and a national APCD-CDL™ Advisory Committee will require additional resources outside of the scope of this project. Engaging national stakeholders and users of the data through a national user group or forum, including HCUP, is a long-term initiative that will require additional resources to fully implement, making this a future action item as resources become available.

Data Quality Benchmarking Pilot Project

States requested that NAHDO establish a process for collecting and comparing cross-state data quality metrics----"So we all are not working in a vacuum". A set of data validation metrics that all states could calculate and report that will shine a light on data quality across states' data systems.

In response, NAHDO's Data Quality Forum[6] focused its 2019 activities to develop and implement a pilot data quality benchmarking project. The objective of the pilot is to **"identify cross-state metrics to compare data quality benchmarks and to test the feasibility and utility of shared data quality metrics for SHDO data sets."**

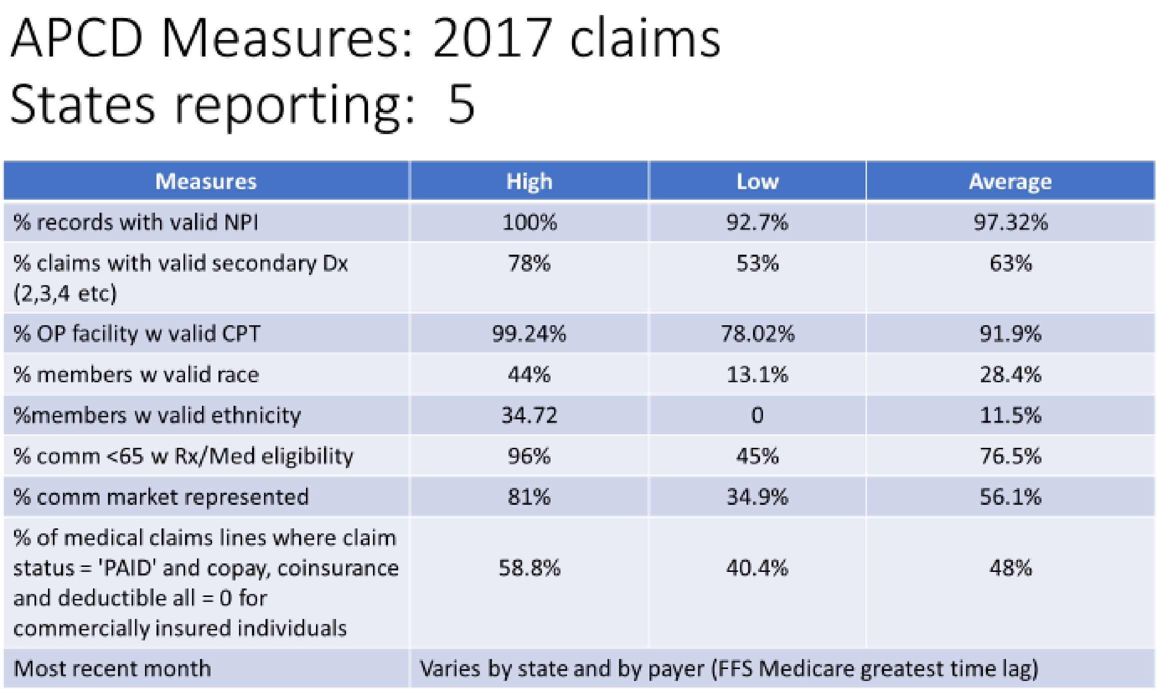
The Forum developed a set of proxy measures of data quality for each major data set (hospital[7] discharge and APCD[8]) that each state should be able to readily calculate and report to NAHDO for compilation. A list of the consensus data quality benchmarks are included in Appendix 1.
Through several meetings over several months via webinars, we gathered input and consensus on an agreed upon set of top (10 or so) data validation metrics (pre and post-production) to compare across states.

Following a state test of these metrics with their state data, documenting the time required to run the 2017 data sets and documenting assumptions made for generating the measures, a data call issued June 25, 2019 for CY 2017 APCD and Hospital Discharge Data Sets. The pilot results are displayed below.

[6] https://www.nahdo.org/data_resources/data_collection_management

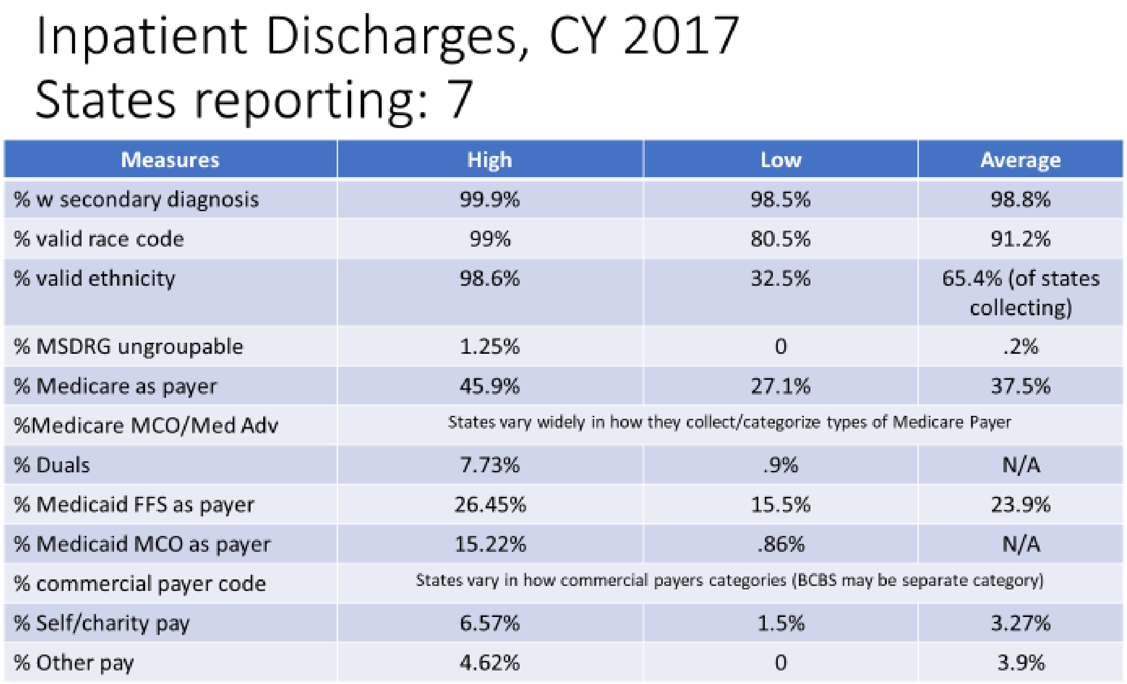
[7] QA Metrics for Discharge Data Systems: https://www.nahdo.org/sites/default/files/2019-09/Discharge%20Data%20Validation%20Metrics.pdf

[8] QA Metrics for APCD data systems: https://www.nahdo.org/sites/default/files/2019-09/APCD%20Data%20Validation%20Metrics.pdf



APCD Measures: 2017 claims
States reporting: 5

Measures	High	Low	Average
% records with valid NPI	100%	92.7%	97.32%
% claims with valid secondary Dx (2,3,4 etc)	78%	53%	63%
% OP facility w valid CPT	99.24%	78.02%	91.9%
% members w valid race	44%	13.1%	28.4%
%members w valid ethnicity	34.72	0	11.5%
% comm <65 w Rx/Med eligibility	96%	45%	76.5%
% comm market represented	81%	34.9%	56.1%
% of medical claims lines where claim status = 'PAID' and copay, coinsurance and deductible all = 0 for commercially insured individuals	58.8%	40.4%	48%
Most recent month	Varies by state and by payer (FFS Medicare greatest time lag)		



Inpatient Discharges, CY 2017
States reporting: 7

Measures	High	Low	Average
% w secondary diagnosis	99.9%	98.5%	98.8%
% valid race code	99%	80.5%	91.2%
% valid ethnicity	98.6%	32.5%	65.4% (of states collecting)
% MSDRG ungroupable	1.25%	0	.2%
% Medicare as payer	45.9%	27.1%	37.5%
%Medicare MCO/Med Adv	States vary widely in how they collect/categorize types of Medicare Payer		
% Duals	7.73%	.9%	N/A
% Medicaid FFS as payer	26.45%	15.5%	23.9%
% Medicaid MCO as payer	15.22%	.86%	N/A
% commercial payer code	States vary in how commercial payers categories (BCBS may be separate category)		
% Self/charity pay	6.57%	1.5%	3.27%
% Other pay	4.62%	0	3.9%

Benchmark Pilot Findings[9]

[9] Data Quality Forum slides of Pilot Findings: https://www.nahdo.org/sites/default/files/2019-09/DQ%20benchmark%20slides.pdf

Based on the early submissions of data quality metrics, there is much less variation in quality for inpatient hospital discharge data (which is not surprising, given that these data systems have existed longer than APCDs). There are findings in which cross-state sharing of stand-out practices might be indicated, including the following:

- Why are hospital discharge databases able to collect race and ethnicity at a very high level in most states but APCDs across the board cannot?
- States with 100% valid NPI for APCD should share with other states how they achieve this level of accuracy.
- We'd like to learn from SHDOs that release APCD data for periods without Medicare FFS (due to time lag of Medicare data extracts). Should SHDOs wait for CMS data before doing more comprehensive reporting?
- There is wide variation in the percentage of commercial market represented in some state APCDs versus others. What are states that are getting the highest percentages doing to drive opt-ins of the ERISA Self-funded data populations?
- States are coding commercial payer categories differently, challenging comparisons and benchmarks. There is a high congruence of Medicaid, Medicare, self/charity payer categories but payer categories FFS, MCO, MHO are not comparable across states

Data Quality Benchmarking Pilot Project Next Steps

For the purposes of the pilot, we focused on hospital inpatient and APCD benchmarks to start, with the intent to expand the benchmarking efforts to other data sets (ED and Ambulatory Surgery) in the future.

We obtained permission from participating SHDOs to share the results of their benchmarks for hospital discharge and APCD data with other participating states.

Next steps for the Data Quality Forum include:

- Recruit additional states to submit data quality metrics in order to increase the reliability of the benchmarks.
- Use webinar presentations by states doing well in key high variance measure to share their practices with other states.
- Present and discuss benchmark project at NAHDO's November meeting in Little Rock?
- Future work around commercial payer categories is needed and consultation with the Payer Typology Workgroup may be indicated.

Finalize and Disseminate the APCD-CDL™

National payers want state APCDs to migrate from state-specific formats and reporting requirements and adopt a set of common data elements and a common set of edits to reduce payer reporting burden. Beginning in May 2016, the APCD Council Learning Network[10] has been working with states, payers, and state APCD vendors to harmonize the data collected across state APCDs to develop a common core set

[10] The APCD Council Learning Network is a joint collaboration between the National Association of Health Data Organizations (NAHDO) and the University of New Hampshire's Institute for Health Policy and Practice.

of APCD elements in a common layout called the APCD-CDL™[11].As of this writing, the APCD Council has finalized the APCD-CDL™, aligning it with the X12N Post-Adjudicated Claims Data Reporting Guide (PACDR). The APCD-CDL™ was made widely available through the APCD Council website as of December 1, 2018. without cost, to registered users. To date, over 60 CDLs have been distributed to users across the country. Several states are in the process of adopting the APCD-CDL™ for their state reporting requirements.

There are cost considerations for states to change from their current state-specific reporting specifications to the APCD-CDL™; however, adoption of this standard format is expected to result in improved data quality and less administrative burden to maintain state-specific lists/tables of valid values. States that adopt the CDL may not collect every data element represented in the CDL, but they agree to follow the CDL format and not vary from that format.

Because APCDs are evolving rapidly in states, so are the data and information needs evolving. Therefore, states and the APCD Council have define a process to maintain and improve the APCD-CDL™ every two years has been established, with version two expected to be released in January 2021. Future plans for establishing reporting thresholds for reporting will be implemented as resources permit. The Data Quality Forum Benchmarking Pilot, discussed above, will help guide future discussion about setting data quality threshold guidance for states.

More information at https://www.apcdcouncil.org/common-data-layout)

Conclusion

The tools, methods, and processes discussed in this paper will facilitate SHDO assessment of their own data quality policies and practices. This paper also provides a roadmap for multi-state collaboration to develop shared resources going forward, beginning with the Data Quality Forum Benchmarking Pilot Project and the APCD-CDL™ maintenance process.

Data will never be perfect. The objective is to know at the point when it's "good enough". States have established their own rigorous data quality assurance and improvement activities but there is a high degree of interest in multi-state collaborative activities. During the Data Quality Workshop, states suggested a host of collaborative activities to help SHDOs improve data quality that ranged from data quality benchmarks, data quality threshold development, and a national users group and APCD user training modules.

Because of resource constraints, NAHDO has focused this project year's priorities on the Data Quality Forum Benchmarking Pilot Project and disseminating the APCD-CDL™. These activities lay the foundation for additional collaboration. As resources become available, the establishment of a national users group designed to promote APCD use and gather feedback about how to improve data utility and quality is high would be considered if a national partner/supporter could be identified.

[11] https://www.apcdcouncil.org/common-data-layout