Best Practices for Data Analytics Reporting Lifecycles: Quality in Report Building and Data Validation

While the importance of data quality in providing high-quality clinical care in today's healthcare setting is typically well understood, the quality of data for report building and validation activities is often not well articulated—and potential data quality issues that impact the accuracy of reports are a frequent, unwanted outcome. Quality data for reporting and validation is critical to ensure that business decisions based on data have positive outcomes. As a result, data quality must be fully understood and continually managed to avoid possible false conclusions or, even worse, negative outcomes.

This Practice Brief outlines best practices regarding data quality characteristics. Application of these characteristics can be applied to healthcare data to ensure success when building reports, validating data, planning methodologies, and analyzing data for both clinical and operational business needs.

Data Quality Management Model

AHIMA's Data Quality Management Model references a list of characteristics of data quality, which include:

- **Data Accuracy:** The extent to which the data are free of identifiable errors
- **Data Accessibility:** The level of ease and efficiency at which data are legally obtainable, within a well-protected and controlled environment
- **Data Comprehensiveness:** The extent to which all required data within the entire scope are collected, documenting intended exclusions
- **Data Consistency:** The extent to which the healthcare data are reliable, identical, and reproducible by different users across applications
- **Data Currency:** The extent to which data are up-to-date; a datum value is up-to-date if it is current for a specific point in time, and it is outdated if it was current at a preceding time but incorrect at a later time
- **Data Definition:** The specific meaning of a healthcare-related data element
- **Data Granularity:** The level of detail at which the attributes and characteristics of data quality in healthcare data are defined
- **Data Precision:** The degree to which measures support their purpose, and/or the closeness of two or more measures to each other
- **Data Relevancy:** The extent to which healthcare-related data are useful for the purposes for which they were collected
- **Data Timeliness:** The availability of up-to-date data within the useful, operative, or indicated time

Source: Davoudi, Sion et al. "Data Quality Management Model (2015 Update)." *Journal of AHIMA* 86, no. 10 (October 2015): expanded web version. http://bok.ahima.org/doc?oid=107773.

Data Quality Defined

To understand how to improve the quality of data reporting, one first must understand what is meant by the term data quality. Data quality simply means that the data that is being reported is meaningful and serves its intended purpose. The Centers for Disease Control and Prevention (CDC) has defined the six core data quality dimensions as:[1]

- **Completeness.** The data is comprehensive and complete. All data values are recorded.
- **Uniqueness.** Data is unique and one-of-a-kind. Duplicates are avoided.

- **Timeliness.** Data represents reality at the point in time in which it is collected.
- **Validity.** Data measures what it is intended to measure.
- **Accuracy.** Data is reflective of real-world values.
- **Consistency.** Data values are consistent across data sets. Data can be matched. There is no conflicting information.

These six dimensions can be managed through data quality management. Data quality management refers to "the business processes that ensure the integrity of an organization's data during collection, application (including aggregation), warehousing, and analysis."[2] Both data quality and data quality management are essential to the success of report building and data validation.

Data Collection and Report Building

Building a report starts with the data collection. This is especially important and pertinent for current health information management (HIM) practices with increased information technology and rapidly growing mountains of information. The first step is understanding the purpose of the data collection, different types and sources of data, and key factors that relate to building a report.

Purpose of Data Collection

Healthcare organizations collect healthcare data for different purposes, including:

- The ability to compare hospitals' performance with a peer group, especially with an organization of excellence, is beneficial in today's competitive environment. Benchmarking has become a common tool used even at the departmental level.
- Clinical decision support provides expert knowledge to healthcare providers to assist them in making the best decisions regarding patient treatment and care.
- The Medicare.gov website Physician Compare maintains information on hospitals, doctors, nursing homes, home health agencies, dialysis facilities, and drug and health plans for the Medicare beneficiary. All of this information is available to the public. One can compare information about the quality of care and services these providers and plans offer and obtain helpful tips on what to look for when comparing and choosing a provider or plan.
- Any healthcare organization or vendor collects and uses information to understand a population and make operational decisions with the purpose of improvement. It can be for quality, payment, productivity, accuracy, financial, resource management, or trending. Software is designed around data elements necessary to capture the information necessary for use. More health information and informatics management professionals are needed in the development of health information technology because HIM professionals have knowledge of the information they are trying to capture. Excluding HIM in decisions causes other problems instead of providing a solution, which ends up costing organizations more money.

Data Types

Healthcare data includes different data types and data from different sources. Major types of healthcare data are usually clinical data, administrative data, financial data, operational data, and population health data.

- **Clinical Data**—Information related to the patient's condition, course of treatment, and progress.
- **Administrative Data**—Includes financial data and operational data. Financial data is non-clinical and operational data can either be clinical or non-clinical data. They are both related to the administrative aspects of the healthcare facility.
    - **Financial Data**—Information related to the financial health of the organization.
    - **Operational Data**—Information related to the organization's efficiency and productivity. Clinical operational data could include readmission rates, adverse incidents, Centers for Medicare and Medicaid Services (CMS) reporting, and average length of stay (ALOS) information.
- **Population Health Data**—Often used for research to study people as a whole instead of individual patient data, such as data used in epidemiological studies, national or regional surveys, or indexes.

Data can be further defined as structured data and unstructured data. This categorization of the data can help the users to better understand how to collect, manage, and use the data for report building.

- **Structured data** (also called discrete data) refers to data that have been predefined in a table or checklist.
- **Unstructured data** includes narrative notes as well as images (such as of scanned documents or medical images like x-rays).

Source of Data

Healthcare data come from different sources. Some are derived internally within the organization while others are originated externally. The major healthcare data contributors include patients, providers (organizations and individuals), vendors, and researchers.

- **Internal:** The organization's data that can be found in electronic health record (EHR) systems, ancillary software, financial software (registration, billing, coding), or other software used to manage a process or people. A facility or provider is the center of healthcare data. It is where information about a patient is created and transferred to a payer. Both a healthcare organization and a payer submit information to government agencies, which use the information for payment purposes and/or creation of regulatory requirements.
- **External:** Data collected and used by multiple sources, such as vendors or any business associate contracting with the healthcare organization. Also included is required reporting data, such as regulatory reporting. Additionally, healthcare organizations also find other partners to build relationships with to obtain more information, which they can use to better understand and predict the target population for which services are provided.

Basic Guidelines for Tables and Schema

Each table should have a purpose and the representative diagram (also known as schema) should be structured in a way to eliminate redundancy of data. Every table should be a part of an entity relationship diagram (ERD) and each schema should show how it relates to other schemas within the organization if there is a relationship. If not, it still needs to be documented so the person maintaining or accessing the information understands how to appropriately use the data captured in the tables/schema.

Once the purpose of the report has been established, the user should select the appropriate tables and understand the constraints of each table to know how to join the tables together and extract the necessary data elements. Users need to understand each data element captured in the table and how it relates to the other information within a report. This is especially true if there are multiple rows written but only one column is different or needed. Understanding the data elements within the table will also help the report designer use the information that doesn't fall within the standard. Every piece of information is useful. If information is not used or weeded out, users have a potential for the non-standard data to grow or numbers to be an estimate and not an accurate picture of what is truly happening.

Types of Reports

There are different kinds of reports, based on the data type and purpose of the reports.

Operational reports include:

- Financial: These reports typically monitor denial management-related activities, outstanding items sent to collections, and payment history for various payors in relation to procedure/diagnosis codes.
- Resource-based: People-related items, such as productivity, accuracy, staffing additions or reductions, and inventory-related reports.
- Quality/Data Integrity: Ensure information being collected in any software application is accurate and processes or standards are being followed when it comes to data input.

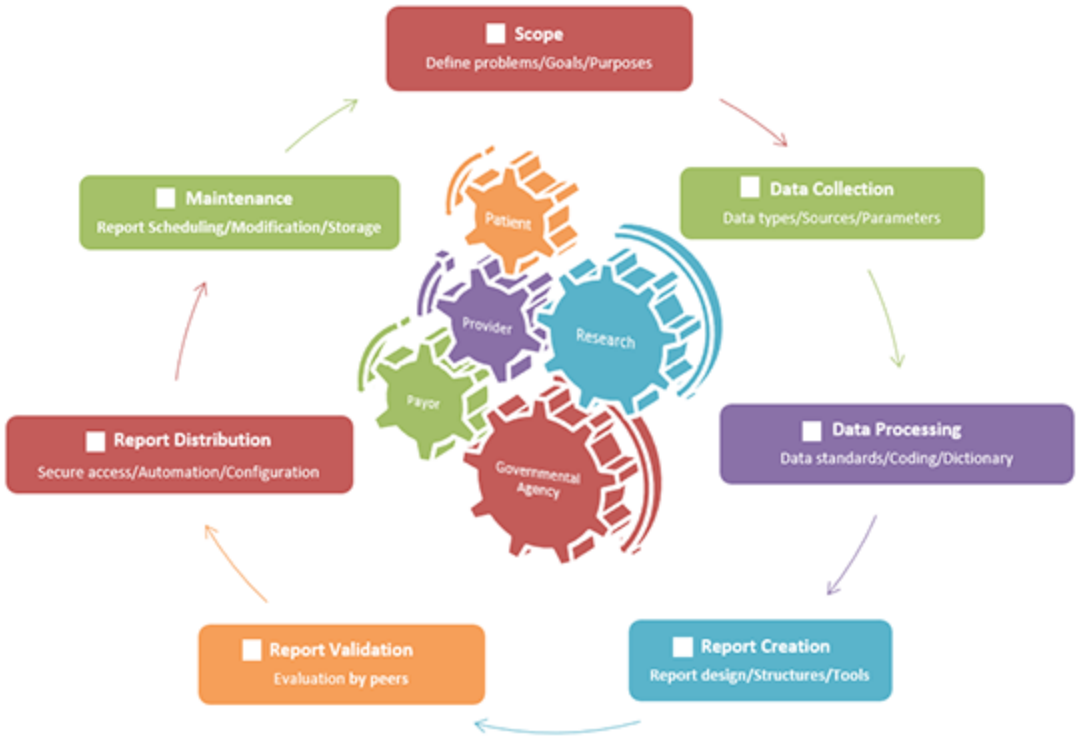
Clinical reports include:

- Treatment/Standards: For example, did each person with the same diagnosis have the same wait times? Were the same resources used?
- Accuracy: For example, a population pulled based on diagnosis codes for manual review of charts to ensure accurate documentation (such as for audits).
- Quality of Care/Improved Outcome: For example, reports required for public reporting initiatives.

Process for Reporting

The report process lifecycle, as shown in Figure 1 below, includes the following steps: Scope, Data Collection, Data Processing, Report Creation, Report Validation, Report Distribution, and Maintenance. These steps are further examined in the following sections.

Figure 1: Data Reporting Lifecycle



Scoping of Request

- **Clear explanation of the problem/situation the data request is expected to solve.** Ask the requestor, "What will you be using this data for?" Starting a conversation about how the customer plans to use the data and who will review the information will provide insight on whether they have requested all the information, or the right information, for their purpose.
- **Are there any business processes/procedures that would impact how the data is being captured?** Understanding whether a field is discrete (structured) or unstructured, and exactly how that data is populated, will often have implications for the report. As an example, one emergency department wanted to know how long it takes for a patient to be seen by a physician for the first time. The business practice is for the physician to "sign up" for the patient themselves upon entering the room. When they sign up, an automated timestamp is placed on the patient record. However, what was found was that some nurses and registration staff were not familiar with the policy and would "help" the physicians by signing them up. This created a break in the automation logic, which led to missing timestamps for these patients. The report writers then included a data quality section on the report identifying who was responsible

for the break-in process. This helped management to educate staff members and greatly improved the overall quality and accuracy of the report.

- **What are the data sources that will be used for this request (i.e., EHR, claims data, satisfaction survey data)?** Having an understanding of not only where the data need to come from, but also how the information will be accessed, is critical. Sometimes a data element might be captured in a source system but is not readily available in the reporting system. Offices that use a data warehouse might only include 70 percent of EHR data in that data warehouse. Based on the urgency of the request, will there be enough time to get that data element added to the reporting system, or are there suitable alternative measures that could be used instead? It is also important to understand the limitations of reporting resources. Some programs may have set date ranges and it is not possible to set a custom date range. Some programs have limited reports available. For example, each EHR only certifies for certain clinical quality measures (CQMs), so when selecting CQMs for reporting it is important to make sure the EHR can pull the measures the data analyst selects to report on. Otherwise, the data analyst will have to do claims reporting instead. Many reporting software types also vary on how quickly the report runs and how up-to-date the data will be. For example, users may select a report to run, but are then unable to see the report until 24 hours later. Some EHRs only update their data to reporting modules weekly, so a report requested on Wednesday will not include any data after the prior Sunday. Therefore, it is important to understand the limitation of the programs in order to better customize workflows.

Collecting the Data

- **Review data standards and/or data dictionary.** To ensure the relevancy and precision of the data healthcare organizations are collecting, report writers must be sure to check available reference materials for data definitions or standards. The difference between an effective date stamp on an order and the order submission date stamp can have important implications for the validity of the report.
- **Plan for reporting methodology.** Identify whether the request will be a one-time ad hoc request or if it will turn into a recurring (daily, weekly, monthly) report. Use this information to inform the methodology around collecting, manipulating, and reporting out on the data. Using Excel for some quick Pivot tables and charts can be a quick win for a one-time request, but the extra time invested in automation of the manipulation and visualization of the data will be well worth it for a recurring report.

Data Processing

- **Does the data story match with user understanding of the business process?** If the report writers are told that a process should only ever result in a certain set of values, they should watch the data to ensure that data or set of values is what they are seeing. Report writers should investigate the cases that don't fit the process. They should reach out to the requestor, clinical informatics, and/or other stakeholders to help identify the cause of the difference. The report writers should be able to provide a summary of what they are seeing and look for patterns. Report writers should not just assume these extra values should be filtered out of the population denominator.
- **What to do with missing data.** When report writers encounter missing data, they should follow the same steps as when they are checking their results against the business process. Do users expect to find missing data? If so, how much do they expect to be missing? Does this affect the accuracy of the report? If the user needs to know how many times a consult request is entered for Dr. X, but the physician name is missing on 60 percent of the consult orders, then the report writers are not going to be able to guarantee the accuracy of the request. In these cases, it is important to document on the report output any caveats or limitations that are noted surrounding the collection of this data.

Report Creation

Now that the report writers have established the purpose of the request, identified the audience who will view the data, and collected and manipulated the data, the writer needs to build the final report. Even this stage should come with careful planning and consideration of best practices.

- **End User Considerations:** In the book *The Design of Everyday Things*, author Donald Norman refers to the Fundamental Principles of Interaction.[3] These principles apply not only in system design, but in designing anything that a human may interact with. An affordance is a relationship between an object and an agent that defines possible uses or

makes clear how it can or should be used. This means that the person and the skill, education, position, or experience level of the person using the report to make decisions must be taken into consideration. What makes sense to the report writer may not make sense to every person using the report to drive a decision. Each end user will have a varied skill set, understanding, and ability. This is why report writers have to understand the process from beginning to end to make sure that they can provide the information necessary at the level of the person using the information. Understanding the users of the report will help report writers understand what actions are needed and how those actions need to be communicated to the end user.

- **Inclusion/Exclusion Criteria:** As a best practice, report writers should include a definitions page either on a cover sheet or as an appendix to each report. This should be where it clearly states the definitions used to define the population, assumptions made, inclusionary criteria, and/or exclusionary criteria. It is also best to include on the report when it was last updated. Having this information attached to the report reduces the likelihood of misinterpretation and confusion for recipients who may not have been part of the initial report creation. See Appendix A for an example of a report cover page, available in the online version of this Practice Brief in AHIMA's HIM Body of Knowledge at http://bok.ahima.org.
- **Filters:** When developing an interactive report, report writers need to understand what filters will be needed to allow the user to extract the information they need. These filters will need to have the appropriate labels, which help the end user understand how to interact with the report. Since each individual can have a different background, position, and experience level, it is important to ensure that the mapping of the information is consistent and appropriate so that each user that interacts with the report does not have to build their own interpretation of the information displayed in the report. If the information is not clear, a user can build assumptions or interpret the data based on their own experiences —which may not be the intention of the information displayed in the report. As with anything the report, if interactive, should have the appropriate feedback to let the end user understand if the report is working correctly or if the report is current.
- **Data Security:** When creating any report, consideration should always be taken regarding protected health information (PHI). If the report is automated, consider if the user can drill down into patient level details and if so, whether or not the report writer locked down the report so that only approved users can view this data. There may also be times in which report writers may want to consider sending the final report as a PDF document so that the values cannot be altered. Even the most seasoned analysts have on occasion typed over a cell in Excel by accident.
- **Data Visualization Considerations:** Data visualization can really take a report to the next level when done properly. Heat maps are a great example and have been popularized in reports emulating the balanced scorecard design. Done poorly, however, these visualizations can result in frustration and confusion for the end users. For a summary on the basic principles of data visualization best practices, see the article "Simplifying Your Data Visualizations" by Braden Tabisula, MBA, RHIA, CHDA, published in the May 2018 issue of the *Journal of AHIMA*.[4]

Report Validation and Distribution

After a report has been created, it should be validated before deployment and distribution. Depending on the report, there may be a few levels of validation. Initial validation should be performed by the report writer, wherein the basic build and results of the report are validated to ensure that the report is extracting data correctly. Next, the report writer should share the results with the requester for validation that the data transforms into meaningful information. At this stage, subject matter experts such as clinical informatics professionals, business analysts, finance analysts, or other department staff may be sought to further contextualize understanding of the data. As an example, consider a report looking at CPOE utilization. To better understand the data, context of order types, and workflows, a member of the clinical informatics team may provide valuable insight. Finally, the overall end user experience should be validated. If end user interaction with the report is not validated, it may render the report less valuable, as the end user may not use the data within the report appropriately or misunderstand the overall purpose of the report. Modifying column placements, titles, or overall report structure may create a more meaningful end user experience, which will add to the overall value of the report.

Ongoing Maintenance

EHRs represent a dynamic ecosystem. This fact is especially important to remember as there are always more data to consume. Once a report has been created, validated, and deployed it becomes a resource that must be maintained. Report maintenance should be approached collaboratively, as there are technical and operational components that impact the relevance and validity of a report. Business analysts and report writers must work collaboratively from a technical perspective

to ensure that system updates, modifications, or new builds do not impact reports. Additionally, as new systems or modules are brought online, reports should be updated to incorporate new and pertinent data fields. End users must also be proactive and inform technical teams of upcoming regulatory and/or operational changes that impact reports.

Addressing the many requirements for quality in the report building and data validation process requires a carefully planned methodology to ensure that the necessary activities and tasks involved with acquiring, processing, analyzing, and sharing reports are managed effectively. The best practices outlined in this Practice Brief can be applied to all types of healthcare data to assist in successful report building and data validation to meet both the clinical and operational business needs of healthcare organizations.

Notes

1. Centers for Disease Control and Prevention. "The Six Dimensions of EHDI Data Quality Assessment." www.cdc.gov/ncbddd/hearingloss/documents/dataqualityworksheet.pdf.
2. Ibid.
3. Norman, Donald. *The Design of Everyday Things* (Revised and Expanded Edition). New York, NY: Basic Books, 2013.
4. Tabisula, Braden. "Simplifying Your Data Visualizations." *Journal of AHIMA* 89, no. 5 (May 2018): 36-39. http://bok.ahima.org/doc?oid=302493.

Prepared By

Jeannine Cain, MSHI, RHIA, CPHI
Lesley Clack, ScD, MS
Shannon Houser, PhD, MPH, RHIA, FAHIMA
Lesley Kadlec, MA, RHIA, CHDA
Raymound Mikaelian, MSHI, RHIA
Amanda Spears, MA, CHDA
Annemarie Wendicke, MPH, CHDA

Acknowledgments

Jane DeSpiegelaere
Diana Flood, MS, RHIA
Dawn Paulson, MJ, RHIA, CHPS
Laurie Peters, RHIA, CCS
Amy Richardson, RHIA, CHDA
Donna Rugg, RHIT, CDIP, CCS-P, CCS
Margaret Stackhouse, RHIA
Robyn Stambaugh, MS, RHIA
Maria Ward, MEd, RHIA, CCS, CCS-P
Lou Ann Wiedemann, MS, RHIA, CDIP, CHDA, CEPHR, FAHIMA
Jami Woebkenberg, MHIM, RHIA

Appendix A: Sample Report Cover Page

[TITLE]

[SUBTITLE]

Reporting Period: MMM-YYYY to MMM-YYYY

Lines of Business: [HMO, PPO, etc.]

[Claims Data as of MMM-YYYY]

Prepared by:

[Company Name]

[Company Address]

[Company City/State/Zip Code]

Appendix B: Tools for Reporting

Healthcare organizations may run standard (routine) or scheduled reports, as well as ad hoc (individual, on an as-needed or requested basis) reports. There are a variety of tools that can be used for reporting. One tool may be preferable than another due to cost, operational considerations, or functionality.

Some of the more common reporting tools used in healthcare organizations include:

- Microsoft Excel: According to Webopedia, "Microsoft Excel is a spreadsheet program included in the Microsoft Office suite of applications. Spreadsheets present tables of values arranged in rows and columns that can be manipulated mathematically using both basic and complex arithmetic operations and functions. In addition to its standard spreadsheet

features, Excel also offers programming support via Microsoft's Visual Basic for Applications (VBA), the ability to access data from external sources via Microsoft's Dynamic Data Exchange (DDE), and extensive graphing and charting capabilities."[1]

- Microsoft Access: According to Techopedia, "Microsoft Access is a pseudo-relational database engine from Microsoft. It is part of the Microsoft Office suite of applications that also includes Word, Outlook and Excel, among others. Access is also available for purchase as a stand-alone product. Access uses the Jet Database Engine for data storage. Access is used for both small and large database deployments. This is partly due to its easy-to-use graphical interface, as well as its interoperability with other applications and platforms such as Microsoft's own SQL Server database engine and Visual Basic for Applications (VBA)."[2]
- Crystal: According to Your Dictionary, Crystal is a "Popular reporting and analysis software for Windows from SAP that is used to retrieve data from more than 30 types of databases. Queries and reports can be made via a Web browser, and the functionality can also be added to proprietary programs written in languages such as C, C++, J++, Delphi and Visual Basic. Crystal Reports was originally a product from Crystal Decisions (formerly Seagate Software) and was acquired by Business Objects in 2003. It later became part of the SAP BusinessObjects family."[3] Some EHR systems use Crystal reports for their reporting modules.
- SAS: SAS is a software suite developed by the SAS Institute. This software enables users to perform advanced analytics, multivariate analyses, business intelligence, data management, and predictive analytics.
- Tableau: According to the University of Washington, "Tableau is easy-to-use business intelligence software used for data analysis, providing visual tools to help users see and understand their data. Users can connect to data in a few clicks, then use drag and drop tools to visualize and create interactive dashboards that can then be shared with Tableau Public. Visual analytics are at the core of the Tableau experience, allowing users to ask questions and see data in meaningful ways to understand it. These tools allow users to easily visualize data and spot trends to answer their own questions."[4]
- Qlikview: According to Tutorials Point, "QlikView is a leading Business Discovery Platform. It is unique in many ways as compared to the traditional BI platforms. As a data analysis tool, it always maintains the relationship between the data and this relationship can be seen visually using colors. It also shows the data that are not related. It provides both direct and indirect searches by using individual searches in the list boxes. QlikView's core and patented technology has the feature of in-memory data processing, which gives superfast results to the users. It calculates aggregations on the fly and compresses data to 10 percent of the original size. Neither users nor developers of QlikView applications manage the relationship between data. It is managed automatically."[5]
- Power BI: According to Microsoft, "Power BI is a suite of business analytics tools that deliver insights throughout the healthcare organization. Connect to hundreds of data sources, simplify data prep, and drive ad hoc analysis. Produce beautiful reports, then publish them for each organization to consume on the web and across mobile devices. Users can create personalized dashboards with a unique, 360-degree view of their business. And scale across the enterprise, with governance and security built-in."[6]
- R: According to The R Foundation, "R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering…) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity. One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control."[7]
- Python/Machine Learning: According to The Python Foundation, "Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms and can be freely distributed."[8]

Article citation:
Cain, Jeannine et al. "Best Practices for Data Analytics Reporting Lifecycles: Quality in Report Building and Data Validation." *Journal of AHIMA* 89, no. 9 (October 2018): 40-45.