CLINICAL QUALITY MEASUREMENT: A VALIDATION STUDY

By

OLUBUMI AKIWUMI

A DISSERTATION

Presented to the Department of Biomedical Informatics and Clinical Epidemiology
and the Oregon Health & Science University
School of Medicine
in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

March, 2017

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the PhD dissertation of

Olubumi Akiwumi

has been approved

Mentor/Advisor: Michael Chiang, MD

Member: Joan Ash, Phd

Member: Emily Campbell, RN, MS, PhD

Member: Karen Eden, PhD

Member: Thomas Hwang, MD

TABLE OF CONTENTS

List of Tables

Acknowledgements

I have many people to thank, particular my friends and family who supported me throughout this journey. I cannot thank you enough for your boundless love and support. I am thankful for my advisor, Dr. Michael Chiang, and committee members, doctors Joan Ash, Emily Campbell, Karen Eden, and Thomas Hwang. Thank you for your invaluable instruction and for helping me to achieve this goal. I would also like to thank all of the individuals in the various IT departments who lent their expertise and time to this project. Dr. Lieberman and Robert Schuff, the resources you shared made this project possible. Much cannot be accomplished without the staff of the Department of Medical Informatics and Clinical Epidemiology. Thank you all for guiding me through the warren of rules and forms and for the assistance you provided over the years. Last, but not least, I would like to acknowledge the sponsor of my fellowship, the National Library of Medicine (Training Grant T15LM00708822).

Thank you one and ALL.

Abstract

CLINICAL QUALITY MEASUREMENT: A VALIDATION STUDY

Olubumi Akiwumi, B.S., M.S., PH.D.
The Oregon Health and Science University at Portland
School of Medicine, 2017

Background and purpose

Today, in spite of significant advancements that have been made in information technology and medicine, the U.S. health care system continues to struggle with achieving consensus and implementing standards for the delivery of clinical quality across the health care system. As a result the U.S. health care community continues to struggle with uniformly disseminating quality care across all patients. Challenges with quality of care are compounded by the high cost of care. One way to address the issues of high cost and varying quality was to implement legislation to help obtain better health information to support clinical care. To this end, the Centers for Medicare and Medicaid Services (CMS) have instituted a policy requiring Medicare providers to report on clinical quality. The primary goal of the study was to assess the accuracy of measure guidelines using electronic health record (EHR) data and to compute quality measures using an EHR-based algorithm and the NQF guidelines.

Methods

This is a cross-sectional and mixed methods study. Clinical quality measures were implemented using two sets of measure criteria for the quantitative analyses. The diabetic retinopathy (DR) and primary open-angle glaucoma (POAG) clinical quality measures were implemented separately using the EHR algorithm and NQF guideline. Patients were classified to determine eligibility (denominator) and success in passing the measure (numerator) for each condition and guideline. Classifications were compared between the two methods of implementation and a reference standard for each condition. The reference standard was developed at the study site and

accessed all data including clinical notes that could not be used to implement the guidelines. Physicians were interviewed at the study site and their responses were analyzed for the qualitative evaluation.

Results

Quantitative Analysis: Detailed comparison of the three methods exhibited differences in patient classification for diagnosis, denominator and numerator. In all but one instance, accuracy in the categorization of patients for the denominator and numerator under the National Quality Forum (NQF) guideline superseded EHR algorithm classifications when compared to the reference standard. Another key finding was considerably higher accuracy for POAG over DR in the classification of patients across both methods. Denominator and numerator classifications for the POAG measure implementations were notably more similar to the reference standard, unlike the DR measure classifications.

Qualitative Analysis: Providers may be aware of national reporting programs through internal initiatives but are not necessarily exposed to or aware of the NQF ophthalmology quality measures. Clinicians participated in a formal training course on general navigation of the EHR system but did not participate in training that was specific to the ophthalmology domain. Lastly, all participants expressed the fact that the NQF measures assess activities that are routinely conducted during clinical care and that adherence was absolute.

Conclusion

The NQF guidelines demonstrated greater accuracy in the classification of patients in contrast to the EHR algorithm. One of the major findings of the study was that physicians may overestimate their adherence to best practice guidelines. The discordance between perceived performance on the quality measures and the computed measure outcomes could be due to documentation habits. This study underscored the drawbacks and challenges to clinical quality reporting and provides a path forward to advance the process.

1. Introduction

Information, which stems from data, is the foundation of decision-making. The need for and use of information, particularly standardized reports, on clinical care has been present for over a century. As quoted by Florence Nightingale in 1863 (1):

> "I am fain to sum up with an urgent appeal for adopting …some uniform system of publishing the statistical records of hospitals. There is a growing conviction that in all hospitals, even in those which are best conducted, there is a great and unnecessary waste of life … In attempting to arrive at the truth, I have applied everywhere for information, but in scarcely an instance have I been able to obtain hospital records fit for any purposes of comparison … If wisely used, these improved statistics would tell us more of the relative value of particular operations and modes of treatment than we have means of ascertaining at present."

Florence Nightingale eloquently pointed out the importance of robust documentation for the purpose of identifying evidence-based care. A similar premise applies to clinical quality measurement. Standardized and accurate information is needed to determine whether or not the process of care delivery and the actual care delivered meet minimum standards. Clinical quality measures offer a means to quantitatively assess adherence to standards and are vital to monitoring and facilitating improvement in the delivery of care. Trustworthy reports on clinical quality are, therefore, compulsory when determining the consistency of delivering "the right care, to the right patient, at the right time" (2). As part of the effort to improve the quality of clinical care, the government has mandated all providers in the Medicare program to report on clinical quality. The measurement of and reporting on quality (adherence to best practice guidelines) shows the extent to which appropriate care is delivered to patients. Upon identification, steps can then be taken to mitigate areas of inadequate compliance to improve the quality of care. In addition, governmental agencies (and providers) can use the reported information to evaluate the quality of care among providers and reimburse providers accordingly. Data are a prerequisite to computing and reporting on quality measures. To fulfill this need, a federal initiative was instituted to encourage providers to transition from paper charts to electronic health records (EHRs). One of the anticipated benefits of the shift towards EHRs was improved access to and use of clinical data to support both clinical care and quality measurement, among others. Given this positive outlook, electronic clinical quality measures were developed to take advantage of data expected to be available from EHRs. The purpose of this study is to assess the use of electronic measures to evaluate clinical quality.

2. Background

Status of the U.S. Health Care System

Today, in spite of significant advancements that have been made in information technology and medicine, the U.S. health care system continues to struggle with achieving consensus and implementing standards for the delivery of clinical quality across the health care system. As a result the U.S. health care community continues to struggle with uniformly disseminating quality care across all patients. Deficiencies in care have been highlighted in reports compiled by the Institute of medicine (IOM). The excessive number of preventable deaths was underscored in the 1999 report, *To Err is Human* (3). A subsequent report, *Crossing the Quality Chasm*, details the fragmented nature of the health care system and the inequalities in the care delivered (4). Several studies provide strong evidence of disparities in the care dispensed. A study by McGlynn et al. on adherence to standard of care practices revealed that only fifty-five percent of adult participants received the recommended treatment for acute, chronic or preventative care (5). A similar study conducted among pediatric patients indicated even lower (46.5%) observance of standard of care guidelines (6) Additional studies among gastroenterology patients produced analogous results (7) These studies point to poor compliance with standards of care or best practices. In addition, external comparisons of the U.S. health care system to other developed countries indicate that the U.S. well below the top ranks (twenty-seventh out of thirty-five countries) when considering population health statistics such as life expectancy (8) Life expectancy in the U.S. is currently more than a year below the average of all countries included in the report; it is worth noting that in 1970 life expectancy in the U.S. was one year above the average of all participating countries (9) Infant mortality is also relatively high with the U.S. ranked seventh highest overall (10) Challenges with quality of care are compounded by the high cost of care. According to the Organization for Economic Co-operation and Development, the U.S. has the highest expenditure in health care and spends almost twice the average percent of gross domestic product (GDP) among all countries (11). The Medicare Payment Advisory Committee predicted that expenditures on Medicare (primarily for the elderly) and Medicaid (for the low-income bracket) public health insurance programs and private health insurance will reach approximately twenty percent of GDP by 2020 (12). Singularly, Medicare is the largest provider of health insurance in the U.S.; the program is overseen by the federal government and accounted for $525 billion dollars (23%) of all expenditures on personal health care in 2011 (12). The high cost of care does not necessarily translate to high quality of care. The health care system, as noted above, has significant gaps in the quality of the care dispensed.

Initiative to Measure the Quality of Clinical Care

A major legislative proposal to force improvement in the health care domain was passed to address the unsustainable spending on health care and unrealized benefits in improved health of the population. One way to address the issues of high cost and varying quality was to implement legislation to help obtain better health information to support clinical care. In 2009, the U.S. government took a major step towards closing the gap in the quality of care by passing the American Recovery and Reinvestment Act (ARRA) (13). The Health Information Technology for Economic and Clinical Health (HITECH) Act was enacted as part of ARRA and puts forth a government initiative to increase the "meaningful use" of health information technology in the U.S (14). Twenty-seven billion dollars were allotted to incentivize clinicians and hospitals to adopt and use electronic health records (EHR) meaningfully, including decision support tools and participation in health information exchange (14,15). U.S. hospitals have made significant strides in the adoption of EHRs. As shown by a survey of acute care hospitals in 2012, almost 60% had either installed a basic or comprehensive EHR system, which is over a four-fold increase when compared to the number of reported installations in 2010 (16). Results from the 2012 National Ambulatory Medical Care Survey showed nineteen percent (double the percentage from 2007) adoption of EHRs that meet the majority of the Meaningful Use Stage I objectives for physician practices; however, seventy percent (a twenty percent increase from 2007) employ some type of health information technology in their practice (17). The ultimate goal of the HITECH Act was to "improve the quality, safety, and efficiency of care, while reducing disparities"(18) Attaining these goals is partly dependent upon broadening adoption and leveraging the functionalities present in EHRs. EHRs offer potential benefits: Improved legibility of records – clinical information is recorded using typed instead of handwritten text; and the use of e-prescribing and clinical decision, as advocated by the Institute of Medicine, to reduce medical errors (4). Of particular importance is the use of clinical data from EHRs to support quality measurement.

The goal of the HITECH Act was to advance the federal government's effort to improve quality of care and control health care spending. To realize this goal, the Department of Health and Human Services (HHS) partnered with the National Quality Forum (NQF) to develop a national strategy on quality(19) The NQF is an organization that sets priorities for national strategies on performance improvement and endorses standards for measuring and publicly reporting on performance, to fulfill the mandate(20). One way to manage the deficiency in quality is to monitor the quality of the care dispensed by providers using objective and quantitative methods and to hold providers accountable for the quality of the care they dispense. To this end, the Centers for Medicare and Medicaid Services (CMS) have instituted a policy requiring Medicare providers to report on clinical quality (15). The reporting requirement is central to fulfilling the mandate for a national strategy on health care quality. The federal government oversees the clinical quality reporting program, which is limited to Medicare providers because the respective states manage the Medicaid programs. The reporting mandate is part of a pay-for-performance program in which eligible Medicare providers receive financial incentives for demonstrating quality care. Beginning in 2015, financial penalties- in the form of withheld payments- will be assessed against providers who do not demonstrate delivery of quality care. The mandate to report on clinical quality leverages the infrastructure –electronic health records- put in place by the HITECH act. Access to clinical data is necessary to measure clinical quality. Data captured via EHRs was expected to support patient care and secondary uses such as clinical quality measurement.

What is a Clinical Quality Measure?

Clinical quality measures (CQM) offer a means to quantitatively evaluate clinical care and derived from best practice guidelines. CQMs are typically presented in the form of a proportion. The denominator includes all patients with a condition of interest and the numerator, a subset of the denominator, includes all patients who actually received the recommended care. For example, a quality measure that assesses the percentage of diabetic patients who have undergone a retinal exam would include all diabetic patients in the denominator and the numerator would include only those diabetic patients who had undergone an examination of their retina. Performance measures are used to evaluate quality in three main areas: structure, the physical and staffing characteristics of the health care facility; process, the means by which care is delivered; and outcomes, the result of the care delivered (21).

Format of Clinical Quality Measures

Originally, quality measures were paper-based. Meaning criteria were descriptive in nature and based on how data were recorded manually. Paper-based measures were converted to electronic measures or eMeasures subsequent to the introduction of the EHR incentive program. The National Quality Forum decided to move from paper-based to electronic measures also known as eMeasure criteria. eMeaures were developed because the paper-based measures were mostly descriptive in nature and did not provide fine detail to guide implementation of a measure. eMeasures, on the other hand, do not only provide a description but the criteria can be operationalized to retrieve the data. The electronic measure criteria are structured and use a standardized format.

Structured data consists of discrete data values such as numeric values (age), dates or standardized vocabularies. Discrete data can be counted. Standardized vocabularies include unique identification codes that represent various data concepts; each identification code is accompanied by a description of the concept it represents. Gender, for example, could be assigned the values of male and female and each would be tied to a unique identification number. In contrast to text or narrative entries (assessments and progress notes), structured data (coded values) can be easily aggregated across large patient populations to support automated quality reporting or other analytical initiatives. Text (unstructured) data, on the other hand, is not appropriate for analytics and would require manual abstraction or the use of complex techniques such as natural language processing to obtain the necessary data. Extra steps would have to be taken to obtain and transform text into structured data before a report could be developed. The existence of structured data eliminates the need to expend resources to transform the data.

Each measure definition is composed of data elements categorized as a diagnosis, procedure, or event. A data element consists of value sets drawn from various terminology standards. Examples of standardized terminologies include: International Classification of Diseases, Ninth Revision (ICD9, Systemized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) and RxNorm, used to define diagnosis, clinical concepts and medications, respectively. The eMeasure format makes use of standardized terminologies to define clinical concepts and activities (such as diagnoses and medication administration) needed to identify patient populations to compute a measure. Logic combining data elements and boolean operators are also provided to specify population criteria for the numerator and denominator of a quality measure.

Source of Data for Measure Computation

Chart Abstraction

Historically, manual abstraction has been the primary method used to obtain data from paper medical records for quality measurement. Paper charts contain hand-written notes documented by clinicians who have provided care to a patient. Illegibility of records due to poor handwriting was therefore a persistent issue with regard to paper charts. In addition, charts can become voluminous, particularly for patients with chronic conditions. Identifying specific data points can be difficult when examining voluminous paper charts. In spite of these challenges, paper records had to be abstracted to obtain data for the measurement of quality, and each patient's medical chart had to be manually searched for pertinent data. The disadvantage of manual review was it both time and labor intensive and prohibitively expensive to abstract paper records (22). Manual chart review offers the benefit of access to all clinical data. However, this method of obtaining data from patient records cannot be routinely conducted on a large-scale to support quality reporting, which requires access to data from hundreds or thousands of patient records.

Administrative Data

Administrative data are typically sourced from applications that support the operational aspects of delivering care as opposed to direct clinical care. Claims data became available when hospitals and insurers began to use computer systems for billing and other operational purposes. Billing applications were solely intended to process reimbursements for care rendered to patients. Limited information, patient identifiers, diagnosis and procedure, was required for reimbursement. All of these factors could be computerized. Patient identifiers, identification number, date of birth, gender, are inherently discrete and conform to discrete data types in electronic systems. Diagnoses and procedures were represented using codes from standardized vocabularies, International Classification of Diseases, Ninth Revision (ICD9) and Current

Procedural Terminology (CPT). These vocabularies were also compatible with computer data formats.

There are definite benefits to using administrative data: “the data are readily available, are inexpensive to acquire, are computer readable and typically encompass large populations” (22). This is particularly true of claims data. No additional costs were incurred and the information was available because billing is an intrinsic part of the health care delivery process. The data can be easily aggregated and manipulated for analytical endeavors such as quality reporting. Nonetheless there are also drawbacks to administrative data. Errors have been observed in the coding of diagnoses and health plan data were incomplete for the measurement of physician performance (23–28). Coding errors and missing data result in the misrepresentation of a patient’s clinical state. This is because a comprehensive picture of a patient's medical status may not be captured in the billing process as the claim form only supports entry of a minimum numbers of codes (22). Furthermore, administrative data focuses on aiding the billing process as opposed to presenting a complete characterization of a patient's medical conditions.

The task of coding is dependent upon the content of a patient’s record, which primarily originates from the physician. Physicians document observations made during a clinical encounter with a patient. These findings are then used to guide code selection. It is therefore unsurprising that coding errors have been linked to the quality of documentation in medical records (29). Agreement was lower between administrative data and abstracted data for records that were categorized as having sub-par documentation. Other factors, in addition to the quality of physician documentation, were determined to affect the quality of administrative data. Incorrect physician diagnosis, missing records or data entry errors negatively affected the validity of administrative data (25). Simborg has described a phenomenon, “code,” which can also lead to coding inaccuracies. Code creep is a major issue and occurs when there is a “deliberate and systematic shift in a hospital's reported case mix in order to improve reimbursement" (30). Essentially, the more complex the disease state, the higher the compensation. The resulting effect is a biased representation of clinical manifestations.

At the other end of the spectrum, coding of conditions deemed to be preventable may be minimized or underreported to avoid financial loss because reimbursement rules were modified to exclude coverage for treatment of conditions deemed to be preventable (31). Clinical activity must also be coded as part of the billing process. In contrast to diagnoses, coded procedures in administrative data demonstrated greater validity (26,32). Yet inconsistencies in procedure coding were found for discharges (27), and a closer review revealed that coding was more accurate for major surgical procedures as minor procedures tended to be underreported (26). Coding was influenced by process rules and patient characteristics. Romano highlighted the adverse effect of limiting the number of diagnosis and procedure codes that could be reported on a claim form and the underreporting of comorbidities among inpatient deaths (27).

The use of operational data to identify disease populations, chronic obstructive pulmonary disease, osteoporosis, acute myocardial infarction, osteoporosis, diabetes, and heart disease and related conditions was not optimal (33–38). In addition, detection was particularly poor when diagnosis codes alone were used; but combining factors and including additional requirements, more than one claim with the specified diagnosis or prescription data, improved case detection ((33,37). Operational data were similarly used in the quality sphere. However, results have been less than stellar for computing quality indicators in the geriatric (assessment of care for vulnerable adults) and diabetes (hemoglobin A1c and LDL testing and retinopathy screening) care and surgical and medical complications (39–43). One study did showed coding of administrative data was accurate in aggregate but the level of accuracy differed by disease (32).

The benefits, access to structured data at a low cost, of administrative data was indisputable. The ability to extract data electronically and on a large scale was an improvement over manual chart review. Yet, there were major drawbacks that cannot be ignored: inaccurate coding and a limited data pool, particularly absent of clinical details. Assessment of various aspects of clinical care required access to clinical content that was generally absent from billing data. Few measures evaluating processes and outcomes at the core of clinical quality can be supported using administrative data. In essence administrative data reflected how we paid for care and was not designed to capture clinical findings and care delivered to a patient. Because of the benefits described, administrative data have been used support quality reporting but may not be a good proxy for clinical data. On the other hand, electronic health records were seen as a viable source of clinical data.

Electronic health records (EHR)

Coupled with the development of eMeasures, EHRs offered a new paradigm for quality reporting. In all, EHRs would embody both the analytical capabilities of administrative data and contain rich clinical content. EHR data is comprised of structured and unstructured formats. The underlying assumptions were EHR data would be structured, scalable and available at a level of granularity that would support reporting. In some instances, billing data might be captured in an application separate from the clinical record system or as part of a suite of applications that includes administrative and clinical modules. Data from clinical modules can be accessed either via a user interface (can only view one record at a time) or a data warehouse (can access data across numerous patients). A clinical data warehouse contains a copy of all or a subset of the data recorded in an EHR system. Data warehouses are used for analytical tasks (e.g. quality reporting) because the chief purpose of EHR systems is to support individual transactions and not data analytics across hundreds or thousands of patients.

A number of studies have examined the use of EHR data in the quality domain. All of these studies either preceded the meaningful use directive or did not involve NQF meaningful use measures. Six organizations with an EHR system, installed prior to the national race to expand the use of EHRs, participated in a study to assess the feasibility of implementing a set of developmental screening measures (44). The aim of the study was to determine the availability of electronic data to implement the measures. None of the measures were actually computed during the study. Potential data sources were identified for two of the three measures across the six organizations. The fact that characteristics of the data and sources identified from the differing EHR systems were heterogeneous was striking. EHR systems diverged in terms of what, how and where data for specific data components were recorded. For example, performance on a screening test was determined via billing codes (most common source), free-text data field containing the score and interpretation, or an amalgamation of free-text fields, scanned documents, and problem list. These are prime examples of the varying forms of electronic data and the challenges of accessibility to structured clinical data. Hazlehurst et al. demonstrated that a homogenous clinical process and, by extension, data collection resulted in better performance on quality indicators (45). Furthermore, EHR data proved to be problematic for a broad spectrum of measures (cancer screening, BMI, blood pressure), resulting in underestimation of performance (26).

Disease registries are another source of electronic clinical data. Registries are a unique data source because the content is focused a particular condition and deliberate effort and stringent is put selecting, defining and collecting data on elements most relevant to the disease of interest. Conversely, EHR data covers a broad range of conditions and do not have to abide by stringent guidelines typical of registries. Data from a registry enabled the successful implementation of 19 out of 22 prostate cancer measures; three of which were sourced exclusively from the registry (46). Measures computed using registry data also had high concordance with chart review data

(47). Barkhuysen brought the advantage of registry data over EHR data to light in an assessment of the two systems within a single primary care group (48). Physicians were routinely required to populate a registry in addition to documenting in an EHR system. With the registry as the reference standard, analysis showed significantly higher agreement for outcome measures than process measures (48). Suggesting that documentation in the EHR was comparable to the registry for some measure elements but not others and that it is possible to capture relevant information for quality reporting.

There is also compelling evidence for the use of EHR data. EHR data had a strong performance against chart review in determining eligibility for cervical cancer screening (49) and measuring administration of standard treatment for myocardial infarction (50). Ninety-nine percent of 100 women randomly selected for manual review were accurately classified for cervical cancer screening and sensitivity ranged from 83 – 100% in the categorization of treatment for myocardial infarction. Other evaluations of EHR data were promising: slightly more than 80% of the data (compared to 95% using chart review) could be sourced directly from the EHR for three diabetes indicators and results did not deviate significantly from chart review (51). The performance of automate computation of heart failure measures using EHR data was promising, with the exception of one measure which was negatively impacted by the lack of access to exclusionary data (52). It should be noted that chart review was not comprehensive and was only conducted if a case failed to pass the measure under the automated method. Lastly, an organization cut the amount of resources and time spent manually reviewing charts in half (53).

Coupled with the less than positive outcome reported from some studies, there are evaluations that have indicated that EHR data can potential advance the measure computation. Many of the studies stated the use of electronic medical record or electronic health record data. But the source of the electronic data differed across some of the studies. Some studies referenced data from clinical modules, while others included pharmacy, billing, laboratory, and registry data under the umbrella of EMR or EHR data.

Evaluation of NQF eMeasures

The EHR incentive program was viewed as a means to promote the adoption of EHRs and subsequently improve access to clinical data for quality reporting. The clinical quality reporting program consists of 113 CQMs that have been endorsed by the NQF (54). The reporting program includes both process and outcome measures; however, a majority of the measures fall into the process category. Hospitals must report on at least 16 of 29 measures and eligible professionals are required to select a minimum of 9 measures from a pool of 64 (54). eMeasures were developed to take advantage of the structured data anticipated to be captured in EHRs.

The use of eMeasures to measure clinical quality is nascent, and few studies have been conducted to assess the use of eMeasures and clinical data from the EHR to support quality measurement. An evaluation of a subset of the paper-based measures identified inaccuracies between electronic reporting (structured data) and manual chart review (55) The above study is limited because the evaluation was conducted using paper-based or descriptive measures. eMeasures had yet to be developed. Researchers have attempted to assess the level of complexity and difficulty with implementing eMeasures by enumerating the terminologies and codes used to define individual eMeasures (56). A survey of quality experts was conducted to solicit feedback on whether or not eMeasures could be calculated without difficulty. The results of the study indicate that half of the respondents categorized execution of the eMeasure criteria as moderately difficult. The evaluation does shed light on the perceived challenge of executing eMeasure criteria but some of the respondents had not implemented the measures under evaluation. Thus, only a subset of the responses to the survey was based upon observations stemming from the experience of implementing eMeasures. Another study on eMeasures involved assessment of the currency and accuracy of the standardized value sets used to identify clinical concepts for the respective measures (57). Each value set consists of a pre-defined list of codes used to identify clinical concepts; for example, a range of ICD9 codes for the identification of diagnoses related to diabetes. The study findings show that 79% of measures had at least one value set with errors. The studies discussed above provide valuable insight on eMeasures; however, the fundamental question regarding the accuracy of eMeasures remains unanswered. The use of eMeasures in a specialty domain is of particular interest. To date, an in-depth analysis of the use of eMeasures to assess adherence to ophthalmology best practice guidelines has not been conducted.

Statement of the Research Problem

It is important to note that although measure definitions are standardized, the documentation of clinical data captured during the process of patient care may or may not reflect the standards used to define the measure criteria. Furthermore, the granularity of the data components in the measure criteria may also differ from data recorded in EHR systems. While some elements in the EHR are recorded in a structured manner that allows for retrieval, the granularity needed for quality measurement may be absent without deliberate effort to capture and extract the data (58,59). Subsequently, the translation of the NQF criteria based on the configuration of the data captured within an institution's EHR system may be necessary to implement an eMeasure. This is because the standardized criteria cannot be applied directly against existing EHRs systems and requires transformation for execution of the requirements. A limited number of the NQF eMeasures have been translated and criteria have been compiled based on the data captured in a commercial EHR system. The drawback is that the accuracy of measuring clinical quality using the interpreted criteria has yet to be determined. Currently, there are no known studies that have examined how well the interpreted criteria measure clinical quality. Ascertaining how well quality is measured is necessary because, as pointed out by Florence Nightingale, reliable information is needed to support analysis and decision-making.

The ultimate goal of the clinical quality reporting program is to facilitate improvement in clinical quality by monitoring patient care and mitigating identified deficiencies. Action cannot be taken to improve clinical quality without the use of accurate information to inform decision-making, hence the need to ascertain the accuracy of measuring quality using the interpreted criteria. The assumption that the necessary clinical data would be readily available for measure implementation if an EHR system were in use requires further exploration. But access to data may not be the only issue at hand as the manner in which data are recorded in the EHR directly affects the appropriateness of the data available for quality measurement. The data recorded in the EHR is influenced by the documentation practices of clinicians, which could vary widely. Studies have shown that standardization and documentation play an important role in data availability, reliability and accuracy (26,60).

Clinical templates, for example, contain structured components to facilitate the capture of discrete data. The study focused on assessing the use of data recorded in the clinical template during the care process. None of the data were sourced from administrative systems or other sources outside of the template used for clinic visits. The consistency with which clinicians employ structured elements in templates could influence the quality of the data available for automated measurement. The perceptions and behavior of clinicians with respect to clinical documentation and the use of electronic data for quality measurement is consequently of great significance. Exploring how and why clinicians document clinical information, particularly those relevant to quality measures, would be of benefit. The primary goal of the study was to assess the accuracy of measure guidelines using EHR data and to compute quality measures using an EHR-based algorithm and the NQF guidelines. These concerns were translated into three aims.

Aims

Our primary goal was to assess the availability and accuracy of using EHR data to compute quality measures using an EHR-based algorithm and the NQF guidelines. This goal would be achieved by the following aims: The first two aims involve comparison to a gold standard.

I. Evaluate the use of interpreted NQF measure criteria for the computation of clinical quality measures.
II. Assess the implementation of NQF eMeasure guidelines using data captured in an electronic health record system. Implementation involved using data to classify patients as "met" or "did not meet" according to the rules outlined in the guidelines.
III. Assess clinician views and reported documentation practices in relation to clinical quality measurement.

Study Factors

The study was conducted at a general ophthalmology clinic, which offers a broad spectrum of services, including medical to surgical care. The eye center records clinical information acquired during the process of patient care via a commercial electronic health record system. Information recorded via the user interface was captured in a database. The user interface or clinical template consisted of individual data elements for recording information for specific clinical components and a section for summarizing findings. The fundus section contained individual data elements for the eye exam components that were pertinent to the study: macula, disc, vessels, periphery and cup to disc ratio. The assessment and plan section contained a text narrative of the ophthalmologist's diagnosis, treatment plan and other relevant impressions for an encounter. Specific data elements were also present within the clinical template for a coded diagnosis and the level of service.

The study was based on the assessment of two ophthalmology quality measures endorsed by the National Quality Forum (NQF). The measures were:

NQF eMeasures

I. Diabetic retinopathy: Percentage of patients aged 18 years and older with a diagnosis of diabetic retinopathy who had a dilated macula or fundus exam performed which included documentation of the level of severity of retinopathy and the presence or absence of macula edema during one or more office visits within 12 months.

II. Primary open-angle glaucoma (POAG): Percentage of patients aged 18 years and older with a diagnosis of primary open-angle glaucoma (POAG) who have an optic nerve head evaluation during one or more office visits within 12 months.

The goal was to evaluate the availability of data to support quality reporting in the ophthalmology domain and to estimate the performance of the interpreted NQF guidelines, termed EHR algorithm. The NQF guidelines were also implemented and comparisons were made between the EHR algorithm and the NQF guidelines. In addition, a reference standard was developed as a benchmark against which both guidelines were compared. Quality measures are typically summarized to a proportion. For this study, analysis focused on comparing the classification of patients under the three methods EHR algorithm, NQF guidelines, and reference standard. Shifting the analysis from comparing proportions to the classification of patients was essential to examining and understanding the impact of the underlying EHR data on quality reporting. The study was conducted at the chosen site because the clinic was in the medical domain of interest; an electronic health record system had been in use for several years; and the researcher had access to medical records via the user interface and existing data captured in the EHR database. The study used a mixed-methods approach, combining quantitative and qualitative methodologies.

3. Methods

Design

This is a cross-sectional and mixed methods study. Clinical quality measures were implemented using two sets of measure criteria for the quantitative analyses. The DR and POAG clinical quality measures were implemented separately using the EHR algorithm and NQF guideline. Patients were classified to determine eligibility (denominator) and success in passing the measure (numerator) for each condition and guideline. Classifications were compared between the two methods of implementation and a reference standard for each condition. The reference standard was developed at the study site and accessed all data including clinical notes that could not be used to implement the guidelines. Qualitative evaluation involved interviewing physicians at the study site and analyzing their responses.

Setting and Study population

The study was conducted at an outpatient general eye care clinic staffed by attending physicians. The clinic provided comprehensive medical and surgical ophthalmology services. A commercial EHR system was in use at the time of the study and had been in use for seven years. The study population included patients with a diagnosis of DR or POAG who were seen at an outpatient eye center between January 1, 2014 and December 31, 2014. Diagnosis criteria from the EHR algorithm and NQF guideline were applied separately to identify patients with diagnosis of DR or POAG. The study period mirrors the annual reporting timeframe outlined in the NQF measure guidelines. The year 2014 was chosen because it was the most recent and complete calendar year for which data were available. The study population was defined as follows:

- 18 years or older on the start date of the study period
- Outpatients seen at a general eye care clinic during the study period
- Diagnosis of DR or POAG per NQF guideline or EHR algorithm

Sample Size

All patients classified with a diagnosis of DR or POAG under the NQF guideline or EHR algorithm were included in the study. A sample of the patients who were classified as not having a diagnosis of DR or POAG under the NQF guideline or EHR algorithm was analyzed to determine whether diagnosis status had been misclassified.

Data Source

The study was conducted using data from a commercial EHR system. Ophthalmologists and other clinicians entered data (exam observations and findings) into the EHR system via a clinical template (customized graphic user interface). The data entered into the EHR system were stored in a transactional database. Data and medical records assessed for the study were limited to those recorded at the study site during the study period. Data and medical records outside of the study period or the site clinic were not evaluated. Data from the data warehouse were used to implement the measure guidelines and data abstracted via human review were used for the reference standards. Data were accessed in two ways:

1) data warehouse containing data exported from the EHR system
2) human review of individual records via the EHR system's user interface

Data Manipulation

Excel files containing data extracted from the EHR system for measure implementations were uploaded to a MySQL database. Structured Query Language (SQL) was used to construct queries to apply the measure guidelines against the study data. Sourced data were used to determine whether or not an encounter had the appropriate diagnosis, was eligible for the denominator, and met the numerator criteria under the EHR algorithms and NQF guidelines.

Statistical Methods

Quantitative Analysis

The kappa statistic was used to compare patient classifications under the NQF guideline and EHR algorithm for DR and POAG. Kappa assesses the level of agreement beyond chance between two entities. In this case, we compared denominator and numerator classifications for DR and POAG patients between two implementation methods. The value of kappa ranges from -1 to 1. A value of -1 indicates that agreement is less than chance and a value of 0 indicates no agreement. The ultimate target is 1, which indicates a 100% agreement. The percent agreement, proportion of patients with matching classifications between the two methods, was also computed. The level of agreement should positively correlate with the kappa statistic: the higher the percent agreement, the higher the kappa and vice versa. Due to low prevalence of particular outcome, a paradoxical relationship might be observed between percent agreement and kappa (61). That is, high percent agreement but low kappa. Prevalence-adjusted bias-adjusted kappa (PABAK) is therefore used in such circumstances to correct for the effect of prevalence on kappa (62).

Qualitative Analysis

Participant responses were annotated and analyzed using the template method. The template method offers a narrow and proficient means of evaluating text though the use of a coding manual (63). The codes are derived from pre-existing knowledge of the researcher and responses from interview participants. And the index of codes can be modified and refined during the review process.

Template analysis was used to identify relevant themes and code in the responses to each question. The template method was chosen over a grounded theory approach (63) because the study questions were targeted and centered on specific topics. Though the questions were open-ended, the questions were narrow in focus and the range of possible responses to the questions was limited to some extent. A grounded theory approach would be better suited to interview questions that garner a broad range of responses in which respondents were not inherently limited in response due to the very nature of the question. The format and scope of the questions inherently leaned toward a more focused method of analysis such as the template method.

Quality Control

The researcher and an expert in field of quality reporting reviewed the implementation guidelines. These guidelines were then translated to SQL queries. The components of each query and the results were reviewed multiple times to identify and address any anomalies. The review process ended when no new errors were identified. A pilot study was conducted to evaluate the proficiency of the abstractor to conduct manual chart reviews under the reference standard. A physician expert reviewed the results. Data abstracted under the reference standard were reviewed in entirety for a patient of cases and compared to the source record in the EHR system. Additional reviews were conducted each time there was any suggestion of an inconsistency in the abstracted data.

Human Subjects

The Institutional Review Board at the study site approved the project. Individual consents were not required for the data used in the quantitative analysis. Verbal consent was required and sought from each interview participant.

Study Details

Quantitative Analysis

Measure Implementations

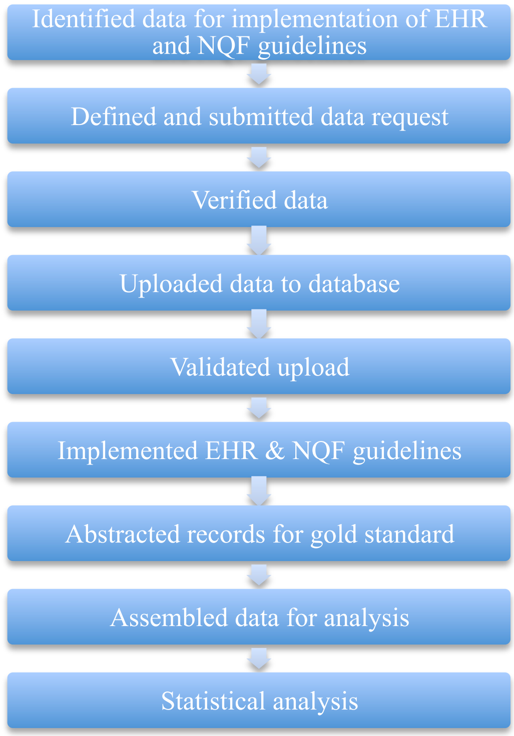
In brief, the EHR and NQF guidelines for the diabetic retinopathy (DR) and primary open-angle glaucoma (POAG) clinical quality measures (CQM) were applied against data on outpatient encounters. The data used for this project were extracted from a commercial EHR. Four sets of implementations (See Table 1) were carried out using data extracted from the EHR. The criteria for each measure were implemented strictly as stated with minimal interpretation.

Table 1. Clinical quality measure guidelines implemented

CQM/Implementation Method	Clinical quality measure	Implementation method	Classifications
DR EHR algorithm	Diabetic retinopathy	EHR algorithm	Denominator Numerator
DR NQF guideline		NQF guideline	
POAG EHR algorithm	Primary open-angle glaucoma	EHR algorithm	
POAG NQF guideline		NQF guideline	

EHR algorithm and NQF guidelines were applied to identify patients with a diagnosis of DR or POAG. Encounters (outpatient appointments) were classified to indicate whether or not each met the denominator and the numerator criteria separately for the four sets of implementations. Next the classification of encounters according to each denominator and numerator criteria was summarized at the patient level. The process of classifying encounters and patients for the denominator and numerator was repeated using the reference standard. Finally, the summarized denominator and numerator classifications for each of the four implementations were compared to the reference standard. An overview of the steps completed for the quantitative analysis is presented immediately below. The details of each step are described in subsequent sections.

Figure 1. Overview of implementation procedures



Data Acquisition

Data for the quantitative analysis were obtained from the data warehouse. The data warehouse is a centralized repository of data from the EHR system. Data in a data warehouse are organized into tables, which are made up of columns (define the data) and rows (contain records). A data request was created and submitted to the information technology (IT) group for extraction of relevant data from the data warehouse. Data extracted from the data warehouse were used to implement the EHR and NQF guidelines for the DR and POAG CQMs. Data corresponding to components in the EHR algorithm (64) and NQF guideline (65,66) for the DR and POAG CQMs were identified from the data warehouse. The following tasks were completed to construct the data request:

1. Identify pertinent data from EHR & NQF guidelines
2. Identify corresponding data in data warehouse
3. Identify ancillary data to lend context

Data Parameters for EHR Algorithms

The parameters of the study data and the location of pertinent data within the data warehouse were determined to create a data request. A metadata dictionary (contains descriptions for tables and columns in the data warehouse) was used to determine the location of pertinent data in the data warehouse. The EHR algorithms for the CQMs were established interpretations of the NQF guidelines and tailored to the general configuration of data in the EHR system. Each EHR algorithm contained thirty-eight identifiers representing data concepts in the parallel NQF guideline. The data concepts were

- Birthdate
- Diagnoses
- Procedures
- Eye exam components
- Provider information
- Procedure code lists
- Diagnosis code lists

The identifiers in the DR and POAG EHR algorithms were identical, with the exception of algorithm specific components. The components specific to the DR EHR algorithm were documentation of macula edema and severity of retinopathy; the POAG specific component was optic nerve head (same as optic disc) evaluation. From the outset, it was not easy to find the location of data in the data warehouse. Ascertaining the link between identifiers and actual data in the data warehouse and acquiring the necessary information to transition from identifiers to data columns was arduous. It turned out that the identifiers in the EHR algorithm served as pointers to specific data columns in the data warehouse. The next test was obtaining information on the columns and tables connected to each identifier. An attempt was made to look up each identifier to find related columns via the user interface of the metadata dictionary. But this approach was inefficient and discarded because it was discovered that many of the identifiers were associated with more than one data column in the data warehouse. To speed up progress, a request was made for the IT group to query the metadata dictionary and provide the list of tables and columns associated with each identifier. Information was retrieved on 36 of the identifiers. No information could be found on two identifiers associated with eye exam components but information was available on two other identifiers related to eye exam components. One thousand three hundred and seven data columns were identified among 36 identifiers. A high number of columns were returned because many identifiers were linked to more than one data column in the data warehouse. Some identifiers had multiple matches because a named column was present in more than one table and/or supplementary columns had been included. Here are examples of both scenarios (Table 2):

Table 2. Example of identifiers with multiple matching columns

Example 1: Same column multiple tables Identifier: *Diagnosis*	Example 2: Supplementary column Identifier: *Diagnosis*
ENCOUNTER_DIAGNOSIS table: DIAGNOSIS_ID LAB_DIAGNOSIS table: DIAGNOSIS_ID	Primary column: DIAGNOSIS_ID (contains a numeric identification number for each diagnosis code)
ALL_DIAGNOSIS table: DIAGNOSIS_ID	Supplementary column: DIAGNOSIS_ICD9 (contains the actual ICD9 code corresponding to the numeric value in DIAGNOSIS_ID)

If an identifier was linked to only one data column in the data warehouse then that specific data column was selected as the data source for the identifier. On the other hand, for identically named columns in multiple tables, the column and table metadata were reviewed and the column (and table) with a description most relevant to the identifier was chosen. For example: The identifier named *diagnosis* in the EHR algorithm represented the *diagnosis of diabetic retinopathy* criteria for the DR measure. The *diagnosis* identifier in the EHR algorithm pointed to multiple data columns across various tables in the data warehouse. Each potential data column contained diagnosis data in the form of an International Classification of Diseases, Ninth Revision (ICD9) code and were in disparate tables such as ENCOUNTER_DIAGNOSIS and LAB_DIAGNOSES. In this setting, the diagnosis column in the ENCOUNTER_DIAGNOSIS table was selected over the LAB_DIAGNOSIS table because the study was centered on outpatient encounters. Supplementary columns were selected as appropriate. It should be noted that the column selection process was challenging due to the lack of access to data in the data warehouse. The ability to view data in the data warehouse would have made the selection process much easier by providing definitive knowledge on the content of tables and columns in the data warehouse. In the end, as described above, the metadata served as the primary source of guidance in the column selection process.

The DR and POAG EHR algorithms had 38 identifiers. No information was available on two of them. Three identifiers were excluded because they were related to inpatient data and the study was conducted in the outpatient setting. A total of 43 data columns were selected for the remaining 33 identifiers. The identifiers represented included: Birthdate, visit diagnosis, billing information, encounter information, eye exam components, procedure codes, and provider information.

Data Parameters for NQF Guidelines

The NQF guideline for the DR and POAG CQMs were composed of data concepts, logic to guide patient selection for the denominator and numerator, and exclusion criteria. Both CQMs share the same data concepts except for measure specific eye exam components. These included macula edema findings present, macula edema findings absent, and level of severity of retinopathy findings for the DR CQM and optic disc exam and cup to disc ratio for the POAG CQM. The ophthalmology clinical template (EHR user interface) was reviewed to identify discrete data elements from which to source data for concepts outlined in the NQF guidelines. Data concepts in the NQF guidelines were matched to the following data elements in the ophthalmology clinical template as shown in Table 3. Data elements analogous to the NQF data concepts birthdate, diagnosis, encounter type, macula, cup to disc, and disc were identified from the clinical template.

Table 3. Matching NQF components to data elements in clinical template

CQM	NQF guideline data concept	Clinical template data element
DR and POAG	Birthdate Diagnosis Encounter type Medical or patient reason for not performing exam	DOB (Date of birth) Encounter diagnosis Level of service No data element found in clinical template
DR	Macula exam Level of severity of retinopathy findings Macula edema findings absent Macula edema findings present	Macula (Right and left) Encounter diagnosis Macula (Right and left) Macula (Right and left)
POAG	Cup to disc ratio Optic disc exam	Cup to disc (Right and left) Disc (Right and left)

None of the discrete data entry components in the clinical template contained information on *level of severity* for the DR measure. However, the *encounter diagnosis* data element was used to determine *level of severity*. This was done because the data element carried ICD9 codes and a subset of the ICD9 codes for DR designates the level of severity. In other words, an encounter was ascribed a level of severity based on its assigned DR ICD9 code. The DR ICD9 codes are listed below and the codes with a level of severity are marked with an asterisk.

362.01 - Background diabetic retinopathy
362.02- Proliferative diabetic retinopathy*
362.03 - Nonproliferative diabetic retinopathy NOS
362.04 - Mild nonproliferative diabetic retinopathy*
362.05 - Moderate nonproliferative diabetic retinopathy*
362.06 - Severe nonproliferative diabetic retinopathy*

Discrete data components and surrogates were not found in the clinical template for *patient or medical exceptions*. The data warehouse column and table associated with each data element identified from the clinical template were determined using the metadata dictionary as well as the database information menu tied to some of the data elements. A list of the data columns identified for the NQF guidelines was compiled.

Data Request Parameters

To obtain the data needed to implement the guidelines, a data request was created and submitted to a data warehouse analyst. The list of data columns identified for the implementation of the EHR algorithms and NQF guidelines were combined and used to construct the data request. The metadata of each table associated with a column on the combined list of data columns was examined to locate additional data columns that would lend context to the columns that had been selected. Other data tables containing related data were also explored. Dates, location, provider, descriptions (for columns that only contained identification codes; ICD9 codes, for example) and other pertinent data points were identified to add meaning to the selected data columns. Detailed documents were generated to outline the criteria for the patients and encounters to be included in the data request and the data columns to be extracted.

The study population and qualifying encounters were defined as outlined:

Date range (for encounter visit dates):	1/1/2014 to 12/31/2014
Age:	18 years or older on start date of study period
Encounter type:	Outpatient
Location:	General eye care department
Encounter status:	Kept or arrived

Instructions were provided on how to calculate the integer value of age to prevent inflation of the calculated age and subsequent inclusion of patients who did not meet the minimum age requirement of eighteen years. Age was initially calculated in days by subtracting the date of birth from the study start date (1/1/2014). The resulting value was then converted to units of years by dividing by 365.25 (to account for the leap year). Finally, the floor function was applied to the decimal value of age in years to obtain the largest integer for the decimal. Using this methodology ensured that patients who where at least 17.5 but less than 18 years old were not rounded up to 18 years and inadvertently included in the study. One example is a patient who is 17.50 years would typically be rounded up to 18 but the floor function would reflect an age of 17 years.

The data request was sub-divided into datasets by topic including encounter details, eye exam components, diagnoses, orders, and charges. The data columns to be included in each dataset were listed in individual Excel worksheets. The data request was divided into fifteen sets of data by topic. Two of the fourteen datasets comprised of the unique patients and encounters that met the criteria for the study population. The remaining datasets contained data from the selected data columns for the patients and encounters that met the criteria for the study population. The data request was submitted to and completed by the data team at the study site. Consultations were held with the data analyst to clarify and address questions about the data request. The data request was fulfilled by extracting data for the specified columns using source tables in the data warehouse. Each dataset was provided in a comma separated values Excel file and accompanied by a PDF file containing a data dictionary.

Data Verification

The datasets were reviewed to ensure that the data retrieved met the criteria outlined in the data request and to ascertain the internal validity of the data. The completed data request contained 14 datasets with a total of 236 data columns. The number of data columns per dataset ranged from 5 to 44; each dataset had an average of 17 columns. Many of the datasets were quite large and averaged 90,353 rows with a range of 1,412 to 202,861. In all, there were 1,264951 rows of data across all 14 datasets. The following data checks were conducted to assess adherence to the criteria of the data request: Re-calculated age of the patients using floor function to verify age on start date of study period; verified encounters selected had a value of "Kept/Arrived" (indicates patient was actually seen by a provider) for visit status; verified encounters had correct department identification number of the study clinic; compared the list of data columns in each dataset to the list of data columns in the request documents; and compared the list of patients and encounters eligible for the study to the remaining datasets to ensure that the respective datasets only contained patients and encounters that met the criteria of the data request.

The following data checks were conducted on each column in each dataset to evaluate internal validity of the data: Checked for missing values; reviewed values in each column to make sure values were appropriate and fell within a reasonable range; checked dependencies between columns (e.g., Row with ICD9 code in *diagnosis* column had a description in the *diagnosis description* column); and manually compared data between data columns and actual medical records via the user interface of the EHR system. Cases were randomly selected for the manual comparisons. Manual comparisons were conducted to review encounter details, eye exam entries, and diagnoses. All discrepancies (e.g., Omitted columns and missing data) were forwarded to the data analyst and a complete review of the data was conducted each time the data was updated. This process was repeated three times until all discrepancies between the data requested and the data supplied were addressed and errors in the extracted data were addressed.

Study Database Assembly

The study data consisted of tens of thousands of rows comprising over a hundred columns. Due to the large volume of data, a database management system was needed to store, query and manipulate the data to create aggregated datasets for statistical analysis. The MySQL database management system was chosen to support the study because it had the features to support the database needs of the study, it was available at no cost, and the information technology support team was familiar with the software. The finalized study data were submitted to an information technology support team for upload into a MySQL database. Each dataset was uploaded into a separate table within the database. A document detailing the name of each table in the dataset and the name, data type and size of each column within each table was created.

Additional information was provided on specialized formatting for data columns (e.g. medical record number had leading zeros) and on which columns had rows with empty cells so that null values could be assigned appropriately. The instructions for uploading the data to a database were submitted to another analyst in a different section of the IT department.

The process of uploading the datasets into tables was iterative. The uploaded tables were reviewed to verify that the data in the tables matched the data in Excel file from which it was sourced. The following data checks were conducted for each table: Columns in table matched columns in source Excel file, row count in table matched Excel file, values in each column in table matched corresponding column in Excel file – assessed for incorrect, truncated and missing values. Several rounds of data reviews were conducted for each table to verify that the data in the uploaded tables matched the data in the finalized Excel data files. Reports generated during each round of reviews were submitted to the analyst so that discrepancies in the uploaded tables could be addressed. Preparation of data for analysis commenced once the data in the uploaded tables had been validated.

DR and POAG EHR Algorithm Implementations

The EHR algorithms for the DR and POAG CQMs were implemented independently using the data uploaded to the MySQL database. Structured Query Language (SQL) was used to manipulate the data and implement the criteria in each algorithm. Of note, inconsistencies were observed between available data and the guidelines; steps were taken to address each incident. An indicator variable was created for each component in the guidelines to delineate whether (value of "1" assigned") or not (value of "0" assigned") an encounter met the requirements of a component. The individual components were then combined to determine whether or not an encounter met all of the denominator and numerator criteria separately. Encounters were initially assessed to determine whether or not they had the appropriate diagnosis for each CQM and then evaluated to determine whether or not they met eligibility for the denominator. Encounters eligible for the CQM denominators were evaluated to determine whether or not they met or did not meet the numerator criteria. The EHR algorithm for the DR and POAG CQMs were adapted from guidelines (64) put forward by the EHR vendor. The criteria implemented have been summarized.

DR EHR algorithm

The DR EHR algorithm criteria are delineated below.

Denominator

Patients seen in the study clinic during 2014 who were 18 years or older on the start date (1/1/2014) of the study period
AND
Encounters with diagnosis on list of codes provided
AND
Patients with two or more encounters
AND
Encounters associated with a faculty provider
AND
(Encounters with a service procedure code on list of codes provided
OR
Encounters with diagnosis on list of codes provided)

Numerator (Applied against encounters qualified denominator)

Documentation of macula edema
AND
Documentation of severity of retinopathy

DR EHR Algorithm Denominator

As described in the data request section, the study data were restricted to patients who were seen at the study site and were18 years or older on the start date of the study period. A value of "1" (Yes) or "0" (No) was assigned to each encounter to mark performance on the remaining conditions in the algorithm. The DR EHR algorithm was accompanied by a list of ICD9 diagnosis codes. The code list was stored in an Excel file. The list of diagnosis codes was used to identify encounters with or without a DR diagnosis according to the DR EHR algorithm. The list included non-DR ICD9 codes (e.g., codes for diabetes mellitus (250.XX) and profound vision impairment (369.XX). One of the ICD9 codes, 362.00, which fell within the retinal disorder code range, appeared to be an incorrect code because it could not be found in the online ICD9 manual. There were 75 unique ICD9 codes across 1,436 unique rows of singular or grouped codes in the file. Table 4 contains a sample of the DR diagnosis code sets.

Table 4. Examples DR EHR algorithm diagnosis value sets

Singular ICD9 value set	Grouped ICD9 value sets
250.50	249.50, 362.01
250.51	249.50, 365.44
250.52	250.50, 369.17
362.00	250.50, 362.04
362.01	250.50, 369.17
362.02	249.50, 362.05, 369.90
362.03	250.52, 362.07, 379.23
362.04	250.50, 361.81, 362.07, 362.01
362.05	250.53, 362.07, 362.05, 379.23
362.06	249.50, 362.07, 369.21, 362.05, 369.05
362.07	250.50, 362.07, 362.05,369.21,369.05

The number of codes in each grouping ranged from 2 to 5. There were 11 singular codes and 1,425 code groupings. Less than 1 percent of the DR diagnosis code sets contained singular codes; the majority (92%) contained groupings of either three or four ICD9 codes (Table 5).

Table 5. Count of DR diagnosis code sets by number of ICD9 codes

Number of ICD9 codes per code set (n = 1436)	Count of code sets, No. (%)
One	11 (0.8)
Two	104 (7.2)
Three	673 (46.9)
Four	643 (44.8)
Five	5 (0.3)

The format of the ICD9 codes in the code list was compared to the format of the study data because the database software discriminates between decimals with and without the trailing zero (e.g., 111.1 versus 111.10). All of the codes on the code list were formatted with two decimal places. But 5 codes (ending in zero) were changed to 1 decimal place to match the study data. This change was made because the original format of the codes was not known and might have changed when the file was passed from one individual to another. The diagnosis code criterion was applied by comparing each row of ICD9 code(s) on the code list to the diagnosis code(s) assigned to each encounter in the study data. Encounters had to carry all of the diagnosis code(s) present in a specific row on the code list to meet the DR diagnosis requirement under the EHR algorithm. Observe that the diagnosis requirement was listed twice in algorithm: optional requirement in conjunction with the service procedure code and stand alone requirement. Only one indicator variable was created for the diagnosis requirement but the requirement was applied twice as per the algorithm to determine eligibility for the denominator.

The criterion for two or more encounters during the study period was evaluated per patient. And the outcome, classification as “Yes” or “No”, for the condition was applied to all encounters associated with a particular patient. The number of encounters per patient was determined by counting the number of unique encounters per patient during the study period. Only one encounter could be counted per date for patients with more than one encounter on the same date. The remaining criteria were evaluated and applied against individual encounters. Faculty provider was based on whether or not the billing provider on an encounter was assigned the role of faculty as opposed to resident.

The service procedure condition employed a list of codes that covered a variety of services; these services included new patient, consult, and established appointments at different levels of complexity. There were 157 unique procedure codes among 159 records on the procedure list. Two codes on the list were duplicated and the descriptions differed within both pairs of duplicates. A misspelling, absence of “E” in “level”, was the culprit for the discrepancy in the descriptions between one pair of duplicated codes. The descriptions for the second pair of duplicated codes differed because the purpose appears to be different between the two; one description referred to laser treatment and the second referred to evaluation of the cornea. All of the codes on the list were formatted with one decimal place (e.g. 11111.1), which matched the format of the data used to implement the study. The unique list of 157 codes was used to execute the service procedure condition.

Some of the conditions in the EHR algorithm were not executed. The order procedure requirement was not implemented because no information could be found on which codes should be used to execute the requirement. Medical or patient exceptions for not performing an exam were not executed because no discrete data columns were found for this information. The criterion for inpatients was not implemented because the focus of the study is to assess implementation of the measures in an outpatient setting and inclusion of inpatient information would confound the results.

DR EHR Algorithm Numerator

Documentation of macula edema and level of severity were the two requirements for the DR EHR algorithm numerator. For an unknown reason the macula and level of severity requirements were listed twice in the EHR algorithm: Data columns were identified from the data warehouse for macula (R and L) documentation. The data columns contained text entries describing clinical findings for the macula of each eye. But details were not provided on specific values that should be considered as documentation of macula edema; the algorithm only indicated that the data columns should be present. Consequently, the requirement was executed by determining whether or not a record that was not null (empty) existed for both *right macula* and l*eft macula* for each encounter. In the case of the level of severity condition, no data columns were found in the data warehouse for this entity. The encounter diagnosis was the only data column that could serve as a surrogate because some of the DR ICD9 codes included the level of severity. The requirement for documentation of level of severity was executed by determining whether or not an encounter was assigned an ICD9 code for diabetic retinopathy that specified the level of severity.

The ICD9 codes for diabetic retinopathy with specified level of severity are:

362.02 – Proliferative diabetic retinopathy
363.04 –Mild nonproliferative diabetic retinopathy
365.05-Moderate nonproliferative diabetic retinopathy
365.06-Severe nonproliferative diabetic retinopathy

Implementation of the DR EHR algorithm is summarized in Table 6.

Table 6. DR EHR algorithm steps implemented

EHR algorithm rules	Data columns from data warehouse
Denominator	
18 years and older	BIRTHDATE
DR diagnosis	ENCOUNTER_DIAGNOSIS
More than 1 encounter per patient	ENCOUNTER_ID, ENCOUNTER_DATE
Faculty provider	PROVIDER_ROLE
Service procedure code	LEVEL_SVC_CODE
Numerator	
Documentation of macula edema	MACULA_R, MACULA_L
Level of severity	ENCOUNTER_ICD9_CODE

POAG EHR Algorithm

The steps implemented for the DR EHR algorithm were replicated for the POAG algorithm with the following changes: diagnosis code list, format of procedure codes, and replacement of macula edema and level of severity conditions with cup to disc ratio and optic disc exam.

The POAG EHR algorithm is presented below:

Denominator

Patients seen in the study clinic during 2014 who were 18 years or older on the start date (1/1/2014) of the study period
AND
Encounters with diagnosis on list of codes provided
AND
Patients with two or more encounters
AND
Encounters associated with a faculty provider
AND
(Encounters with a service procedure code on list of codes provided
OR
Encounters with diagnosis on list of codes provided)

Numerator (Applied against encounters qualified denominator)

Cup to disc ratio
AND
Examination of the optic disc

POAG EHR Algorithm Denominator

For the POAG measure, the list of diagnosis codes specified for the POAG algorithm was used to execute the diagnosis criterion. Again, the list included non-POAG diagnosis codes such as 249.50 –Secondary stage diabetes mellitus. The POAG diagnosis list consisted of 24 unique ICD9 codes across 98 rows of code sets (See Table 7 for sample).

Table 7. Examples POAG EHR algorithm diagnosis value sets

Singular ICD9 value sets	Grouped ICD9 value sets
249.50	249.50, 362.01
250.50	250.51, 365.44
250.51	365.11, 365.70
365.10	365.12, 365.74
365.11	365.60, 365.70
365.12	250.51, 362.02, 362.07
365.15	365.10, 379.31, 365.71
365.44	365.44, 249.50, 362.05, 362.07
	250.50, 362.07, 362.05, 365.44

There were 8 singular codes and 90 sets of code groupings. Each grouping contained a range of 2 to 4 ICD9 codes. Almost three quarters of the diagnosis code sets for POAG had either two or three ICD9 codes; 19 percent had four ICD9 codes and only 8 percent had one ICD9 code (Table 8).

Table 8. Count of POAG diagnosis code sets (n = 98) by number of ICD9 codes

Number of ICD9 codes per code set	Count of code sets, No. (%)
One	8 (8.2)
Two	34 (34.7)
Three	37 (37.8)
Four	19 (19.4)

The format of one ICD9 code on the list was changed to 1 decimal place to match the study data. Similarly, the format of the procedure codes on the POAG list was changed from two decimals to 1 decimal place to match the study data (See Table 9).

Table 9. Implementation of procedure criterion

EHR Algorithm Procedure Code Formats	Study Data Procedure Code Format	Match Between EHR Algorithm and Study Data Formats
11111.10	11111.1	No
11111.1	11111.1	Yes

POAG EHR Algorithm Numerator

Again, the methods employed for the DR numerator were also applied for the POAG algorithm. The only difference was the measure specific columns. Disc and cup to disc ration records were used to assess whether or not an examination of the disc was conducted. The cup to disc ratio and optic disc exam requirements were executed by determining whether or not an encounter had a non-null (empty) record for *R and L cup to disc ratio* and R *and L disc*. The POAG EHR guidelines are summarized in Table 10.

Table 10. POAG EHR algorithm steps implemented

EHR algorithm rules	Data columns from data warehouse
Denominator	
18 years and older	BIRTHDATE
POAG diagnosis	ENCOUNTER_DIAGNOSIS
More than 1 encounter per patient	ENCOUNTER_ID, ENCOUNTER_DATE
Faculty provider	PROVIDER_ROLE
Service procedure code	LEVEL_SVC_CODE
Numerator	
Cup to disc ratio	CD_RATIO_R, CD_RATIO_L
Optic disc exam	DISC_R, DISC_L

DR and POAG NQF Guideline Implementations

As with the EHR algorithm, the NQF guidelines for the DR and POAG measures were executed using the data columns identified from the data warehouse. Measure components were classified at the encounter level for each component and then combined to assess status on the denominator and numerator as a whole. Indicator variables were created to categorize (1 = Yes, 0 = No) encounters for each component. Although the NQF guidelines were executed using the same process as the EHR algorithm, the list of diagnoses, service codes, and conditions for the numerator differed. Likewise the EHR algorithm, patient and medical exceptions could not be implemented due to the lack of data. In addition, restrictions on the number of appointments per patient and provider role were absent from the NQF guidelines.

DR NQF Guideline

Denominator

> Patients seen in the study clinic during 2014 who were 18 years or older on the start date (1/1/2014) of the study period
> AND
> Encounters with NQF diagnosis code(s)
> AND
> Service encounter type

Numerator (Applied against encounters in denominator)

> Level of severity of retinopathy findings
> AND
> (Macula edema findings absent OR Macula edema findings present)

DR NQF Guideline Denominator

DR diagnosis under the NQF guideline was dependent upon six ICD9 codes. All six codes were applied individually; none of the codes were assembled into groupings. The ICD9 codes provided in the NQF guidelines for the identification cases with DR were:

362.01 - Background diabetic retinopathy
362.02- Proliferative diabetic retinopathy
362.03 - Nonproliferative diabetic retinopathy NOS
362.04 - Mild nonproliferative diabetic retinopathy
362.05 - Moderate nonproliferative diabetic retinopathy
362.06 - Severe nonproliferative diabetic retinopathy

Specific encounter types were outlined in the NQF guidelines and covered several care settings and services: care services in long-term residential facility, nursing facility visit, office visit, fact to face interaction, ophthalmology services, and outpatient consultation. Some of these parameters, long-term care facility and nursing facility, were not applicable to the study setting and not implemented. The fact to face interaction data concept required the use of Systemized Nomenclature of Medicine- Clinical Terms (SNOMEDCT) codes to implement the condition. Except Current Procedural Terminology (CPT) and not SNOMEDCT was used to code services in the study data. Moreover, the study was inherently limited to in-person interactions because it only included "kept" or "arrived" appointments. Meaning that a patient had to see a provider in-person for the encounter to be included in the study. Office visit, ophthalmology services, and outpatient consultation were the only data concepts that were applicable to study and could be implemented using Current Procedural Terminology (CPT) codes.

DR NQF Guideline Numerator

Documentation of either the presence or the absence of macula edema for each eye and documentation of the level of severity were required to meet the numerator criteria. Similar to the DR EHR algorithm, the encounter diagnosis provided a surrogate for level of severity. The level of severity requirement was executed by determining whether or not the ICD9 code(s) assigned to an encounter included a code for diabetic retinopathy that specified the level of severity. The ICD9 codes for diabetic retinopathy with a level of severity were:

362.02 – Proliferative diabetic retinopathy
363.04 –Mild non proliferative diabetic retinopathy
365.05-Moderate nonproliferative diabetic retinopathy
365.06-Severe nonproliferative diabetic retinopathy

Ascertaining documentation on the presence or absence of macula edema was accomplished by searching *Macula right (eye)* and *macula left (eye)* data columns for terms acquired from the Value Set Authority Center (VSAC). VSAC is the official central repository for value sets contained in the electronic clinical quality measures and is overseen by the National Library of Medicine[1]. Data concepts in the NQF guidelines were assigned object identifiers, which pointed to specific value sets in the VSAC repository. Value sets consisted of numeric codes and descriptions from standard vocabularies such as Current Procedural Terminology (CPT), ICD-9-CM, Logical Observation Identifiers Names and Codes (LOINC), and Systematized Nomenclature of Medicine- Clinical Terms(SNOMEDCT) (67).

Table 9 contains the list of descriptions associated with SNOMEDCT codes in the value sets for presence of macula edema and absence of macula edema. The table does not include the actual SNOMEDCT codes because the macula data columns contained clinical findings recorded in text and not a numeric code format. Entries in the macula data columns had to be compared to the descriptions in the value sets to identify whether or not a record contained any of the descriptions in either value set. Matching records to value set descriptions was not an easy prospect because the terms used to document findings were not guaranteed to be an identical match those in the value sets. For example, abbreviations or different phrasing may have been used to represent the same descriptions in the value sets. Even the inclusion of an extra space between identical words would inhibit matching. Tools such as natural language processing could have been used to address the issue but were beyond the scope and resources of the study.

Table 11. DR NQF CQM: Macula edema value sets

Presence of macula edema	Absence of macula
Advanced diabetic maculopathy (disorder) Cystoid macula edema (disorder) Diabetic maculopathy (disorder) Clinically significant macula edema (disorder) Diabetic macula edema (disorder) Postoperative cystoid macula edema (disorder) Autosomal dominant cystoid macula edema (disorder) Uveitis related cystoid macula edema (disorder) Diffuse diabetic maculopathy (disorder) Focal diabetic maculopathy (disorder) Ischemic diabetic maculopathy (disorder) Mixed diabetic maculopathy (disorder) Macula retinal edema (disorder) Diabetic macula edema not clinically significant (disorder) Exudative maculopathy associated with type I diabetes mellitus (disorder) Exudative maculopathy associated with type II diabetes mellitus (disorder) Noncystoid edema of macula of retina (disorder)	Macula edema absent (situation)

A manual approach was taken to solve the problem of classifying records for the macula edema conditions. Approximately 400 records for macula R and L were manually reviewed to identify regular expressions (or phrases) in the study data that matched to descriptions in the value sets (See Table 11). The expressions were then used to write SQL queries to automate the process of classifying records for presence or absence of macula edema (See example in Table 12).

Table 12. Expressions in macula R and L records (study data) matched to value set descriptions

Value set description	Expressions from macula R and L entries
Macula edema present	
Clinically significant macula edema	Clinically significant macula edema
Cystoid macula edema	CME
Macula retinal edema	Marked edema of retina
Macula edema absent	
Macula edema absent	No macula edema, no edema, no frank edema, no definite edema, no frank macula edema

Documentation on the macula for each eye had to definitively state either the presence or absence of macula edema. Encounters lacking documentation on both eyes or with documentation on only one eye were classified "fail" (Assigned value of "0") for documentation of the presence or absence of macula edema.

POAG NQF Guideline

Denominator

Patients seen in the study clinic during 2014 who were 18 years or older on the start date (1/1/2014) of the study period
AND
Encounters with NQF diagnosis code(s)
AND
Service encounter type

Numerator (Applied against encounters in denominator)

Cup to disc ratio
AND
Optic disc exam

POAG NQF Guideline Denominator

Implementation of age, diagnosis, and service procedure criteria were patterned after the DR NQF CQM. The lone change was the list of diagnosis codes. The ICD9 codes provided in the NQF guidelines for the identification of cases with POAG were:

365.10 - Open-angle glaucoma, unspecified
365.11- Primary open angle glaucoma
365.12 - Low tension open-angle glaucoma
365.15 - Residual stage of open angle glaucoma

POAG NQF Guideline Numerator

Disc R and *disc L* and *cup to disc ratio R* and *cup to disc ratio L* had to be completed to designate the occurrence of an optic disc exam.

Reference Standard

Individual medical records containing clinic notes were manually abstracted via the ophthalmology EHR user interface to obtain data for the reference standard. Manual abstraction of the medical records allowed access to information in the clinic notes, which were not part of the data used to implement the EHR and NQF guidelines. Data for the reference standard were acquired by completing these steps:

1. Define reference standard criteria
2. Pilot test abstraction guide
3. Abstract and classify encounters under reference standard

Reference Standard Defined

The purpose of the reference standard was to set a benchmark for comparison against the NQF and EHR guidelines. The reference standard reflected the intent of the NQF summary statements and used sound clinical principles to develop a comprehensive and precise means of identifying measure components for the DR and POAG CQMs.

NQF eMeasure Summary Statements

I. Diabetic retinopathy: Percentage of patients aged 18 years and older with a diagnosis of diabetic retinopathy who had a dilated macula or fundus exam performed which included documentation of the level of severity of retinopathy and the presence or absence of macula edema during one or more office visits within 12 months.

II. Primary open-angle glaucoma (POAG): Percentage of patients aged 18 years and older with a diagnosis of primary open-angle glaucoma (POAG) who have an optic nerve head evaluation during one or more office visits within 12 months.

The standard was developed under the guidance of an ophthalmologist and was reviewed and approved by a second ophthalmologist. The medical record of qualifying encounters was manually abstracted using the reference standard. Medical records were manually abstracted by viewing individual records via the user interface of the EHR system. Qualified encounters consisted of those with a diagnosis of DR or POAG under the NQF or EHR algorithm guidelines. Manual abstraction offered access to documentation recorded by an ophthalmologist or a scribe under the supervision of a physician. Particular sections of the clinical template were reviewed for abstraction: documentation in the ophthalmology exam module and notes in the assessment and plan section. Manual abstraction was advantageous because it afforded access to unstructured clinic notes in the assessment and plan section, which are not viable for large-scale or automated reporting and were not part of the EHR or NQF implementations.

Specific data elements reviewed in the ophthalmology exam module included dilation and right and left eye data elements for cup to disc ratio, disc, macula, periphery, and vessels. These data elements were the data entry points for corresponding data columns used to implement the EHR and NQF guidelines. The reference standard was presented in the form of an abstraction guide and the primary components are described as follows:

DR Reference Standard

A definitive diagnosis of diabetic retinopathy had to be recorded in the Assessment and Plan to meet the denominator criteria. Documentation of a dilated exam, presence or absence of macula edema for both eyes, and level of severity and completion of the disc, macula, vessels and periphery data elements were required to meet the numerator criteria.

Denominator

> 18 years or older on start date of study period
> AND
> Diagnosis of diabetic retinopathy recorded in Assessment and Plan

Numerator

> Dilated exam section completed
> AND
> Presence or absence of macula edema documented for both eyes
> AND
> Level of severity documented in Assessment and Plan
> AND
> Disc, vessels, and periphery completed for both eyes

Exceptions

> All numerator components had to be completed unless a patient or medical exception was documented.

Complete documentation on dilation entailed recording of which eye(s) was dilated, date and time of dilation, and the medication administered for dilation. For documentation of the presence or absence of macula edema both records for the macula had to indicate either the presence (edema, thickening, clinically significant macula edema (CSME), macula edema present, macula edema, or diabetic macula edema (DME)) or absence (no edema, flat, no thickening, no clinically significant macula edema (no CSME), macula edema absent, no diabetic macula edema (no DME)) of macula edema.

Table 13. DR gold standard criteria

Component	Source	Values
Denominator Diagnosis of diabetic retinopathy	Assessment & plan	Diabetic retinopathy, DR
Numerator Dilated exam Dilated eyes Dilation time Dilation drops	Ophthalmology exam module	Both eyes OR left or right with exception for single eye exam Date and time Medication administered
Documentation of presence or absence of macula edema	Ophthalmology exam module or assessment and plan	Macula R and macula L carry either presence or absence of macula edema. Presence of macula edema: Edema, thickening, clinically significant macula edema (CSME), macula edema present, macula edema, or diabetic macula edema (DME) Absence of macula edema: Edema, thickening, clinically significant macula edema (CSME), macula edema present, macula edema, or diabetic macula edema (DME)
Complete fundus exam	Ophthalmology exam module or assessment and plan	Both eyes completed for disc, vessels and periphery
Documentation of level of severity	Assessment and plan	Proliferative diabetic retinopathy or mild, moderate or severe nonproliferative diabetic retinopathy
Disc Vessel Periphery	Ophthalmology exam module	Documentation on both eyes completed

Or the assessment and plan had to document whether or not macula edema was present or absent for both eyes. In addition, the level of severity (mild, moderate, sever, or proliferative) had to be documented in the assessment and plan and disc, periphery, and vessel data elements for both eyes had to be completed. Exceptions, medical or patient, had to be noted in the absence of the required information. Otherwise the encounter was classified as “0” (failed) for the component. The criteria for the NQF DR measure are summarized in Table 13.

POAG Reference Standard

The assessment and plan section of each encounter had to report a diagnosis of primary open-angle glaucoma to meet the reference standard denominator requirement. Right and left disc and cup to disc ratio had to be completed for to meet the numerator criteria for the reference standard (See Table 14). An encounter was labeled as “0” (failed) in the absence of each requirement, unless an exception was documented for the numerator.

Denominator

18 years or older on start date of study period
AND
Diagnosis of primary open-angle glaucoma recorded in Assessment and Plan

Numerator

Disc completed for both eyes
AND
Cup to disc ratio completed for both eyes

Exceptions

All numerator components had to be completed unless a patient or medical exception was documented.

Table 14. DR gold standard criteria

Component	Source	Values
Denominator		
Diagnosis of primary open-angle glaucoma	Assessment & plan	Primary open-angle glaucoma, POAG, glaucoma
Numerator		
Disc		
Cup to disc ratio	Ophthalmology exam module	Right and left eye completed

Abstraction Pilot Test

An abstraction guide was developed for each CQM using the reference standard criteria. A pilot study was conducted to test the clarity of the guides. The guides were edited as necessary and finalized. The researcher, a non-clinician, independently abstracted 10 randomly selected records each for diabetic retinopathy and primary open-angle glaucoma to assess their competency. An ophthalmologist reviewed the abstracted records. Ninety percent of the diabetic retinopathy and 100% of the primary open-angle glaucoma cases were abstracted successfully. The incorrect case was reviewed with the ophthalmologist. Abstraction for the study commenced upon completion of the pilot study.

Classification Under Reference Standards

The medical records of encounters that met the diagnosis criteria under the NQF and EHR guidelines for the DR and POAG CQMs were abstracted according to the reference standard. Encounters were categorized as “met” (1) or “did not meet” (0) for each component. All of the components were combined to determine whether or not an encounter met the denominator criteria. Encounters that qualified for the denominator were also classified for the numerator.

Data Preparation

Data analysis only included patients with a DR or POAG diagnosis under the EHR or NQF guidelines. Diagnosis, denominator eligibility and numerator performance were determined for the corresponding encounters under the reference standards. Each eligible encounter was classified as "yes" or "no" for denominator and numerator for each of the three methods. For example, only patients with a DR or POAG diagnosis per the NQF guideline or EHR algorithm were evaluated for the study. Consequently, patients would have a value of "yes" for the NQF and EHR guidelines for diagnosis but could have a either a value of "yes" or "no", depending on the information documented, for the reference standard. Individual datasets of encounters classified under the EHR algorithm and NQF guidelines and the corresponding reference standard was compiled for the DR and POAG measures. A third dataset was created to compare encounters classified under the NQF guidelines to classifications under the EHR algorithm. The EHR and NQF implementations were compared to the reference standard.

The reference standard, NQF and EHR algorithms were carried out at the encounter level. Some patients were found to have more than one encounter in the respective datasets. But the statistical methods and overall analysis of performance on a measure only allowed for one record per patient. The patient, and not the encounter, had to be used as the unit of analysis. The data were subsequently aggregated so that each patient was only represented once in each dataset. The encounter(s) of each patient were reviewed to determine overall classification of a patient as "yes" (met) or "no" (did not meet) for each diagnosis, denominator and numerator rule. Patients with at least one encounter that met the implemented criteria were categorized as "yes" (met criteria) and patients that did not have any encounters that met the implemented criteria were categorized as "no" (did not meet criteria).

For example, a patient with two encounters in which one met the diagnosis requirement but the other did not was classified as having met the diagnosis requirement. Another example is one in which the patient had two encounters and both had the appropriate diagnosis but one met the denominator criteria and the other did not; the patient was classified as met for the diagnosis and denominator. Last example is for a patient with two encounters that met the diagnosis and denominator criteria but had differing values (met and did not meet) for the numerator. The patient was classified as having met all requirements because it had at least one encounter that met all three indicators. To summarize a patient was required to have at least one encounter that met the diagnosis, denominator or numerator criteria to be categorized as met for these components. Otherwise, the patient was categorized as "0" (did not meet) for the entity. The respective datasets of classifications for diagnosis, denominator, and numerator per patient for each CQM and method combination were used for data analysis.

Data Analysis

The primary goal of the study was to assess the accuracy of using EHR data to compute quality measures using an EHR-based algorithm and the NQF guidelines. To accomplish this, the individual components of EHR and NQF guidelines were implemented using data from the data warehouse to classify encounters for each denominator and numerator. Encounters were also categorized according to the reference standard, using data from manual chart review. CQMs are typically reported in the form of the proportion -proportion of patients eligible for denominator who met the numerator criteria. And statistical analysis could have been set on simply comparing the percentages between the NQF and EHR implementations and the reference standard. But such an approach would have been inadequate because it ignores the underlying data used to compute the each proportion.

Proper analysis involved comparing the classification of individual patients under the EHR and NQF algorithms to the reference standard. The three areas compared were diagnosis, denominator and numerator. Comparisons were also performed between the EHR and NQF guidelines. Statistical analysis was conducted using Stata version 13 (College Station, TX) (68). The following pairwise comparisons were made for the DR and POAG measures: EHR algorithm versus reference standard, NQF guideline versus reference standard, and EHR algorithm versus NQF guidelines. Evaluation of numerator classification was limited to patients that were eligible for the denominator under the two methods being compared.

This is because a valid value (1-met or 0-did not meet) was required for the numerator of each method to execute the statistical comparison. For example, patients eligible for the NQF DR denominator with a value (either yes or no) for the numerator but was ineligible for the denominator under the reference standard, and as a result did not have a value for the numerator under the reference standard, were not included in the analysis.

Descriptive statistics were presented and the kappa statistic was computed to evaluate the level of agreement between the methods used to classify patients. The kappa statistic was used to compare how cases were identified. The percentage was not used because it does to take into account how patients were classified. Sensitivity, specificity, positive predictive value and negative predictive value were computed. The kappa statistic was only computed for denominator and numerator comparisons. It was not computed for diagnosis because the datasets used for analysis were based on patients with ICD9 diagnosis codes under the NQF or EHR guidelines. Patients without either diagnosis were not included for two reasons: 1) Measures are built around patients with a specific disease and these patients did not carry the appropriate ICD9 codes and 2) Resources were not available to abstract the 8729 patients without the diagnoses of interest. Furthermore, kappa can only be computed when binary (0 and 1) values are present for both entities under comparison. Since selection was limited to patients with a value of "yes" for a DR or POAG diagnosis and patients with a classification of "no" were not included, kappa cannot be computed. There were approximately 8729 patients that did not meet either diagnosis guideline. Because resources were unavailable to validate whether or not all (8729) patients were classified appropriately as "did not meet" for the diagnosis guidelines, a limited evaluation (described below) was conducted to estimate the number of cases that were misclassified by either the NQF or EHR algorithm.

Misclassified Diagnosis Estimation

Three percent (307/9036) of the study patients met the DR or POAG diagnosis requirements using the EHR algorithms and NQF guidelines. The remaining 8729 unique patients were not included in the study because they did not have any encounters that met the diagnosis requirement of under the EHR or NQF guidelines; 512 of these patients did not have a diagnosis coded on their encounter. In all, 8217 patients had a non-DR or POAG diagnosis code per the NQF guideline and EHR algorithm. Application of the diagnosis rules was dependent upon the ICD9 codes assigned to encounters. This means it is possible that some patients might have been diagnosed with DR or POAG but their diagnosis was not reflected in the ICD9 code(s) ascribed to their encounter(s).

Resources were unavailable to abstract all 8217 patients to ascertain whether or not a written diagnosis for DR or POAG was present in their records. To address this issue, a limited study was conducted to estimate the percent of patients without the appropriate ICD9 code but had either diagnosis of interest documented in the assessment and plan. Of the 8217 patients without the applicable diagnosis codes, 160 were randomly selected for review. The assessment and plan section in the medical records of the selected cases was manually reviewed to determine whether or not any of the patients had a written diagnosis of DR or POAG. An estimate of the percentage of patients who were excluded from the study but had either a written diagnosis of DR or POAG was computed.

Qualitative Evaluation

Participants

A recruitment letter and information sheet was emailed to all five ophthalmologists affiliated with the study clinic. Four of the five ophthalmologists consented to an interview.

Data collection and analysis

Face-to-face semi-structured interviews were conducted with ophthalmologists at a general eye clinic (same study location as quantitative evaluation) to assess the perceptions and self-reported behavior with respect to documentation and quality reporting. Each interview was approximately thirty minutes long. The interviews took place at a location that was convenient to the participant. An interview guide was developed and used to conduct the interviews. See Table 15 for a list of questions. The questions solicited information on the following topics: Awareness of quality measures, perception of quality reporting measures, training related to documentation, and documentation practices as they relate to specific components associated with the diabetic retinopathy (DR) and primary open-angle glaucoma (POAG) clinical quality measures.

A digital audio device was used to record the responses of each participant. Oral consent was recorded at the start of each interview session, a process approved by the IRB. Visual aids were used during the interview: A print out of the clinical template was used to assist with questions related to documentation practices and a print out of the NQF summary statement for each measure was presented. Participants were allowed to respond openly to each question and were not restricted to a pre-determined set of responses. Some of the questions in the interview guide were revised for clarification after completion of the first interview. Follow-up questions were posed as necessary during each interview to solicit more details from respondents when responses were abbreviated or ambiguous. The transcript of preceding interviews was reviewed to identify additional questions to be posed at subsequent interviews. Questions were added to shed light on areas that required further exploration and had not been covered in the initial interview. The audio recordings were transcribed to text. Each transcript was careful reviewed to identify and tally the range of responses to specific questions. Themes were identified from the responses in the initial set of interviews. Codes were added as additional transcripts were reviewed. The updated code set was then applied to subsequent transcripts. The process was iterative. New themes that emerged in subsequent transcripts were added to the template, which was then reapplied against all transcripts.

Table 15: Interview Guide

- Are you familiar with any quality measures related to patients with diabetic retinopathy/primary open-angle glaucoma?
- Besides normal eye exam features such as vision and pressure, what are the most important clinical components that you specifically document for a patient with diabetic retinopathy/primary open-angle glaucoma?
- Where do you document this information in the EHR?
- Are there any evidence-based guidelines that you follow when documenting for patients with diabetic retinopathy/primary open-angle glaucoma?
- Are there any requirements that you have to adhere to when documenting on a patient with diabetic retinopathy/primary open-angle glaucoma?
- Does a technician assist with documentation during a clinic visit? If so, please describe how you collaborate with the technician when documenting?
- What are your thoughts on documenting specific clinical components that may be needed to compute a particular clinical quality measure?
- What impact would a report on the diabetic retinopathy and primary open-angle glaucoma measures have on clinical quality?

Several rounds of reviews were conducted to ensure that the list of codes was comprehensive. The same codes were used to label similar sentiments expressed across all of the interviews. On the outset, the list of codes was quite detailed and unwieldy. Codes were subsequently consolidated into broader, overarching themes to create a more manageable list. Each transcript was annotated using the finalized list of codes. A second researcher reviewed the data and findings to verify that the summaries were accurate. The results of the qualitative study were presented in line with the identified themes.

4. Results

Quantitative Results

EHR Algorithms

Classification for DR and POAG Measure Denominators

Diagnosis

The denominator criteria consisted of the following: diagnosis, patient with two or more encounters, encounter with a faculty provider, and encounters with at least one of the specified service procedure codes. Each measure was accompanied by a list of International Classification of Diseases, Ninth Revision (ICD9) codes to identify encounters with the conditions of interest. The diagnosis criteria were applied independently against the study data because the list of ICD9 codes differed between the two measure guidelines.

Most of the 191 encounters with a DR diagnosis code under the EHR guidelines had either one or two ICD9 codes: 174 had singular ICD9 codes and 17 had two ICD9 codes. The same was observed for POAG in that 342 encounters met the diagnosis requirement under a singular code and 2 met the criteria with 2 ICD9 codes. Code sets with three or more ICD9 codes were not present because encounters had already qualified under code sets with fewer codes. One of the singular ICD9 codes for POAG, 365.15 - Residual stage of open angle glaucoma, was absent from the study data.

Out of the 15740 encounters (9036 unique patients), 191 encounters (173 unique patients) met the DR diagnosis criteria and 344 encounters (229 unique patients) carried the diagnosis requirements outlined for the POAG measure. Fifteen thousand five hundred and nine encounters (8729 unique patients) did not carry either diagnosis criteria; 730 (512 unique patients) of these encounters did not have a diagnosis code.

Common Denominator Rules

The shared rules for the DR and POAG measure denominators were patients with more than one encounter, encounters with a faculty provider, and encounters with specified service procedure codes. These rules were applied independently against encounters with either a DR or POAG EHR algorithm diagnosis.

Table 16. Count of DR and POAG encounters that met denominator rules

Denominator rule	No. (%)	
	DR (n = 191)	POAG (n = 344)
Number of encounters per patient		
One	106 (55.5)	91 (26.5)
Two	38 (19.9)	84 (24.4)
Three	11 (5.8)	65 (18.9)
Four	10 (5.2)	36 (10.5)
Five	6 (3.1)	19 (5.5)
Six	2 (1.0)	18 (5.2)
Seven	7 (3.7)	20 (5.8)
Eight	5 (2.6)	6 (1.7)
Nine	2 (1.0)	2 (0.6)
Ten	-	3 (0.9)
Eleven	3 (1.6)	-
Twelve	1 (0.5)	-
Faculty	191 (100)	344 (100)
Service procedure code		
New patient visit	23 (12.0)	18 (5.2)
Established patient visit	40 (20.9)	163 (47.4)
Eye exam & treatment	122 (63.9)	133 (38.7)
Post operative visit	6 (3.1)	17 (4.9)
Special procedure	-	10 (2.9)
Vision exam	-	3 (0.9)

The results are presented in Table 16. Forty-five percent of the DR encounters (N = 191) were associated with a patient who had two or more encounters during the study period. The number of encounters per patient ranged from 2 to 12. Slightly more than half (56%) of the encounters were associated with a patient who did not meet the requirement for more than one encounter during the study period. All of the DR encounters met the faculty requirement. Four categories of service procedure codes were observed among encounters with a DR diagnosis: post operative, new patient, established patient, and eye exam and treatment visit. Post-operative visit was not on the list of service codes for either measure.

Eighty-five of the191 DR encounters the patient-encounter count requirement; all 191 encounters met the faculty requirement. Five of the 85 encounters did not meet the service procedure requirement. The final count of encounters that met the denominator criteria would have been 80 but the algorithm invoked the diagnosis requirement a second time in the last line of the denominator rules listed above. The statement includes the “OR” logical operator. This means that an encounter only has to meet either the service code or the diagnosis requirement but not both. Because of this statement, the number of encounters that met the denominator criteria was 85 and includes 5 encounters that did not meet the service procedure requirement but met the diagnosis requirement. In summary, 85 encounters were classified as “1” (met) and 106 were classified as “0” (did not meet) for the DR EHR algorithm denominator.

Table 17. Count of DR and POAG encounters by denominator rule

Denominator rule	CQM measure	
	DR (n = 191)	POAG (n = 344)
MRN with two or more encounters, No. (%)	85 (44.5)	253 (73.5)
Faculty, No. (%)	191(100)	344 (100)
Service procedure code OR Diagnosis, No. (%)	85 (44.5)	253 (73.5)

Of the 344 POAG encounters, 73% were affiliated with patients who had met the two or more encounter count per patient requirement (Table 17). The maximum number of encounters per patient was 10. Seventy percent of POAG patients who met the more than one encounter requirement had between 2 and 7 encounters during the study period. All 344 POAG encounters met the faculty rule. A total of 30 encounters (31%) were assigned a service code (Post-operative visit, special procedure, vision exam) that was not on the list of codes for the EHR algorithm.

The qualifying service codes for new patient, established patient and eye exam and treatment visit were assigned to 91% of the 344 encounters with a POAG diagnosis. The three common denominator requirements were combined as follows:

> 18 years and older AND
> POAG diagnosis AND
> Patient with 2 or more encounters AND
> Faculty AND
> (Service procedure code OR POAG diagnosis)

Likewise the DR algorithm, the "OR" logical operator resulted in the inclusion of POAG encounters that did not meet the service code requirement. 224 encounters would have been classified as "met" for the POAG denominator in the absence of the "OR" operator. In the end, 253 of the 344 encounters with a POAG diagnosis were classified as "1" (met) and 91 were categorized as "0" (did not meet).

EHR Algorithm: Classification for DR and POAG Measure Numerators

The numerator criteria were applied against encounters that met the denominator criteria for the respective measure guidelines. The requirements for level of severity and the presence or absence of macula edema for each eye had to be met as outlined in the methods section: non-null record for macula right and macula left and an ICD9 diagnosis code that specified the level of severity. Among the 85 encounters that met requirements for the DR denominator, less than half (45%) met the level of severity rule. On the other hand, 89% met the documentation of macula edema requirement. Of note, some encounters had more than one record in the dataset for macula right and left even though the EHR user interface displayed only one textbox. Multiple records were present for a singular macula textbox because the text recorded in the front end was quite long; therefore, multiple records were created to hold all of the text. More than half (55%) of the encounters eligible for the DR denominator did not carry an ICD9 code that specified the level of severity.

The remaining forty-five percent of the DR encounters were coded with an ICD9 code designating the level of severity. Mild level of severity accounted for the highest proportion at 21%, followed by moderate at 14%. Proliferative and severe level of severity had the lowest proportions at 7% and 2%. Overall, 39% of denominator eligible DR encounters met both numerator requirements (Table 18).

Table 18. Count of DR and POAG denominator eligible encounters by numerator requirement

Numerator rule	No. (%)
DR measure (n = 85)	
Documentation macula edema	
Macula right	80 (94)
Macula left	79 (93)
Macula right and macula left	76 (89)
Level of severity	
Mild	18 (21.2)
Moderate	12 (14.1)
Severe	2 (2.1)
Proliferative	6 (7.1)
DR diagnosis level of severity not specified	47 (55.3)
Macula edema and level of severity	33 (39)
POAG measure (n = 253)	
Disc	
Disc right	193 (76.3)
Disc left	193 (76.3)
Disc right and left	183 (72)
Cup to disc ratio	
Cup to disc ratio right	204 (80.6)
Cup to disc ratio left	211 (83.4)
Cup to disc right and left	201 (79.4)
Disc and cup to disc ratio both eyes	178 (70.4)

With respect to the POAG numerator rules, a completed record had to be present for disc right and left and cup to disc ratio right and left. A significant proportion, 72%, of the POAG encounters met the disc requirements and higher proportion, 79%, met the cup to disc ratio requirements. When combined, 70% of the POAG encounters eligible for the POAG denominator met the numerator criteria (Table 3).

NQF Guidelines

Classification for DR and POAG Measure Denominators

Diagnosis and service encounter type make up the requirements for the DR and POAG measure denominators. Each measure had a unique list of ICD9 codes, which were applied independently against the study data of 15740 encounters (9036 unique patients). Table 19 shows a breakdown of the encounters per diagnosis code for each measure. Approximately 3% of the study patients had either a DR or POAG diagnosis under the NQF guidelines. The number of encounters with either diagnosis for NQF was 345 (232 unique patients). All six DR ICD9 codes specified were found in the data. Though some encounters had more than one DR ICD9 code, each code was applied independently and not in groupings. The number of encounters with a DR diagnosis was 106 (99 unique patients) and the number of encounters without a DR diagnosis was 14904. Only two encounters had more than one DR ICD9 code. One encounter had two DR ICD9 codes, background and moderate DR, and the second had three ICD9 codes (background, moderate, and severe DR). The remaining 104 encounters were assigned one DR ICD9 code each. Mild non-proliferative diabetic retinopathy (Mild NPDR) had the highest frequency at 46 (43.4%); moderate non-proliferative diabetic retinopathy (Moderate NPDR) was a distant second at 20 (18.9%). Background diabetic retinopathy (BDR) and proliferative diabetic (PDR) had similar distributions at 15% and 16%. Severe non-proliferative diabetic retinopathy (Severe NPDR) was assigned to only 4% of encounters and non-proliferative diabetic retinopathy was coded on only 1 encounter.

Table 19. Count of DR and POAG encounters by diagnosis

Diagnosis	No. (%)
DR measure (n = 106)	
Background diabetic retinopathy	16 (15.1)
Proliferative diabetic retinopathy	17 (16)
Non-proliferative diabetic retinopathy	1 (0.9)
Mild non-proliferative diabetic retinopathy	46 (43.4)
Moderate non-proliferative diabetic retinopathy	20 (18.9)
Severe non-proliferative diabetic retinopathy	4 (3.8)
Background & moderate diabetic retinopathy	1 (0.9)
Background, moderate & severe diabetic retinopathy	1 (0.9)
POAG measure (n = 241)	
Open-angle glaucoma unspecified	7 (2.9)
Primary open-angle glaucoma	212 (88)
Low-tension open-angle glaucoma	22 (9.1)

The POAG diagnosis had higher representation at 241 encounters (135 unique patients) and in fewer diagnosis categories, three. The number of study encounters without a POAG diagnosis was 14769. Primary open-angle glaucoma was well represented at 88% (212 encounters). Nine percent had a diagnosis of low-tension open-angle glaucoma and 3% were assigned a diagnosis code for primary open-angle glaucoma unspecified.

The number of encounters with the appropriate diagnosis and evaluated for the denominator of each measure was 106 for DR and 241 for POAG. The service type constraint reduced the number of encounters that met the denominator requirements from 106 to 103 for the DR measure and from 241 to 216 for the POAG measure. The counts per service code are outlined in Table 20. About 90% of encounters on each diagnosis case list were either established visits or coded as an eye exam and treatment. A very small subset of the encounters, 3% for DR and 6% for POAG, had service codes that were not in the NQF value set for service type. These service types included post-operative visit, special procedure, and vision exam.

Table 20. Count of DR and POAG encounters by service type

Service type	No. (%)
DR measure (n = 106)	
New patient visit	8 (7.5)
Established patient visit	21 (19.8)
Eye exam and treatment	74 (69.8)
Post operative visit*	3 (2.8)
POAG measure (n = 241)	
New patient visit	4 (7.1)
Established patient visit	142 (58.9)
Eye exam and treatment	70 (29)
Post operative visit*	13 (5.4)
Vision exam*	3 (1.2)
Special procedure*	9 (3.7)

*Did not meet denominator criteria

NQF guidelines: Classification for DR and POAG measure numerators

The numerator criteria were applied against 103 DR encounters and 216 POAG encounters that met the denominator requirements. Each denominator eligible DR encounter was required to have documentation of the presence or absence of macula edema and the level of severity.

Table 21. Count of DR encounters by numerator criteria

Numerator criteria	No. (%) (n = 103)
Documentation of level of severity	
Mild	45 (43.7)
Moderate	20 (19.4)
Moderate & severe	1 (1)
Severe	4 (3.9)
Proliferative	16 (15.5)
Level of severity not specified*	17 (16.5)
Documentation of presence or absence of macula edema (ME)	
Documentation ME Macula right	
Macula right presence of ME	1 (1)
Macula right absence ME	6 (5.8)
Macula right presence or absence ME	7 (6.8)
Documentation ME Macula left	
Macula left presence of ME	1 (1)
Macula left absence ME	6 (5.8)
Macula left presence or absence ME	7 (6.8)
Macula right and left documentation ME	4 (3.9)
Documentation of level of severity and macula edema	3 (2.9)

*Included ICD9 codes 362.01- Background diabetic retinopathy and 362.03-non-proliferative diabetic retinopathy

Presence or absence of macula edema was determined by evaluating macula records for the NQF values outlined in the methods section and level of severity was assessed using assigned ICD9 codes as detailed in the methods. Eighty percent of the 103 DR denominator eligible encounters had an ICD9 code that specified the level of severity (Table 21). One patient had two ICD9 codes with differing levels of severity, moderate and severe. A subset, 17%, of the DR encounters was assigned either 362.01(Background diabetic retinopathy) or 362.03 (Non-proliferative diabetic retinopathy), neither of which specified a level of severity. Very few encounters met the NQF DR numerator requirements. Only 7% of encounters had documentation on either eye and an even fewer (4%) had documentation of either the presence or absence of macula edema on both eyes. In combination, only 3% of DR denominator eligible encounters met both numerator requirements.

The numerator requirements for the POAG NQF measure were completed records for right and left disc and cup to disc ratio. Largely, the POAG numerator criteria were easier to meet compared to the DR requirements because about three quarters of the eligible POAG encounters achieved both requirements, unlike the poor performance of 3% for the DR measure. As shown in Table 22, most of the eligible POAG encounters had non-null records for disc and cup to disc ratio for both eyes. The minimum rate of completion across the four components was 78%. Cup to disc ratio for both eyes had a higher completion rate over disc at 87%.

Table 22. Count of POAG encounters by numerator criteria

Numerator criteria	No. (%) (n = 216)
Optic disc documentation	
Disc right	174 (80.6)
Disc left	169 (78.2)
Disc right and left	112 (77.3)
Cup to disc ratio documentation	
Cup to disc ratio right	189 (87.5)
Cup to disc ratio left	190 (88)
Cup to disc ratio right and left	187 (86.6)
Documentation disc and cup to disc ratio	163 (75.5)

Reference Standard Classification

The reference standard guidelines are outlined in the methodology. The medical record of encounters with a DR diagnosis (191) and a POAG diagnosis (344) under the EHR algorithm was manually abstracted to identify whether or not the selected encounters met the reference standard denominator and numerator criteria. The same process was repeated for 106 DR encounters and 241 POAG encounters with the designated diagnoses under the NQF guidelines.

Overall Performance per Measure and Method

According to the reference standard (Table 23), 71.5% of patients with a DR diagnosis under the NQF guidelines met the reference standard numerator criteria and 84.8% NQF POAG patients met the reference standard numerator criteria.

Table 23. Percent of patients that met the requirements under each method

Measure/method	Number patients with Dx. (n)	Met NQF guidelines			Met EHR algorithm			Reference standard		
		Denom. (n)	Num. (n)	% Met num.	Denom. (n)	Num. (n)	% Met num.	Denom. (n)	Num. (n)	% Met num.
DR										
NQF guidelines	99	97	3	3.1	-	-	-	88	63	71.5
EHR algorithm	173	-	-	-	67	31	46.3	135	88	65.2
POAG										
NQF guidelines	135	127	109	85.8	-	-	-	125	106	84.8
EHR algorithm	229	-	-	-	138	113	81.9	129	110	85.3

Abbreviations: Dx., diagnosis; Denom., denominator; Num., numerator.

Dataset for Analysis

The encounter level data reference standard, NQF and EHR guidelines were summarized per the methods described. For the EHR algorithm, 191 encounters (173 unique patients) were identified with a DR diagnosis and 344 encounters (229 unique patients) were identified for with a POAG diagnosis. The NQF guidelines yield 106 (99 unique patients) with a DR diagnosis and 241 (135 unique patients) with a POAG diagnosis. The same encounters were abstracted for the reference standard in each group. Then the data were summarized at the patient level for each guideline and reference standard pairing. Encounters classified under the NQF guidelines for DR and POAG were also compared to classifications under the EHR algorithm and summarized at the data aggregated per patient. The number of patients in each analysis pairing was as follows:

Table 24. Number of patients per analysis pairing

Analysis	Number of patients
DR EHR algorithm versus reference standard	173
POAG algorithm versus reference standard	229
DR NQF guidelines versus reference standard	99
POAG NQF guidelines versus reference standard	135
DR NQF guidelines versus EHR algorithm	99
POAG NQF guidelines versus EHR algorithm	135

DR EHR Algorithm versus Reference Standard Comparison

The categorization of diagnosis matched between EHR algorithm and reference standard for 79% of the 173 patients. All of the false positives had a DR ICD9 code per the NQF rules and did not have a DR diagnosis documented in the Assessment and Plan (reference standard). Of the 173 patients with a DR EHR algorithm diagnosis, only 39% met the denominator requirements (Table 25). Performance under the reference standard was much higher at 78%. The percent agreement 42.2% (95% CI, 0.4 – 0.5) and the kappa -0.026 (p value 0.686, 95% CI -0.134 – 0.082) indicated the absence of agreement on the classification of patients. There were slightly more discordant classifications (58%) than concordant classifications (42%).

Table 25. Comparison of classifications between DR EHR algorithm and reference standard

Parameter (n patients)	EHR Met No. (%)	Reference Met No. (%)	EHR & Reference Concordance No. (%), Concordance Yes, No. (%)	EHR & Reference Discordance No. (%)	EHR & Reference Percent agreement % (95% CI)	EHR & Reference Kappa, P value (95% CI), PABAK
Diagnosis* (173)	173 (100)	137(79)	137(79)	36(21)	-	-
Denominator (173)	67(39)	135(78)	73(42) 51(29)	100(58)	42.2 (0.4 – 0.5)	-0.026 0.686 (0.134 – 0.082) -0.16
Numerator (51)	26(51)	33(65)	28(55) 18(35)	23(45)	54.9 (0.4 – 0.7)	0.093 0.245 (-0.170 – 0.356) 0.10

*Percent agreement and kappa not computed because no data on patients without EHR algorithm diagnosis

Unlike the denominator, percent agreement between the methods for the numerator marginally surpassed the half way mark at 54.9% (95% CI 0.4 – 0.7) for the 51 denominator eligible patients. The kappa statistic was poor at 0.093 (p value 0.677, 95% CI -0.170 – 0.356) and likely due to because the confidence interval includes the null value.   Classification was concordant for approximately half of the patients. There were 24 discordant pairs.

POAG EHR Algorithm versus Reference Standard Comparison

Of the 229 patients with an EHR algorithm diagnosis for POAG, 56% matched the reference standard diagnosis (Table 26). Percent agreement and kappa were not computed for diagnosis because data were not collected on patients without EHR algorithm diagnosis. Similar percentages of the 229 patients met the denominator criteria for the EHR algorithm and the reference standard. The percentage of patients for the EHR algorithm was 60% and the reference standard had 56%.

On the whole, percent agreement was 67.2% (95% CI 0.6 – 0.7) for classification of the denominator between the EHR algorithm and reference standard. Kappa was graded as fair at 0.327 (p value 0.000, 95% CI 0.204 – 0.451). The kappa statistic was above the null as was the lower boundary of the confidence interval. This suggests that the kappa statistic was not due to chance.

Table 26. Comparison of classifications between POAG EHR algorithm and reference standard

Parameter (n patients)	EHR Met No. (%)	Reference Met No. (%)	EHR & Reference Concordance No. (%), Concordance Yes, No. (%)	EHR & Reference Discordance No. (%)	EHR & Reference Percent agreement % (95% CI)	EHR & Reference Kappa, P value (95% CI), PABAK
Diagnosis (229)	129 (56)	129(56)	129(56)	100(44)	-	-
Denominator (229)	138(60)	129(56)	154(67) 96(42)	75(33)	67.2 (0.6 – 0.7)	0.327 0.000 (0.204 – 0.451) 0.34
Numerator (96)	81(84)	82(85)	93(97) 80(83)	3(3)	96.8 (0.9 – 1.0)	0.878 0.000 (0.743 – 1.000) 0.94

+Prevalence adjusted bias adjusted kappa

The percent of patients (n = 96) that met the POAG numerator requirements under each method only differed by one percentage point. The proportion that met the numerator was 84% for the EHR algorithm and 85% for the reference standard. Percent agreement was sizeable at 96.8 (95% CI 0.9 – 1.0). And the kappa 0.878 (p value 0.000, 95% CI 0.743 – 1.000) rated as very good, strongly supported the observed percent agreement, and was statistically significant. Cases classified as "did not meet" for both methods had identical values for disc and cup to disc ratio components of the numerator. A minute subset of 3% had discordant categorizations for the numerator. One case "met" the reference standard but not the EHR algorithm rules because question marks were record for both right and left cup to disc ratio. These entries were viewed as valid under the EHR algorithm, which assessed for records that were not empty. Two cases were categorized as "met" for the reference standard but "did not meet" for the EHR algorithm. The disparities were present because of the application of an exception for one patient and missing records for another. Manual chart review provided access to the exception documented in the Assessment and plan. The patient had had an eye removed and only had documentation on one eye as opposed both eyes as required per the EHR algorithm. The second patient had records that were visible via the user interface but absent from the study data.

DR NQF versus Reference Standard Comparison

Classification of diagnosis, denominator and numerator were compared between the NQF guidelines and the reference standard for 99 patients with a DR diagnosis under the NQF guidelines (Table 27). Ninety percent of the patients with an NQF DR diagnosis had a matching diagnosis under the reference standard. The false positive rate for NQF DR diagnosis was 9%.

Table 27. Comparison of classifications between DR NQF guidelines and reference standard

Parameter (n patients)	NQF Met No. (%)	Reference Met No. (%)	NQF & Reference Concordance No. (%), Concordance Yes, No. (%)	NQF & Reference Discordance No. (%)	NQF & Reference Percent agreement % (95% CI)	NQF & Reference Kappa, P value (95% CI), PABAK
Diagnosis* (99)	90 (91)	90(91)	90(91)	9(9)	-	-
Denominator (99)	97(98)	88(89)	86(87) 86(87)	13(13)	86.8 (0.8 – 0.9)	-0.035 0.693 (-0.084 – 0.014) 0.74
Numerator (86)	3(3)	63(73)	26(30) 3(3)	60(70)	30.2 (0.2 – 0.4)	0.026 0.143 (-0.005 – 0.057) -0.40

*Percent agreement and kappa not computed because no data on patients without NQF guideline diagnosis
+Prevalence adjusted bias adjusted kappa

Agreement between NQF and reference standard classifications for the DR denominator was 86.8% (95% CI, 0.8 - 0.9). All 86.8% had a classification of “Met”; non were concordant with a classification of “Did not meet”. Only 13% of patients (n = 99) had differing classifications between the two methods; these patients were classified as “met” for one method and “did not meet” for the other, and vice versa. The percent agreement fell well within the narrow confidence interval and the lower boundary of the confidence interval was above the null. The kappa statistic of -0.035 (p-value 0.693, 95% CI -0.084 - 0.014) points to no reproducibility between the NQF guidelines and reference standard classifications for the denominator. The lower bound of the confidence interval also crosses the null, which supports the lack of reproducibility. However, the low kappa does not reflect the high agreement observed between the two methods. This discordance was due to an imbalance in the 2X2 table is at the basis of the low kappa because one of the cells has a value of zero. The prevalence-adjusted bias-adjusted kappa (PABAK) of 0.74 was more in line with the observed percent agreement.

Reproducibility (kappa 0.026, p value 0.143, 95% CI -0.005 -0.057) was also poor between the DR NQF and reference standard numerators for the 86 eligible patients. The large difference between the percent met for the numerator of the NQF guidelines (4%) and reference standard (73%) foreshadowed a marked difference in classification. A percent agreement of 30.2% (95% CI, 0.2 – 0.4) for numerator categorization between the two methods substantiated these observations. A total of 26 patients were concordant for the numerator classification: 3 met both numerator requirements and 23 did not. This meant 70% of patients had discordant classifications for the numerator.

POAG NQF versus Reference Standard Comparison

The classifications for diagnosis, denominator and numerator were compared between NQF guidelines and the reference standard for 135 patients. Ninety-three percent of the patients were positive for a POAG diagnosis and had the same classification between the two methods (Table 28). Seven percent were positive for a POAG NQF diagnosis but negative under the reference standard.

Table 28. Comparison of classifications between POAG NQF guidelines and reference standard

Parameter (n patients)	NQF Met No. (%)	Reference Met No. (%)	NQF & Reference Concordance No. (%), Concordance Yes, No. (%)	NQF & Reference Discordance No. (%)	NQF & Reference Percent agreement % (95% CI)	NQF & Reference Kappa, P value (95% CI), PABAK
Diagnosis* (135)	125 (93)	125(93)	125(93)	10(7)	-	-
Denominator (135)	127(94)	125(93)	119(88) 118(87)	16(12)	88.2 (0.8 – 0.9)	0.048 0.285 (-0.160 – 0.257) 0.76
Numerator (118)	102(86)	103(87)	115(97) 101(86)	3(3)	97.5 (0.9 – 1.0)	0.889 0.000 (0.765 – 1.000) 0.95

*Percent agreement and kappa not computed because no data on patients without NQF guideline diagnosis
+Prevalence adjusted bias adjusted kappa

Reproducibility was poor for categorization of patients (n =135) for the denominator between NQF guidelines and the reference standard, kappa 0.048 (p value 0.285, 95% CI -0.160 – 0.257). The low kappa was contradicted by a high percent agreement of 88.2% (95% CI 0.8 - 0.9). Again, the paradox was due to imbalance in the cell numbers of the 2X2 table as seen in the wide range of the marginal totals from a minimum of 8 to a maximum of 127.

Both the NQF and reference standard showed high adherence to the numerator requirements among the 118 eligible patients. The percent of patients that met each numerator was 86% and 87%, respectively, for the NQF guidelines and reference standard. Overall percent agreement was very good for the numerator between the two methods and approached perfect agreement at 97.5% (95% CI, 0.9 – 1.0). An equally high kappa of 0.889 (p value 0.000, 95% CI 0.765 – 1.000) corroborated the strength of the agreement. The kappa had a grading of very good. Only 3 patients had discordant values between the two methods. Discordance for 1 of the patients was because the cup to disc ratio records contained question marks. Since the NQF guidelines checked for the presence of non-null records, this patient was given a pass but was disqualified under the reference standard.

Discordance was observed for the remaining two patients because one had an exception that was accessible under the reference standard but inaccessible under the NQF guidelines, and the other had disc and cup to disc records that were also accessible via manual chart review for the reference standard but were absent from the study data. The exception exempted the patient from having documentation on disc and cup to disc ratio on both eyes because one eye had been enucleated.

DR EHR Algorithm versus NQF Guidelines Comparison

The dataset for analysis was based on 99 patients with a DR diagnosis per the NQF criteria. There was no difference in the classification for DR diagnosis between the NQF and EHR guidelines. All patients were classified as having a DR diagnosis for both the NQF guidelines and EHR algorithm.

Table 29. Comparison of classifications between DR EHR algorithm and NQF guidelines

Parameter (n patients)	EHR Met No. (%)	NQF Met No. (%)	EHR & NQF Concordance No. (%), Concordance Yes, No. (%)	EHR & NQF Discordance No. (%)	EHR & NQF Percent agreement % (95% CI)	EHR & NQF Kappa, P value (95% CI), PABAK
Diagnosis* (99)	99 (100)	99(100)	99(100)	-	-	-
Denominator (99)	38(38)	97(98)	38(38) 37(37)	61(61)	38.4 (0.3 – 0.5)	-0.008 0.633 (-0.054 – 0.039) -0.23
Numerator (37)	31(84)	2(5)	8(22) 2(5)	29(78)	21.6 (0.1 – 0.4)	0.022 0.261 (-0.013 – 0.056) -0.57

*Percent agreement and kappa not computed because no data on patients without NQF guideline diagnosis
+Prevalence adjusted bias adjusted kappa

The percent agreement was abysmal for the classification of the denominator and numerator between the two methods (Table 29). The percent agreement for the denominator was 38.4% (95% CI 0.3 – 0.5) and 21.6% (95% CI 0.1 – 0.4) for the numerator. Thirty-eight percent of the classifications for the denominator were concordant. Concordance for the numerator classifications was lower at 22%. This kappa statistics mirrored the percent agreement. A negative kappa of -0.008 (p value 0.633, 95% CI -0.054 – 0.039) was observed for the denominator comparison, demonstrating that agreement was less than chance. The kappa for the numerator comparison was 0.022 (p value 0.261, 95% CI -0.013 – 0.056). Though the kappa was positive, the lower boundary of the confidence interval crossed the null. And this means that the kappa could possibly be zero (no agreement) because zero fell within the confidence interval. In other words, the observed kappa could be due to chance.

POAG EHR Algorithm versus NQF Guidelines Comparison

Patients (135) with a POAG diagnosis under the NQF guidelines were used to create the dataset for this comparison. Each patient was classified as having a diagnosis of POAG under the NQF and EHR guidelines.

Table 30. Comparison of classifications between POAG EHR algorithm and NQF guidelines

Parameter (n patients)	EHR Met No. (%)	NQF Met No. (%)	EHR & NQF Concordance No. (%), Concordance Yes, No. (%)	EHR & NQF Discordance No. (%)	EHR & NQF Percent agreement % (95% CI)	EHR & NQF Kappa, P value (95% CI), PABAK+
Diagnosis* (135)	135 (100)	135(100)	135(100)	-	-	-
Denominator (135)	103(76)	127(94)	97(72) 96(71)	38(28)	71.9 (0.6 – 0.8)	-0.050 0.779 (-0.154 – 0.055) 0.44
Numerator (96)	83(100)	83(100)	96(100) 83(87)	0(0)	100.0 (0.9 – 1.0)	1.000 0.000 (1.000– 1.000) 1.00

*Percent agreement and kappa not computed because no data on patients without NQF guideline diagnosis
+Prevalence adjusted bias adjusted kappa

The percent of patients that met the denominator criteria for the NQF guideline was roughly 20% higher than the EHR algorithm (Table 30). However, the proportion of patients that met the respective numerator criteria differed was identical at 100%. The percent agreement was 71.9% (95% CI .6 – 0.8) for the denominator and an impresive100% for the numerator. The kappa, -0.050 (p value 0779, 95% CI -0.154 – 0.055), was surprisingly low and did not echo the magnitude of percent agreement. This contradiction was likely due to the imbalance in the 2X2 table. PABAK is much higher than kappa at 0.44 and better reflects the relatively high percent of agreement. As expected, the kappa for the numerator comparison was 1.00, since classification was identical between the two methods.

DR and POAG EHR and NQF Guideline Proportions versus Reference Standard

The proportions are presented in Table 31. The percent of patients who met the implemented guidelines was higher for the POAG measure compared to the DR measure. A higher proportion of patients met the DR EHR algorithm implementation over the NQF guideline. There was only a 1% difference between the proportion of patients who met the POAG NQF implementation and the reference standard. The percent difference between the POAG EHR algorithm and the reference standard was slightly higher at 3%. This disparity in the proportions that met the guidelines was much higher at 20% between the DR EHR algorithm and the reference standard. An even greater difference (69%) was observed when the percent of patients who met the DR NQF guideline was compared to the reference standard. It is important to recognize that regardless of whether or not the proportions were similar or different between the NQF/EHR guidelines and the reference standard, the proportions do not take into the individual classification of patients under the rules that were implemented. In other words, the proportion does not reflect whether or not the classification of patients under the respective guidelines match the reference standard as described in preceding tables.

Table 31. Comparison of DR and POAG EHR and NQF guideline proportions to reference standard

Condition	Diabetic retinopathy		Primary open-angle glaucoma	
	% Met CQM	Absolute difference, P value	% Met CQM	Absolute Difference, P value
EHR algorithm	46.3	19.7	81.9	3.4
EHR algorithm reference standard	65.9	0.007	85.3	0.455
NQF guideline	3.09	68.5	85.8	1.03
NQF guideline reference standard	71.6	<0.0002	84.8	0.818

Sensitivity and Specificity for Denominators and Numerators

Sensitivity and specificity were computed for all of the denominator and numerator comparisons (Table 32). These results complement the agreement analysis described above. The DR NQF guideline out performed the EHR algorithm in the identifying patient who qualified for the denominator (Sensitivity 0.98). The DR NQF guidelines correctly classified (sensitivity 0.98) cases for the denominator. The positive predict value (PPV) was also high (0.89) and supports the positive identification of patients for the denominator. For the numerator, the DR NQF guidelines were better suited (Specificity 1.00) for identifying patients that did not meet the numerator criteria.

Table 32. Sensitivity and specificity for denominators and numerators for all comparisons

Measure/method (n patients)	Sensitivity (95% CI)	Specificity (95% CI)	Positive predictive value (95% CI)	Negative predictive value (95% CI)
DR				
EHR algorithm versus reference standard				
Denominator (173)	0.38(0.30 – 0.47)	0.58(0.41 – 0.73)	0.76(0.64 – 0.85)	0.21(0.13 – 0.30)
Numerator (51)	0.55(0.37 – 0.71)	0.56(0.31 – 0.78)	0.69(0.48 – 0.85)	0.40(0.22 – 0.61)
NQF guideline versus reference standard				
Denominator (99)	0.98(0.91 – 0.99)	0(0 – 0.32)	0.89(0.80 – 0.94)	0(0 – 0.80)
Numerator (86)	0.05(0.01 – 0.14)	1.00(0.82 – 1.00)	1.00(0.31 – 1.00)	0.28(0.18 – 0.38)
NQF guideline versus EHR algorithm				
Denominator (99)	0.38(0.29 – 0.49)	0.50(0.03 – 0.97)	0.97(0.84 – 0.99)	0.02(0.001 -0.154)
Numerator (37)	1.00(0.19 – 1.00)	0.17(0.07 – 0.34)	0.06(0.01 – 0.23)	1.00(0.52 – 1.00)
POAG				
EHR algorithm versus reference standard				
Denominator (229)	0.74(0.66 – 0.82)	0.58(0.48 – 0.68)	0.69(0.61 – 0.77)	0.64(0.53 – 0.73)
Numerator (96)	0.98(0.91 – 0.99)	0.93(0.64 – 0.99)	0.99(0.92 – 0.99)	0.87(0.58-0.98)
NQF guideline versus reference standard				
Denominator (135)	0.94(0.89 – 0.98)	0.1(0.005 – 0.459)	0.93(0.87 – 0.97)	0.13(0.007 – 0.533)
Numerator (118)	0.98(0.92 – 0.99)	0.93(0.66 – 0.99)	0.99(0.94 – 0.99)	0.88(0.60 – 0.98)
NQF guideline versus EHR algorithm				
Denominator (135)	0.76(0.67 – 0.83)	0.13(0.007-0.533)	0.93(0.86 – 0.97)	0.03(0.002 – 0.180)
Numerator (96)	1.00(0.94 – 1.00)	1.00(0.72 – 1.00)	1.00 (0.94 – 1.00)	1.00(0.72 – 1.00)

The DR EHR algorithm was equally good at identifying patients that met and did not meet the numerator criteria. Only 38% of patients were accurately classified as "met" for the denominator. Classification of patients who "did not meet" (true negatives) the denominator was slightly higher at 58%. As expected, the NQF and EHR algorithm for POAG out performed the DR guidelines in the classification of cases that "met" and "did not meet" the numerator criteria. All four point estimates were above 80%. The POAG NQF guidelines performed (Sensitivity 94%) well in classifying cases that "met" the denominator criteria. Denominator performance for the POAG EHR algorithm was not as strong as the numerator. PPV was 69% and sensitivity was 74%.

Estimation of DR and POAG Cases Missed

The results of the assessment to estimate the number of missed DR or POAG cases are outlined in Table 33. Out of the 8217 patients, with an assigned diagnosis, 160 were randomly selected for review. A total of 273 encounters with an appointment at the study site were manually reviewed among the 160 patients. The estimated number of DR cases missed was 225 (2.7%) and 357 (4.3%) for POAG.

Table 33. Assessment of false negative DR and POAG patients

Diagnosis	Number of patients reviewed	Number of patients missed	Percent reviewed patients missed (%)
Diabetic retinopathy	160	8	5.0
Primary open-angle glaucoma		11	6.9

Qualitative Results

Knowledge of DR and POAG Measure

Some participants appeared to be unfamiliar with the term *quality measure*. On the other hand, a segment of the interviewees were aware of national clinical quality reporting programs in some capacity. When asked about their familiarity with quality measures for DR and POAG, half of the participants immediately mentioned entities that are related to the national clinical quality reporting program. These included: PQRS/PQRI (Physician Quality Reporting System/Initiatives) and Medicare.

One of the participants immediately went on to explicitly describe the NQF summary statement for the DR quality measure almost in entirety: "I think it's doing a dilated eye exam every year and documenting retinal and macula health primarily." The description included three of the four criteria of the DR measure: Dilated exam, annual exam, and documentation on the macula; documentation of the level of severity was omitted. Documentation on the anatomical structure of interest (macula) was mentioned; but the specific clinical finding *macula edema* was not mentioned. Although the response to the question about familiarity with a POAG was not as targeted due to the inclusion of unrelated clinical items, the same participant also provided a fairly good description of the POAG measure: "...for patients with open-angle glaucoma we look at the anterior chamber anatomy, we do gonioscopy, do a dilated fundus exam at least once a year, documenting the optic nerve health and anatomy as well as the surrounding vasculature." Two of the three POAG components were stated: Annual exam and documentation on the optic nerve head (same as optic disc); cup-to-disc ratio was omitted. This participant was not only aware of the national quality programs but specifically knew of and was able to pointedly describe at least one of the ophthalmology disease specific measures that were part of the national reporting program. Further questioning revealed that knowledge of the national quality reporting program and ophthalmology disease measures was acquired from organizations external to the study site. In another response, the description provided for the POAG measure was clearly identified as a PQRI metric and contained all three NQF measure criteria but was accompanied by items that were not part of the NQF POAG measure: "…And I think it's something like, that you are of course taking air pressure, vision pressure, doing a full eye exam, that you should be looking at the optic nerve at least once per year, with documentation of nerve findings, like cup-to-disc ratio."

Other responses on knowledge of a DR or POAG quality measures were provided within the context of clinical exam guidelines or requirements for clinical documentation: "Quality measures as far as what we're documenting in our exam?" In such instances respondents did not cite disease specific quality measures. The list of clinical exam guidelines or documentation requirements given included some of the components of the NQF quality measures:

> 1)"…When you see a patient with diabetes, well you wanna have an exam done within a certain amount of time after their diagnosis. So with Type I vs. Type II diabetes we like to do regular screening exams depending on the level of retinopathy, they also have very set follow-up schedules to see a patient. And the we – there's specific gradings that we're looking at, and how to grade retinopathy, what level of retinopathy, the amount, the severity of it, what type it is, whether or not there's edema…"
>
> [A subsequent question on the important clinical factors to be documented for DR garnered an answer that listed the levels of severity.]
>
> 2) "I mean we usually document what their hemoglobin A1C is in our notes and then sort of pertinent negatives on their exam or pertinent positives. So looking at the iris and seeing if there's any evidence of any neovascularization of the optic nerve…And then looking at the macula for the presence of macula edema."

Another participant gave a prompt and succinct response that enumerated most of the components of the DR measure: "…I know they want us to write whether it's non proliferative, proliferative, and that it is mild, moderate, severe and that it is high risk or not high-risk and that there is or not macula edema." This description was not presented in the form of a quality measure and does not include the temporal criterion or dilation. But the description was nearly complete and included both clinical factors of the DR measure and grading for level of severity and the status for macula edema. Other responses to the inquiry about knowledge of a POAG measure were also framed in the form documentation requirements and included mention of the cup to disc ratio and the optic nerve, among other unrelated clinical entities.

Documentation Practices for DR and POAG

Almost all of the providers used a scribe to assist with documentation during a clinical examination. The ophthalmologists guided scribes on what information should be recorded in the clinical template and reviewed and edited entries after the clinical exam. One provider "occasionally" used a scribe due to a varying schedule, which prevented the development of a "stable working relationship with a [particular] technician." All providers indicated that they documented exam findings in the fundus section of the clinical template. The fundus section contains individual textboxes for components of the eye exam. The assessment and plan section within the progress notes was also used for additional documentation, including grading, diagnosis, and notation of communication with the primary care provider. Each provider expressed the fact that clinical practice -content of exams, disease management and treatment- is based on established evidence-based guidelines.

EHR System and Documentation Training

Providers completed a general training course on how to navigate the EHR system. The introductory EHR training course did not include any instruction that was geared towards the ophthalmology clinical template. Participants indicated that they "received kind of a general [EHR system] training…before you start. But in terms of eye documentation, zero." Respondents relied upon informal guidance on documentation in the ophthalmology clinical template and sought assistance from colleagues. "…There is some formal presentation about it [documentation], but there is a lot of informal presentations as well, going to people who are familiar with the medical record." Providers also received informal training on changes to the clinical template and initiatives that impacted documentation during meetings. All respondents indicated that they drew upon previous experiences to guide their documentation practices: "I designed [EHR system] ophthalmology [clinical templates] for the [] health care system before I came here. So I was pretty familiar with it."

The actual information documented was also perceived to be universal across organizations: "But as far as what we're documenting I mean that's pretty standard wherever you were trained." and "It carries over fairly easily though, you know, it's sort of all the same elements it's just where to find them." Training on documentation can be summed up in the following statement: "We all do an [EHR] course that's just general. It's not specific to ophthalmology. And then everything else is kind of it's just on the job in practice. And so as a resident, really I started as a resident doing that." Strong interest in more targeted training and the rationale behind documentation requirements was expressed: "…Showing us exactly what we are using in the EMR and why we are using it, so I would love more of that."

Documentation to Support Quality Reporting

Providers appeared to be willing to record additional information for quality reporting providing it was important ("If it's important I am fine doing it"), they were informed of the requirement ("as long as I am aware that that needs to be done"), guidelines were clearly laid out ("I think it makes it easier if there is something that is already set in the electronic medical record, to say that we are needing this…It is nice to have it just very clearly laid out that that's what we need to do and then we do it, and click on it, and somebody can compile that data."), the need was substantiated ("If there's a convincing case presented and it was helpful") and it had the potential to positively impact clinical care ("…as long as the outcome or the goal of doing that was improved patient care or efficiency.") A second respondent echoed the benefit of customizing the clinical template to support documentation. The respondent found the addition of data fields to the clinical template for a particular reporting initiative to be "really helpful, because we could just look and make sure we were doing those things." Having "structured" guidelines to follow made compliance easy and straightforward. Furthermore, making changes to the clinical template to enable documentation of additional information was presented as a task that was feasible: "You know, especially with the electronic medical records it's fairly easy to add one extra thing or whatever extra thing it is."

Impact of DR and POAG Measures on Quality

There was consensus among all of the ophthalmologists who were interviewed that everyone adheres to the requirements outlined in the NQF summary statement for the DR and POAG measures. "I think in our clinic all of us have. You'll get 100% because all of us do that. I mean it's good but we already do that." The reference to 100% achievement is applicable to both the DR and POAG measures. Along the same lines another provider indicated that "this particular quality measure [DR] for PQRI looks like, wouldn't change anything … I do this anyway…This is one of the things that I so obviously do. It sort of, this is just a given." Other responses include: "…All of us in [study clinic] we all do this" and "I think this is recorded the vast majority of the time." A caveat was noted by one of the ophthalmologists: "Unless they don't show up, then that is another whole other issue." That is, providers have little control over whether or not a patient attends appointments as suggested. The tasks delineated in the measures are described as a fundamental part of the clinical exam process for patients with the conditions: "You know it's just the basic steps of what you do in these exams, so there's not a question as to the need. It doesn't not get done." Further discussion on the topic discloses a view that the measures assess the quality of documentation but has no effect on clinical care: "This is more just how reliably do people document this. They should but it doesn't affect patient care otherwise."

Exposure to Quality Reports

Some participants mentioned a patient satisfaction report and the federal meaningful use (MU) reporting program when asked about their participation in any initiative to improve patient care. The MU reports mentioned assessed general patient care processes tied to communication and safety: Provision of after visit summaries to patients, electronic prescribing of medications (as opposed to handwritten prescriptions), review of medication, and review of allergies. There was no mention of any disease related measures similar to the ones under study.

Credibility of Quality Reports

The trustworthiness of past reports on provider performance was called into question: "It didn't seem like they reflected my practice very well and I wasn't convinced that the data…the source and the manner of compiling the data wasn't communicated to me. It looked like it wasn't something that was convincingly accurate so…" This perception did influence the individual's use or disuse of the reports: "We all get a lot of information every day sent to us, so unless you are convinced the information is quality a lot of times you don't have time to assimilate it."

Improving Measure Definitions and Quality Reporting

A few potential means by which the NQF measures and the reporting process could be improved were gleaned from the interviews. The first is to consider additional methods of evaluating the optic nerve head for the POAG measure. The initial response from one of the participants upon reading the measure description was the following: "…so do they mean just looking at it? Or do they mean a scan of it or what is that? There's a couple of ways we can evaluate the optic nerve head." Ocular coherence tomography (OCT) was mentioned by a number of the participants as a diagnostic tool that is used in the evaluation of DR and POAG cases. The measure population could be applied against alternate patient populations for greater impact: "So this would probably be more helpful because sometimes patients sneak into our clinic with a diagnosis of diabetes and we don't specifically address presence or absence of diabetic retinopathy in those patients. So if you are going by percentage of patients with a diagnosis of diabetes then I think this would be helpful." A reduction in documentation requirements and harmonization of said requirements across entities within the domain would be beneficial. As one respondent pointed out "There are a bunch of different reasons why you document." Some of the purposes for documenting included clinical care, credentialing, board certification, and billing (financial reimbursement). Each sector was noted to have different requirements and information needs.

In summary, providers may be aware of national reporting programs through internal initiatives but are not necessarily exposed to or aware of the NQF ophthalmology quality measures. The DR measure was promptly and definitively described by one of the participants. In most cases, the clinical factors in the NQF DR and POAG measures were described in some capacity, as a best practice guideline or as a documentation or exam requirement. Clinical exam findings were consistently documented in the fundus and assessment and plan sections of the clinical template. A majority of the participants used a scribe to assist with documentation and the ophthalmologist reviewed the entries before signing off on a visit. Ophthalmologists were also open to documenting additional information that may be needed for quality reporting. Of note, some of the ophthalmologists indicated that modifying the clinical template to support additional documentation was feasible and simple.

Clinicians participated in a formal training course on general navigation of the EHR system but did not participate in training that was specific to the ophthalmology domain. All of the participants drew upon past experiences with EHR systems to inform their use of the clinical template and sought assistance on documentation from knowledgeable colleagues whenever necessary. Lastly, all participants expressed the fact that the NQF measures assess activities that are routinely conducted during clinical care and that adherence was absolute. Each interviewee expressed the view that reporting on these measures would have a negligible impact on quality.

5. Discussion

Outcome of Guidelines Implemented

Few studies have assessed eMeasure guidelines in general or its components (57,69). This is the first known study to explore the use of translated NQF electronic measure guidelines for EHR quality reporting. The results of this study underscored similarities and differences in classification between interpreted NQF guidelines for EHR implementation (EHR algorithm) and the NQF guideline and reference standard methods of computing the DR and POAG clinical quality measures. In all but one instance, accuracy in the categorization of patients for the denominator and numerator under the NQF guideline superseded EHR algorithm classifications when compared to the reference standard (Table 25 – 30). The single anomaly was the comparison of guidelines for the DR numerator. Both the EHR algorithm and the NQF guideline underperformed – the strength of the agreement with the reference standard was poor. But the almost double (55% versus 30%) DR EHR algorithm percent agreement over the NQF guideline was noteworthy. Underperformance under the NQF numerator guideline was due to the literal application of the value set for absence of macula edema. Since no liberties were taken when the value set was applied, a significant number of cases failed to meet the numerator requirement. A broader interpretation of the value set did result a higher percent agreement and kappa for the NQF numerator. Under those circumstances, the NQF numerator implementation would have outperformed the EHR algorithm.

Another key finding was considerably higher accuracy for POAG over DR in the classification of patients across both methods. Denominator and numerator classifications for the POAG measure implementations were notably more similar to the reference standard, unlike the DR measure classifications. Examination of the classification of patients that passed the EHR algorithm, NQF and reference standard implementations was correspondingly higher for POAG compared to DR. The simplicity (numerator only sought to determine occurrence of an optic disc exam) and limited need for clinical findings in the numerator requirements of the POAG measure probably contributed to the better performance. DR numerator requirements, on the other hand, were composed of clinical findings (documented presence or absence of macula edema and level of severity) and had to ascertain specific clinical content.

The proportion of patients that passed each implementation method was determined (Table 31). The proportion of patients who passed the DR measure under the NQF guideline and EHR algorithm did differ significantly from the reference standards. Conversely, the difference between the NQF guideline and the EHR algorithm and the reference standard for the POAG measure was marginal, 3% for the EHR algorithm and 1% for the NQF guideline. An important factor to consider is that comparing the proportion of patients that met the respective implementations is an improper method of assessing the accuracy of said methods. Examination of the classification of individual patients under each implementation against the reference standard (a discussed in the preceding paragraph) is a more robust means of assessing the precision of the guidelines and understanding the underlying reasons for discordance.

Disparities in Classification Among Guidelines

Detailed comparison of the three methods exhibited differences in patient classification for diagnosis, denominator and numerator. Diagnosis misclassification occurred when patients were misrepresented as either positive or negative for a diagnosis of DR or POAG. Misclassification of diagnosis stemmed from two main sources: inclusion of extraneous ICD9 codes and inaccurate assignment of ICD9 codes to encounters. The identification of cases with a DR or POAG diagnosis using the EHR algorithm was hampered by the inclusion of codes that did not reflect the diseases of interest. The issue of extraneous codes was limited to the EHR algorithm. For example, the code list for DR included ICD9 codes for diabetes mellitus (DM). Although DM is a comorbid condition and precursor to DR, ICD9 codes for DM do not necessarily equal a diagnosis of DR. Therefore, DM codes should not have been included on the list. The same issue was observed for the POAG measure under the EHR algorithm. Moreover, ICD9 codes for completely unrelated conditions were found on the EHR algorithm diagnosis code lists for DR and POAG. ICD9 codes for diabetes mellitus were present on the code list for POAG and diagnosis codes for glaucoma and other retina conditions were included on the code list for DR. Errors (deprecated codes, for example) in the value sets of NQF eMeasure guidelines have been cataloged (57) but not for EHR translated guidelines.

The second source of diagnosis misclassification was miscoding. Miscoding involved the assignment of an ICD9 code that misrepresented either the presence or absence of a diagnosis of DR or POAG. An encounter with a coded diagnosis for "Borderline glaucoma" but a written diagnosis of POAG is one such example. It must be noted that the ICD9 code data (and all data for that matter) used for this study were sourced from entries captured in the ophthalmology clinical template. And the clinical template was populated by or in the event a scribe was used, with the guidance of a physician. None of the data were sourced from an administrative or billing module within the EHR. As shown, there was a clear disconnect between the assigned ICD9 code and the written diagnosis in the clinical template for a noticeable number of encounters.

The difference in diagnosis classification was quite large between the EHR algorithm and the reference standard. Twenty percent of cases with a DR diagnosis and 44% of cases with a POAG diagnosis under the EHR algorithm were false positives based on the reference standard (Table 25 & 26). A diagnosis of diabetes mellitus was documented in the assessment and plan for all 36 false positives for DR under the EHR algorithm; ten of these patients had additional documentation that clearly stated the absence of retinopathy. POAG EHR algorithm discordance with the reference standard stemmed from the inclusion of ICD9 codes for diabetes mellitus and other unrelated conditions in the diagnosis code specifications for the EHR algorithm. Glaucoma suspect, diabetes mellitus, cataracts, pseudophakia, diabetic retinopathy, and ocular hypertension were some of the reference standard diagnoses documented for discordant pairs. In contrast, the false positives for diagnosis between the reference standard and the NQF guidelines were much smaller: 9% for DR and 7% for POAG (Table 27 & 28). Despite the use of only DR or POAG ICD9 codes for the NQF guidelines, differences were observed because the ICD9 code assigned to the cases did not match the diagnosis documented in the assessment and plan under the reference standard. Diagnosis of diabetes mellitus was documented (reference standard) for some of the cases with an NQF guideline ICD9 DR diagnosis. For POAG NQF guideline diagnosis, patients had a reference standard diagnosis such as glaucoma suspect or borderline glaucoma.

Additionally, the inappropriate assignment of ICD9 codes to encounters led to the underrepresentation of patients who had a diagnosis of DR or POAG under the NQF and EHR guidelines. Both of which employed a select list of ICD9 codes to identify patients with the diagnoses of interest. As demonstrated in the sample study (Table 33) to examine cases classified as not having the appropriate diagnosis under the NQF and EHR guidelines, there were indeed patients who would have been considered for the denominators of the CQMs if they had been assigned an ICD9 code that matched their written diagnosis of DR or POAG.

Incorrect ICD9 codes for the EHR algorithm played a role in the misclassification of cases for the DR and POAG CQM denominators, as did the requirement for two or more encounters. Neither the NQF guidelines nor the reference standard included the two or more encounters criterion. Though the EHR algorithm was an interpretation of the NQF guidelines, the NQF guidelines did not contain the aforementioned criterion. The requirement appears to have been added on to the

EHR algorithm. The reasoning behind the inclusion of this rule was unknown. One possibility is that the rule was carried over from administrative or paper-based reporting guidelines that preceded the electronic measure specifications (39,70). The required count of two or more patients played a large role in the undercounting of cases eligible for the denominator of the DR and POAG EHR algorithms. Forty-nine percent of DR (Table 25) and fourteen percent of POAG (Table 26) cases did not meet the minimum number of encounters for inclusion in their respective denominators.

Cases that would have otherwise been eligible for the POAG NQF denominator were excluded because of the procedure code requirement used to select for ophthalmology services. As a result, five percent of POAG cases were ineligible for the denominator because their encounter lacked the specified NQF procedure codes. The same requirement was present in the DR NQF denominator guideline; however, none of the DR cases were affected. Exclusions also played an important but limited role in the identification of eligible patients for both measures. Two patients were misclassified because exceptions could be applied under the reference standard by reviewing text but not for the DR NQF and EHR algorithm guidelines. It should be noted that restricting provider role to members of faculty had no effect on the denominators because the study was conducted in a clinic that was only staffed by attending physicians.

Absence of appropriate data and inadequate instructions made implementation of the numerator criteria difficult. The EHR algorithm pointed to specific data columns but numeric values or text descriptions were not provided to implement the rules. The lack of guidance was not a major issue for the POAG measure because it simply checked for whether or not an optic disc exam was performed. Verification of the occurrence of the exam was centered on the disc and cup to disc ratio components for all three methods. If records were present and contained entries on clinical findings for disc and cup to disc ratio then the criteria had been met. Specific clinical findings were not required. As would be expected, reproducibility was well within the "very good" range for all three POAG numerator comparisons: 0.878 between EHR algorithm and reference standard, 0.889 between NQF and reference standard, and 1.00 between EHR algorithm and NQF (Table 26, 28, 30). Classification of the numerators was perfect between the POAG NQF guidelines and EHR algorithm. Discrepancies were observed in the application of an exception (enucleation), missing records in the study data, and notation of improper entries under the reference standard. Access to clinic notes highlighted these differences.

Impact of Guideline Definitions

Agreement with the reference standard proved to be evasive for the DR numerator. The DR numerator criteria required the documentation of the presence or absence of macula edema and level of severity. Again, the EHR algorithm pointed to the macula columns but did not specify values for selection. Therefore, the rules were applied similarly to POAG by identifying whether or not a record was present. Regardless, implementing the macula edema criteria would have been difficult –even with specified values- because documentation was not standardized. Clinical findings were entered in text format and were neither limited by patient nor terminology. No data columns were found for level of severity. Hence, the ICD9 diagnosis coded to the encounter was substituted. As mentioned in the discussion on diagnosis, the assigned ICD9 code did not necessarily mirror the diagnosis in the assessment and plan but was the only viable source of level of severity. Indeed there were DR cases with a documented level of severity for the reference standard but an ICD9 code for DR that did not designate or mirror the written level of severity.

Though statistical significance was not achieved and the point estimate suggested poor reproducibility, the DR EHR algorithm versus reference standard numerator comparison yielded a higher kappa at 0.093 (p value 0.245, 95% CI -0.170 – 0.356) among the DR comparisons.

Agreement with the reference standard was 54.9%. The observed level of agreement was likely due to the lack of specificity in the EHR algorithm numerator criteria for DR. The lack of specificity made it much easier for patients to meet the EHR algorithm numerator criteria. Discordance (reference standard = pass and EHR algorithm = fail) with the reference standard was attributed to: ICD9 code for DR that did not reflect the documented level of severity and missing macula records from data extract. Patients positive for the EHR algorithm but negative for the reference standard were missing documentation for the presence or absence of macula edema, dilation, level of severity or other exam components.

Evaluations involving the DR NQF guidelines resulted in worse agreement. Less than a handful of patients met the DR NQF guideline numerator criteria. Having such a small number meet the numerator criteria did not bode well for comparisons with the other methods. DR NQF guideline versus reference standard numerator garnered a dismal percent agreement of 30.2% (0.2 – 0.4) and an equally low kappa of 0.026 (p value 0.143, 95% CI -0.005 – 0.057) (Table 27). There were 60 discordant pairs. The patients were flagged as "met" for the reference standard but failed the numerator under NQF guidelines. Discordance was observed because 9 lacked an ICD9 code specifying a level of severity and the appropriate values for presence or absence of macula edema. In contrast, 51 had an ICD9 code with a level of severity but did not carry the specified entries for macula edema under the NQF rules. An incomplete value set was the cause of the poor performance for DR NQF numerator. There was only one entry for this value set (Table 21), "macula edema absent" and did not include the negated values of the conditions listed in the macula edema present value set. One example is "no clinically significant macula edema," the negation of "clinically significant macula edema," was nonexistent in the value set for "macula edema absent." The single entry of "macula edema absent" could have been applied one of two ways: literally (only consider values such as "no edema" or "no macula edema," as was done in the study) or broadly (include literal application plus negated entries for “macula edema present” values).

Indeed analysis of the broad interpretation identified twenty-five additional cases (excluded from the literal interpretation) that passed the DR NQF criteria and matched the reference standard. In turn, inclusion of these cases would have had a positive impact by increasing the number of concordant pairs and kappa statistic between the NQF guideline and reference standard implementations. Omitting the negated values in the value set for the presence of macula edema was problematic because the incomplete listing suggested that there was one valid value for the absence of macula edema. At the very least, the value set for macula edema absent should have included the negated version of the values on the macula edema present listing. Including all of the negated values would alert implementers to the potential presence of additional values to consider for the absence of macula edema finding. Agreement was even lower when the NQF guideline and EHR algorithm were compared. Again, the poor point estimates originated from the low number of patients that met the DR NQF criteria.

On the whole, the NQF guidelines performed better over the EHR algorithms in the classification of cases for the measures. The single exception was the classification of cases for the DR numerator. As described, the value set for absence of macula edema was incomplete but this issue can be easily remedied by adding negated versions of the values for the presence of macula edema. The POAG measure appeared to produce numerator classifications that more closely matched the reference standard in comparison to the DR measure. But these results were primarily due to the fact that the numerator criteria for POAG was much easier to meet as it simply checked for the presence of specific records and did not require documentation of specific clinical findings.

Measure Definition and Interpretation
NQF Guidelines

Anomalies were found in the measure logic, data concepts and value sets of the NQF guidelines. Some of the inconsistencies were applicable to both DR and POAG measures, while others were measure specific. The first item noted was an incongruity between the parameters referenced in the summary statement of the DR measure and the measure logic and data concepts. Dilation was mentioned in the summary definition of the DR measure but omitted from the measure logic and accompanying data concepts. DR diagnosis, presence or absence of macula edema, level of severity and the measurement period were in both the summary definition and the detailed instructions for the measure. Dilation was glaringly absent from the instructions and was not considered during the implementation of the DR NQF guidelines. The omission of dilation was significant because, in its absence, all patients were automatically given a pass on the requirement, despite the fact that dilation is a necessity to examine the macula. Patients who did not undergo a dilated exam automatically met the criterion. Such cases fell into the category of false positives (assuming all of the other numerator criteria had been met) when compared to the reference standard. Inclusion of the dilation requirement in the NQF guidelines for DR would have reduced the false positive rate in the classification of patients for the numerator between the guideline and the reference standard.

The difference noted between the summary definition for the POAG NQF guideline and the measure details were less impactful. The summary statement indicated the patients must undergo an optic nerve head evaluation and the relevant exam entities (disc and cup to disc ratio) were listed in the measure details. But the summary definition did not explicitly state that an optic nerve head evaluation consisted of an optic disc exam and cup to disc ratio. While this is a small point, it may be a critical piece of information that allows accurate reporting of the quality measure.

The incomplete value set for “macula edema absent” under the DR NQF guidelines was a significant factor. The value set only contained the singular value: macula edema absent. The negated terms under “macula edema present” were not included on the list of terms for “macula edema absent”. “Macula edema present” value set included the following terms: advanced diabetic maculopathy, cystoid macula edema, clinically significant macula edema, diabetic macula edema, macula retinal edema, and postoperative cystoid macula edema, among others. The negated version of two of these terms was represented in the study data: no cystoid macula edema (no CME) and no clinically significant macula edema (no CSME). Both terms were used in the abbreviated format. Records with these two terms were not identified because the value set for “macula edema absent” did not include the terms. Subsequently, only three patients met the DR NQF numerator criteria under the value sets implemented for “macula edema present” and “macula edema absent” (Table 27).

A nine-fold increase, 25 more patients, with “no CSME” in the number of patients that met the DR NQF numerator would have been observed if the term “no clinically significant macula edema” had been included in the value set for “macula edema absent”. The percent of patients who met the DR NQF numerator criteria would have increased from 3.49% (N =3) to 31.40 %(N =27). The percent agreement with the reference standard would have been 85% higher: 30.2 (95% CI 0.2 - 0.4) to 55.8 (95% CI 0.5 - 0.7). And the kappa, though still not at an ideal level, would shift from poor to fair grading. Original kappa of 0.026 (p value 0.1434, 95% CI -0.005 - 0.057) would have been 0.247 (p value 0.0005, 95% CI 0.119 - 0.374). Not only would the agreement have increased between the DR NQF guideline and the reference standard, but also outperformed the DR EHR algorithm, which had a kappa of 0.093 (p value 0.677, 95% CI -0.170 – 0.356). The incomplete value set for “macula edema absent” had a resounding negative effect on the results of the DR NQF numerator. The value set should have had been comprehensive and at the very least included the negated versions of the values in the “macula edema present” value set or instructions should have included an explicit statement to that effect.

Thirdly, the NQF rules on exclusions were not exhaustive. The rules on exceptions were not implemented for either measure because there were no data elements that captured this information. However, the assessment is still relevant in the event the measure guidelines are modified. Measure guidelines assumed an exam was either completed in its entirety or not done at all. The rules did acknowledge medical or patient exceptions for not performing an exam in its entirety by allowing exceptions for partial exams. One such example was a patient who had undergone an enucleation of the right eye. As expected, only a subset of the exam components was completed. All of the components for the enucleated eye were not applicable and did not have any documentation. In addition, a review of the value set for “medical reason” did not identify any terms referring to the absence of an anatomical site. The value set for medical reasons may need to be modified to accommodate ophthalmology specific exceptions. A second example of a partial exam was one in which some of anatomical sites could not be examined because the view was obstructed. There were also instances in which an exam was attempted but documentation for all exam components indicated a “poor view.” For this particular scenario the question would be whether or not such an exam in which the structures could not be evaluated should even be considered as an exam. With these examples in mind, guidelines on exceptions may need to be augmented.

EHR Algorithms

Like the NQF guidelines, inconsistencies were found in the logic as well as the values associated with the data concepts. The diagnosis code sets for the DR and POAG EHR algorithms contained unrelated ICD9 codes, including codes for diabetes mellitus and cataracts. Extraneous codes accounted for 81% of false positives under the DR EHR algorithm and 90% under the POAG EHR algorithm. These patients were incorrectly included in the study because the ICD9 code sets for DR and POAG contained codes for diabetes mellitus. ICD9 codes for diabetes mellitus should not have been included in the EHR algorithm diagnosis code sets because the diseases of interest were DR and POAG for the respective measures and not diabetes mellitus. Though still incorrect, inclusion of diabetes code in the DR diagnosis code set may be reasonable, since the two conditions are related. But inclusion of diabetes mellitus ICD9 codes in the POAG diagnosis code set cannot be justified. The DR EHR algorithm code set also included one ICD9 code, 362, that could not be found in any ICD9 manuals online. The code does not appear to be valid and no records in the study data carried the unfounded code.

The EHR algorithm had a requirement of more than one appointment per patient, but did not specify which appointments should be used to apply the criterion. Patients may have been seen at various clinics within a single department or across the institution, in which case a decision would have to be made about which appointments should be used to fulfill the criterion. For this study, it was restricted to appointments at single clinic within the ophthalmology department, eliminating the dilemma of deciding which appointments should be used to apply the rule.

NQF Guidelines and EHR Algorithms

Neither rule unequivocally stated laterality for any of the eye exam components. Guidelines did definitively indicate whether or not presence or absence of macula edema had to be documented for just one eye or for both eyes. The same goes for documentation on disc and cup to disc ratio. Common sense would dictate that each eye had to have the proper documentation. The guidelines should be precise to avoid misinterpretation or incorrect implementation of the requirements.

Attention should be drawn to the presence of service procedure codes that were similar but not identical to values in the EHR algorithm and NQF guidelines. Some of the codes had alphanumeric values and descriptions that were quite similar to those in the guidelines. Service codes found in the study data with a likeness to values in measure details were not selected because the alphanumeric values were not an exact match. Here is a hypothetical example:

	Service procedure codes	
Study data	EHR algorithm	NQF guidelines
81004 – EYE EXAM, NEW PATIENT, COMP	81004 – EYE EXAM, NEW PATIENT, COMP	81004 – OPH SVCS EXAM; COMP, NEW PATIENT
81004.1 –VISION EXAM, COMP, NEW PATIENT	81004.1P – VISION EXAM. COMP. NEW PATIENT	

The 81004 service code was in the NQF service procedure value set as well as in the EHR algorithm. The code was also present in the study data. The descriptions are similar across all three columns and it should be safe to assume that all codes in the table refer to the same service in the absence of the alphanumeric codes. But note the difference between the top and bottom rows. The EHR algorithm code has an extra character, the letter "P," which is absent from a similar code in the "Study data" column. And the NQF guidelines are represented by a single code in the form of integer. Visits with the 81004 code would be identified from the study data for under the NQF and EHR definitions. No cases would be identified for 81004.1P because it is not present in the study data. Cases in the study data with 81004.1 would not be selected because this code is absent from both measure definitions. The reason for the difference in the format of the alphanumeric values is unclear and whether this is an isolated variation or a common occurrence across various database systems is unknown. But small inconsistencies in measure guidelines such as this can cause patients to be excluded inadvertently resulting in inaccurate measurement. For this study, the criteria were applied as presented against the study data. Steps were not taken to identify and include codes in the data that appeared to be similar to those in the criteria, as taking such liberties with the criteria would lead to differing implementations of the measure guidelines.

The final area of concern was the need for more exact instructions. Guidelines should have explicitly stated that implement would occur in two stages. Stage 1 would involve identification of the eligible encounters for the denominator and numerator. Stage 2 would entail aggregating the data at the patient level. Otherwise, patients with more than one eligible encounter would be over represented when the measure is summarized to a proportion. Secondly, instructions should clearly indicate that each eligible encounter should meet all of the numerator rules to be classified as “met” of the numerator of a measure. Measure compliance should not be achieved by meeting one criterion on a particular encounter and achieving the second criterion on another encounter belonging to the same patient. All rules must be applied against each encounter and each numerator requirement should be met for a denominator eligible encounter to be classified as compliant. Lastly, measure definitions should exclude missed appointments when computing quality measures. Only encounters during which a patient actually saw a clinician should be included.

Precise, relevant, and comprehensive guidelines are crucial to the proper and uniform execution of measure rules. Errors and extraneous requirements in the measure rules will lead to incorrect reports. And unclear rules can lead to misinterpretation and misapplication of measure rules and ultimately erroneous results. Each measure definition must be scrutinized to insure internal validity: the purpose and summary definition must be aligned with the detailed instructions and each instruction should stand on its own merit. Every detail matters to produce quality reports that are trustworthy and useful. Any inconsistencies in the guidelines would introduce bias in the computation of quality measures and shift reports away from the truth.

Data Identification

Data is at the core of any quality reporting initiative and data selection can have a significant impact on the integrity of reports. Data acquisition was therefore a vital step in the implementation process. This step was critical because it set the foundation for the data that was used to execute the guidelines. If the report did not appear to reflect the perceptions of the clinicians upon whom the reported was based or if the there was not enough knowledge about the data that was used to explain differences between perceptions and the report then buy-in from the clinicians would be a challenge. Data acquisition did not simply refer to the act of creating a SELECT query (listing column and table names) to retrieve data using Structured Procedural Language (SQL). Acquisition involved the critical process of shifting through thousands of prospective data columns to determine which columns would best serve as the source for each criterion in the guidelines. A keen understanding of the context of each criterion in the measure guidelines and the ability to translate and marry each to individual data columns in the local EHR database were key establishing a credible dataset. Defining the data to be extracted from the local database was a monumental exercise and riddled with numerous decision points.

An important part of data retrieval was understanding the underlying structure of the data captured in the EHR. Information was also needed to map identifiers in the EHR algorithm to data columns in the EHR database. With some effort data dictionaries were used to accomplish both endeavors. Fields with static content such as *date of birth* were fairly easy to process but others such as diagnosis code, which could be sourced from multiple data columns that differed in context, posed the most difficulty. An inordinate amount of time and labor was expended to map identifiers to data columns, investigate each potential column and then determine the column that was best suited for each application.

Attention to detail was mandatory to successfully implement the quality guidelines and produce credible results. Other factors that could easily be overlooked had to be in the purview: verifying the use of left joins between fact and dimensional tables and setting default values for nulls in the extracted data. Raw data (data in the same state as recorded by clinicians in the EHR user interface) as opposed to normalized data (cleaned and errors addressed) were used to conduct the study. The use of raw data was necessary to explore and identify similarities and gaps between primary data available from the EHR and the data essential to compute quality measures. Findings can shape proposals to improve data quality at the point of entry. Otherwise, deficiencies in the data would persist and resources would continue to be squandered on normalizing the data. The benefits would be two-fold: reduce inference and address any data needs at the source.

Data Availability and Format

All of the data used for the analysis were tied to data elements in the ophthalmology clinical template. Diagnosis and service procedure data were readily available in a structured format. Structured format refers to data that are recorded in discrete columns and limited to pre-defined values. The pre-defined values are tied to numeric codes and make it possible for data to be aggregated for analytical purposes like quality reporting. ICD9 and CPT codes are examples of terminology standards for diagnosis and procedures and are engrained in the health care system. These codes are not only used to support the reimbursement process but in addition are applied towards secondary purposes such as quality reporting and research.

Secondary to this prerequisite for structured data is the format of the actual data values captured via the user interface and stored in the database. It could be assumed that structured data such as ICD9 codes and CPT codes for service type would be stored in a uniform format. But it would be detrimental to make such an assumption. Structured data contain codes representing individual values of a data element. However, the format of the codes themselves is important. The codes must have a uniform format. Take ICD9 codes, which can have up to two decimal places. Values for 365 and 365.1 could be stored in a few ways: 365 versus 365.0 versus 365.00 or 365.1 versus 365.10. Unless constraints were put in place to force all values to two decimal places, the format of these ICD9 codes could vary across individual entries in a data column.

The same goes for the values in the measure definitions. These should be reviewed to ensure that values have a consistent format and that the format is identical to the data against which they will be applied. This is a crucial point because database management software literally seeks identical values when executing an exact match between two sets of values. The format of the data is monumental to obtaining comprehensive results. Checking the format of values in the data columns as well as in the measure definitions was imperative and highlighted the very issue described above. Most of the ICD9 diagnosis codes in the study data had two decimal places but a few codes only had one decimal format. The ICD9 codes in the EHR algorithms also had the same issue along with the service procedure codes in the measure definition. Fortunately the study data had a consistent format of one decimal place for the service procedure data. Differences in format in the study data may have been an artifact of data handling during extraction from the EHR database, saving the extracted data to an Excel file or accessing the data files for analysis. It is also possible that the data were actually stored in this manner in the source database. The same goes for the format of values in the EHR algorithm definitions. Regardless, of the origin, extra steps had to be taken to address the differing formats.

Unlike diagnosis and procedure data, clinical findings were not recorded using a standard format. Some of the crucial clinical concepts in the measure definitions could not be found in a structured data format. These items were tied to the numerator criteria for the DR measure: presence or absence of macula edema and the level of severity. Implementation of the DR NQF numerator criteria was hampered by the lack of structured data. Not so for the DR EHR algorithm because the criteria only assessed for the presence of a record and not specific clinical values.

On a positive note, separate textboxes were present in the clinical template for recording findings on the macula of each eye and other fundus exam components. But the textboxes permitted the entry of text (unstructured data) and did not constrain the terminology used for documentation or the patient matter of the content. Entries in the macula data columns included clinical findings that were both related and unrelated to macula edema and recorded using various terms. Furthermore, text data is not conducive to large-scale reporting because it lacks uniformity. Without the structured entries, the NQF guidelines were implemented by manually comparing the descriptions associated with the designated numeric codes for macula edema to entries in the macula data columns. This approach was taken because the dataset was limited to a few hundred patients and entries but would be inefficient as the prevailing method for routine reporting across large patient populations. The situation was dire for level of severity because there were no data entry components for this clinical entity. The visit ICD9 code was used as a surrogate level of severity in the absence of a dedicated data element.

Medical reason and patient reason for not performing an ophthalmology exam were components for which structured data were unavailable. Limited information was available in the clinical notes but was only accessible via manual review. This information is key to preventing the inclusion of patients who would otherwise be ineligible for the denominator of the measures. In one instance, documentation indicated that a DR patient had refused dilation, an exclusionary reason. Under the reference standard the information could be considered and the patient was classified accordingly for the denominator. Plus exceptions were important to absolving patients from numerator criteria that were not applicable to them. One example is a patient who had had an eye removed and consequently only had documentation on one eye. The patient would not have met the numerator criteria if the exception had not been taken into account under the reference standard. Exclusionary criteria are necessary to ensure that patient membership in the denominator is classified correctly. Failure to account for exceptions could adversely affect performance on a measure.

Ideally, information on the two clinical components for DR and exceptions should be captured in a structured format using a standardized vocabulary. Entries for each data element would be limited to a pre-approved set of terms; and each term would be associated with a specific code to enable easy retrieval and data analysis. In the absence of a structured data element, textboxes should be available for each clinical component and uniform terms used to document findings. Documentation on the macula has one of the two suggested conditions for documentation that is not structured: individual data fields to document on the macula of each eye. The singular outstanding item is the use of standardized terminology to describe the presence or absence of macula edema.

Documentation Practices

Use of Clinical Template

All of the physicians interviewed reported use of the established ophthalmology clinical template to document exam findings. The study data corroborated the reported behavior. Providers consistently recorded the various exam components using the same data elements within the clinical template. Furthermore, data elements relevant to the study were customarily populated. The exceptions were instances in which data elements for pertinent study components (macula, disc, cup to disc ratio etc.) were not completed for some encounters. Documentation indicating the reason for the incomplete data elements was recorded in a limited number of these instances. To reiterate, incomplete records were far from the norm. Most patients had data elements containing at least one word describing the exam findings.

Absence of Standardization

As mentioned previously, lack of standardization was a hindrance in the implementation of some measure rules. The absence of standardization had a significant impact on the DR measure, which required specific values for the documentation of presence or absence of macula edema and level of severity. POAG measure rules were not affected by the lack of standardization because specific values were not needed to implement these rules. POAG numerator rules only assessed for the presence of specific records. With regard to the DR numerator, macula findings were recorded in separate textboxes for each eye. No data entry elements existed for level of severity. As expressed during the interviews and validated during chart review, when documented, grading (level of severity) for DR was routinely recorded in the primary notes section (Assessment and plan) of the clinical template.

The data entry components for macula were free-text and contained a myriad of clinical findings. Entries were neither restricted by patient matter nor terminology. And recordings included findings that were both related and unrelated to determining the presence or absence of macula edema. Some of the entries included text that clearly described the status of macula edema, while others did not. To compound the problem, various descriptors were used to document the presence or absence of macula edema. Standard terminology such as “no clinically significant macula edema” and “clinically significant macula edema” was used in in some cases. The format used to record these terms also varied: the terms were either spelled out or abbreviated (e.g. no CSME, CSME, NCSME etc.). “No edema”, “no macula edema”, and “no frank macula edema,” among many others, were used in the absence of the standard terms. Alternative terms that would not be readily recognized as designators of the presence or absence of macula edema were present in the clinical notes. “Thickening” and “flat” are samples of synonyms that were used to describe the presence and absence of macula edema, respectively. Of note, the NQF value sets for presence and absence of macula edema did not include the alternate terms employed in the clinical notes. Thus leading the underreporting of these characteristics when the NQF guideline was applied.

The lack of standardized documentation was evident during manual chart review. The interview responses explained why documentation standards varied among the physicians. Absence of established guidelines appeared to be the primary reason for the diversity in documentation. Each interviewee indicated that they utilized evidence-based guidelines to inform their clinical practice.

Major research studies and professional organizations were cited as the source of clinical guidelines. But this was not true for documentation. No guidelines were cited with respect to documentation. All of the ophthalmologists cited the fact that they drew their documentation habits from past experiences and had not participated in ophthalmology specific documentation training at the study site. In the absence of formal guidelines, assistance was sought from colleagues, primarily to determine where specific information should be recorded in the template. Establishing formal guidelines for what and how to document findings would be of value to clinical practice and other initiatives. Specific clinical components could be targeted to control the scope of the guidelines. At the very least, this would harmonize terminology usage. Harmonization would reduce ambiguities in documentation because pertinent findings would be described using definitive terms. This could be a positive for clinicians as they review documentation by colleagues. The process of detecting and cataloging specific findings across patients for analytical processes would also be easier. Some ground has already been made because some of the records already exhibit standardized language.

Misclassification of Diagnoses

ICD9 code assignment is another area that could benefit from review and guidance. There is room for improvement in the congruency between the assigned ICD9 code and the documented diagnosis. The discussion primarily focuses on the NQF guidelines to assess the compatibility between assigned ICD9 codes for DR and POAG and the documented diagnosis. The EHR algorithms were not discussed because guidelines included none DR and POAG ICD9 codes, allowing for the inclusion of cases without an actual DR or POAG ICD9 code. The NQF guidelines were restricted to actual DR and POAG ICD9 codes. Some encounters had a diagnosis code for DR or POAG that was not present in the written notes (Assessment and plan). Nine percent of patients with a diagnosis of DR under the NQF guidelines, which used valid ICD9 codes, were false positives when compared to the reference standard (Table 27). Patients with a false positive had a DR ICD9 code assigned to at least one of their encounters but documentation by physicians for these encounters did not corroborate the coded diagnosis. Five of the 9 false positives had an ICD9 code of 362.04 (Mild NPDR) and four had an ICD9 code of 362.01 (BDR). All nine patients with a false positive NQF DR diagnosis had a diagnosis of diabetes mellitus (DM) listed in the Assessment and Plan under the reference standard. In addition to a diagnosis of DM, documentation for the reference standard explicitly stated the absence of retinopathy for 8 of the patients. "No retinopathy" and "without retinopathy" were used to indicate the lack of retinopathy. The percentage of false positives for POAG diagnosis under the NQF rules was 7% (Table 28). Patients with disagreeing diagnosis classifications between the NQF guidelines and the reference standard for POAG had ICD9 codes of 365.11- Open-angle glaucoma (n = 9 of 10) and 365.12- Low-tension open-angle glaucoma but a reference standard diagnosis of cataracts, glaucoma suspect, ocular hypertension, borderline glaucoma, corneal irregularity or keratoconus.

Other encounters had a documented diagnosis of DR or POAG without a corresponding ICD9 code as shown in the estimation study of patients without a DR or POAG ICD9 code (Table 33). The medical records (Assessment and plan section) of a subset of patients without an ICD9 code for either DR or POAG under the NQF or EHR rules were reviewed to determine whether or not they accurately classified as none DR or POAG cases. An estimated 3% of patients with a DR diagnosis and 4% of patients with a POAG diagnosis were excluded from the study because of a mismatch between the assigned diagnosis code and the documented diagnosis. These patients were misclassified as not having a DR or POAG diagnosis because the study diagnosis was documented in the Assessment and plan but the ICD9 assigned to the encounter did not match. ICD9 codes were the primary means by which cases were selected for inclusion in study; selection was based on the ICD9 codes outlined in the EHR algorithm and NQF guidelines. The aforementioned patients were excluded from the study because their encounters did not have the correct ICD9 code.

Incorrect assignment of ICD9 codes to encounters lead to the inadvertent inclusion of unqualified cases and exclusion of relevant cases in the study. Inaccuracies in coding may be due to system processes in which codes were carried forward from the scheduling process. Physicians may also have limited time to devote to selecting the appropriate codes or may need to be trained on which codes are appropriate. Another possibility is the fact that scribes were used to assist with documentation and may have inadvertently selected the incorrect code. Although it should be noted that physicians typically review the information documented by scribes before closing the encounter. No matter the cause, increasing the accuracy of coding is vital because it would improve the identification of disease populations for quality reporting and other endeavors such as credentialing, board certification, and research.

Dichotomy of Measure Performance

Interestingly, all physicians expressed absolute adherence to the best practice guidelines upon which the measures are based. Assessment of the measures under the reference standard, however, revealed discordance between actual documentation practices and reported behavior. Seventy-one percent of patients with a DR diagnosis identified using the NQF guidelines met the reference standard denominator criteria and numerator criteria. The percentage for DR cases under the EHR algorithm was slightly lower at 65%. Cases with a POAG diagnosis under the NQF or EHR guidelines had the same met rate of 85% for their respective numerators. Neither set of percentages matched the 100% performance reported by the physicians during the interviews.

The observed performance on the reference standard points to a disconnect between the perceived and actual behavior. Each measure assessed performance on two levels:

1. Examination of specific components
2. Documentation to demonstrate examination was conducted or reason for not performing the exam

Based on the above levels of assessment, observed performance under the reference standard could be due to the fact that the exams may not have been formed; the exams were performed but documentation was inadequate; or the exams were not performed and documentation did not mention an exception. If the first scenario were true then patients would have been accurately classified as "did not meet" for the measure. Assuming that physicians did adhere to the best practice guidelines and patients underwent the appropriate evaluations, it stands to reason that the disagreement would then be due to insufficient documentation. In other words, for DR, patients might have been evaluated for macula edema and assessed for level of severity but the information was not documented in a manner that could be easily recognized. For example, some of the macula records contained the value of "normal" or other text that did not definitively specify the presence or absence of macula edema. Or the level of severity was not specified. The latter scenario was applicable to both the DR and POAG measures. Encounters with missing or blank records for exam components without a recorded reason for the absence of documentation were categorized as "did not meet" for the numerator. Encounters without the appropriate documentation in the Assessment and plan and a documented reason would also fall into this category.

The lack of specificity and incomplete documentation contributed to the performances observed in the reference standard. Differences in documentation style were also evident among physicians. Some documented the clinical components needed for the measures, while others did not. Moreover, some physicians employed standard terminology when documenting on components that were pertinent to the measures. The disparate methods of documentation could be due to the lack of formal training at the study site or is user interface that permits variation. Some of the interviewees indicated that their style of documentation was developed during their medical training. Documentation guidelines likely differed across the various training programs and mentors, which was apparent in the lack of uniformity across records. However one physician expressed the following sentiment:

"But as far as what we're documenting I mean that's pretty standard wherever you were trained." and "It carries over fairly easily though, you know, it's sort of all the same elements it's just where to find them."

This viewpoint suggests that standards for documentation were universal. But the information in the records suggest otherwise. Disparities were present in terms of what information was documented and how the information was documented. Some were trained to document in a more standard format that facilitated quality reporting. The DR measure components offered striking examples of both issues. Undoubtedly, physicians primarily documented to support clinical care and to convey their observations to others. The principal purpose was not to support quality reporting. Guidelines may be needed be established a level of uniformity across physicians. It would be difficult to set standards for each and every clinical component. The best approach would be to focus on select clinical components that are of high value to clinical care and most relevant to reporting obligations. These standards may be desirable to improve the specificity and completeness of documentation for targeted clinical components. A majority of the interviewed ophthalmologists did not object to adding specific components to clinical documentation suggesting that physicians may accept such an endeavor. A thoughtful user interface for clinical documentation may improve compliance with the standards.

Physician awareness of measures

From the interviews we learned that, in general, physicians were unaware of the DR and POAG clinical quality measures. When asked on the outset about DR or POAG measures, most seemed unfamiliar with the term "quality measures" and cited various exam or documentation requirements. Some of the responses were on point and directly related to the components of the measures under study. However, at the end of the interview, when shown the summary definition of each measure, each ophthalmologist was familiar with the best practice guideline outlined in the NQF summary statements for the DR and POAG measures. And all participants immediately indicated that they routinely performed the tasks outlined in the statement.

The dichotomous responses, initial unfamiliarity and emphatic acknowledgement, were striking. It is possible that interviewees might have misinterpreted the initial question on awareness of measures due to unfamiliarity with the term "quality measures," leading to some of the convoluted responses. Another conceivable explanation is that physicians use a different term to refer to the concept embodied by quality measures. But despite unfamiliarity with the term, one of the participants immediately described all of the components of the DR measure, suggesting that physicians can have detailed knowledge of preferred practice patterns without familiarity with the concept of a quality measure. In a perfect world, physicians would be aware of and knowledgeable about best practice guidelines, quality measures and quality improvement. At a practical level, efforts should focus on increasing broad understanding of the symbiotic relationship among the three entities. Cognizance of best practice guidelines is important, as these guidelines should dictate clinical practice. Knowledge of best practice guidelines would thus take precedence over familiarity with the fine points of each quality measure because 1) each measure echoes the components of the guidelines it represents and 2) expecting physicians to have in-depth knowledge of quality measures would not be feasible in the current landscape of competing interests. It is far more vital for physicians to have a strong grasp of quality measures as an effective tool for improvement and the link between documentation habits and the quality of data available for reporting.

6. Recommendations

The following propositions should be considered to improve the NQF, EHR and the process of computing clinical quality measures. These steps would reduce ambiguities in the instructions (standardize implementation by limiting the need for implementers to interpret guidelines) and improve the accuracy and completeness of the guidelines. Achieving a high standard of precision and clarity in the instructions would curtail variations in the implementation of the guidelines. Uniformity in the execution of the guidelines would support level comparison of measure outcomes within and between organizations.

Implications for NQF Guidelines

Guidelines should be carefully reviewed to identify and address discrepancies. The detailed instructions and listing of measure data elements should reflect the intent and parameters outlined in the summary statement for each clinical quality measure and vice versa. Thus, ensuring that the results from implementing the guidelines encapsulate and mirror the purpose of the clinical quality measure. The requirement for service type (procedure) should be discontinued as it inadvertently excludes patients who have the disease of interest but for whatever reason do not have a procedure code that matches the guidelines. If the focus is to improve the quality of patient care, then guidelines should be inclusive (and comprehensive) rather than exclusive. Patients with the condition under consideration should only be ineligible for a measure if they have a valid exclusionary criterion and should not be excluded based on a technicality. Value sets should be examined to verify that content are complete and instructions should be explicit and leave little or no room for misinterpretation. Consideration should be given to adding vital status to the measure guidelines for patients who expired during or prior to the measurement period. Mortality could be added to the list of exclusions.

Implications for EHR Algorithms

Interpretations of the NQF guidelines must adhere to the directives in the NQF guidelines and should only be augmented with information on the location of or specific data columns from which the relevant data can be retrieved. In this vein, removal of extraneous rules, two or more encounter requirement and limitation on physician role that are absent from the NQF guidelines, should be considered. The patient pool should be all encompassing. All patients, regardless of the number of appointments or whether an attending physician or resident saw them, should qualify, if they have the conditions under assessment. The two or more encounters requirement may be a way to identify patients who routinely seek care from a particular location. But patients who had minimal clinic visits should probably be assessed as part of the measures to verify that they received proper care. The same goes for patients who were seen by a resident. Although it was not implemented because codes were not provided to apply the criterion, the order procedure condition is another factor that would unnecessarily exclude valid patients. The reasoning behind excluding cases without a specific set of orders is not known. Code sets for diagnoses proved to be problematic and would have to be amended to improve the accuracy of identifying patients with DR or POAG. Irrelevant ICD9 codes should be eliminated reduce false positive for the diagnosis of DR or POAG.

Implications for Documentation

Standardization of ophthalmology documentation, particularly for high priority diseases for clinical practice and reporting, would be an important step toward improving the availability of data for reporting. Uniform documentation practices would ease identification of clinical components and data retrieval. This would include development of formal guidelines and staff training on documentation. Documentation requirements could also be harmonized across clinical care and regulatory and other reporting responsibilities, simplifying the goal of "good documentation" for the clinician and simultaneously mitigating the burden of adhering to various documentation guidelines. Setting minimum data capture per disease to dually support clinical practice and reporting (quality and other applications) would be advantageous. Rather than attempting to set guidelines for all clinical entries, focus can be placed on clinical factors that would be of most value for specific conditions. In addition to disease-specific findings, two areas that should be addressed are: documentation of exceptions for examinations and assignment of diagnosis codes. Diagnosis codes are the most efficient means of identifying patient populations, without reviewing each an every record to determine whether or not a patient has a particular condition. Clinicians should be made aware of why diligence in documentation is so important and that improved documentation would not only benefit reporting but also improve access to quality information for patient care.

Implications for Data Acquisition

The ability to efficiently and effectively locate data columns linked to components in the EHR interface would ease the data discovery process that precedes construction of a quality report. Therefore, mapping between the user interface and the data warehouse should be comprehensive and easy to follow is paramount. This can be achieved by maintaining comprehensive (current and historical) documentation on the entire data flow process that would facilitate the identification of pertinent data columns for retrieval. Consistency in the formatting of data values within individual data columns should be emphasized. Uniform data formats would facilitate matching of pre-defined values for measure components against the data in the data warehouse.

7. Study Limitations

There are some limitations to the study. The selection of patients with a diagnosis of DR or POAG under the EHR and NQF guidelines was dependent upon ICD9 codes. Patients who did not meet the ICD9 code diagnosis requirements under either guideline were not included in the study. There were 8217 patients who fell into this category. Inclusion of these patients was beyond the scope of the study because each measure focused on a specific disease. Patients without the conditions (based on the assigned ICD9 code) under study were ineligible for the study. A limited sample of the excluded cases was evaluated to determine the presence of false negative cases for DR and POAG. False negative cases included patients with a documented diagnosis of DR or POAG without a corresponding ICD9 code on the encounter. The evaluation of 160 randomly selected patients identified five percent and seven percent false-negative cases for DR and POAG separately. If ICD9 coding had been accurate, these cases would have been part of the study. Though the use of ICD9 codes is not flawless, they provide the most efficient and a fairly effective means of identifying disease populations.

The DR and POAG EHR algorithms did not provide definitions of some data concepts and values for implementation. No information was available on two identifiers cited in the EHR algorithm for the numerators. Inquires by the analyst and reviews of the EHR data dictionary were fruitless. Therefore, the section of the algorithm associated with these identifiers was not implemented because no data columns could be located for these identifiers. It was presumed that the identifiers were related to eye exam components because they were located within the section of the algorithm that contained these components. Beyond that, no other conclusions could be drawn about the purpose of the identifiers. Information may not have been available on the identifiers because they had been deprecated from the EHR system or were never installed in the

EHR system, since the configuration of EHR systems varies across deployments. The impact of not implementing the rules associated with these identifiers is unknown.

The EHR algorithms also did not provide the values to implement the order procedure requirement. The requirement was not implemented as a result. Implementation of the order procedure requirement would likely have caused under performance of the EHR algorithms. Patients who were not positive for the order procedure requirement would have been misclassified as "did not meet" for the denominator of the EHR algorithms and resulted in discordance with classifications under NQF guidelines and reference standards. The strength of the agreements would have been lower because neither the NQF guidelines nor the reference standards included restrictions on order procedures. Application of the requirement would have unnecessarily excluded patients who would have otherwise been eligible (assuming all other criteria for denominator had been met) for the denominator of the measures.

Scribes were used to assist with documentation. Though under the guidance of the physician, documentation may not reflect the exact work of the physician. Physicians do have the opportunity to edit information documented by a scribe, as it is standard practice for the physician to review documentation before closing the encounter.

A non-clinician conducted the manual abstraction for the reference standard. Having a clinician who was knowledgeable of the ophthalmology domain might have been advantageous over an individual who had not completed formal training in the field. However, an abstraction guide was developed and the abstractor successfully completed a training exercise to verify their competence. An expert clinician clarified issues and questions encountered during abstraction. The effect of having a non-clinician complete the abstraction is probably minimal but the possibility remains that it could have impacted the accuracy of the manual abstractions.

Generalizability of the results beyond the study site should be considered. The study was conducted within an environment that was shaped by the local culture and information technology. Therefore, some of the findings may not be applicable outside of the study site. At the same time, there are discoveries that the informatics community can draw from.

8. Conclusion

The goal of the national health care initiatives was to promote the adoption and meaningful use of EHRs. The adoption phase could be considered a success as EHRs are fairly ubiquitous and can be found in many health care organizations. Achievement of the meaningful use objective, however, is still in progress. The results of this study point to that fact but also emphasize the gains that have been made and offer a glimmer of hope for the future.

The study demonstrates that there is a divergence between the data required to compute CQMs and the data captured in the EHR. This is particularly true for data concepts that embody clinical findings such as level of severity and stems from the fact that clinical findings have not been prominently featured in data acquisition initiatives. Data, primarily diagnoses and procedures, used to support the economic arm of health care have typically been at the forefront of data collection processes. As a result, ICD9 and procedure codes were accessible and in a format that enabled analysis on a large-scale. But access to clinical data in a structured format is tenuous because much emphasis had not been placed on its collection. However, the tide is changing with the focus on the secondary use of clinical data for analytical applications including quality reporting. Structured clinical findings are also difficult to acquire because collection requires purposeful planning. Purposeful planning entails identifying and prioritizing critical data concepts across obligations (patient care, board certification, quality reporting) and disease populations and instituting guidelines for the recording and storage of said data.

One of the major findings of the study was that physicians may overestimate their adherence to best practice guidelines. The discordance between perceived performance on the quality measures and the computed measure outcomes could be due to documentation habits. Interviews with physicians, manual chart reviews, and data extractions were in accord: all ophthalmologists used all components (structured and unstructured) of the clinical template to document. However, the content and terminology used to record study related components differed among ophthalmologists. Comprehensive use of the clinical template means that resources do not have to be expended to encourage use of the template. Efforts could therefore focus on key initiatives that have the potential to positively impact data quality:

- Standardization of documentation practices, particularly for high-priority diseases for clinical practice and reporting
- Formal training on documentation in the ophthalmology clinical template and navigation of ophthalmology related modules in the EHR.

Lack of standardized documentation limits the ability to capture data in EHR to compute clinical quality measures. Subsequently, accessibility of data for the computation of CQMs based on clinical findings was challenging in comparison to activity-based CQM. The establishment of uniform guidelines would reduce disparities in the documentation of pertinent clinical information and possibly improve the accuracy of reports on adherence to best practice guidelines.

When compared to the reference standard, the NQF guidelines demonstrated greater accuracy in the classification of patients in contrast to the EHR algorithm in most instances. The NQF guidelines were more inclusive, had fewer extraneous criteria, and almost all of the value sets were accurate and comprehensive. The only exception was the value set for absence of macula edema, which contributed to the underperformance of the DR numerator. The EHR algorithms, a translation of the NQF guidelines accounting for the configuration of data in a particular EHR system, contained criteria that were absent from the NQF guidelines. The added criteria negatively impacted the accuracy of EHR algorithms when compared to the NQF guidelines and the reference standard. The EHR algorithm provided important information to guide the implementation of the NQF rules locally but there is room for improvement to ensure that the content is representative of the NQF rules in their entirety. Inconsistencies and ambiguities were present in both sets of guidelines and hindered implementation of the clinical quality measures. Warranting improvements to the NQF definitions to create guidelines that fully reflect the intent of the measure, are more inclusive, and exclude fewer patients.

The goal of the CMS quality reporting initiative was to transition from a fee-for-service model for reimbursement to one that is quality-based. Reliable and accurate reports on quality are mandatory to achieve this objective. The findings from this study have important implications for quality reporting and made a meaningful contribution to the limited knowledgebase on eMeasures in its exploration of modified NQF guidelines for a commercial EHR system to compute clinical quality measures. EHRs have the potential to significantly impact quality measurement and improvement. This study underscored the drawbacks and challenges to clinical quality reporting and provides a path forward to advance the process.

References

1. Nightingale, Florence. Notes on Hospitals. London: Longman, Green, Longman Roberts, and Green; 1863.

2. HHS. What is Health Care Quality and Who Decides? [Internet]. Washington, DC: U.S. Department of Health and Human Services; 2009 Mar. Available from: http://www.hhs.gov/asl/testify/2009/03/t20090318b.html

3. Institute of Medicine. To Err Is Human: Building a Safer Health System [Internet]. Washington, DC: The National Academies Press; 2000. Available from: https://www.nap.edu/catalog/9728/to-err-is-human-building-a-safer-health-system

4. Institute of Medicine. Crossing the Quality Chasm: A New Health System for the 21st Century [Internet]. Washington, DC: The National Academies Press; 2001. Available from: https://www.nap.edu/catalog/10027/crossing-the-quality-chasm-a-new-health-system-for-the

5. McGlynn EA, Asch SM, Adams J, Keesey J, Hicks J, DeCristofaro A, et al. The quality of health care delivered to adults in the United States. N Engl J Med. 2003;348(26):2635–45.

6. Mangione-Smith R, DeCristofaro AH, Setodji CM, Keesey J, Klein DJ, Adams JL, et al. The quality of ambulatory care delivered to children in the United States. N Engl J Med. 2007;357(15):1515–23.

7. Spiegel BMR, Farid M, Van Oijen MGH, Laine L, Howden CW, Esrailian E. Adherence to best practice guidelines in dyspepsia: A survey comparing dyspepsia experts, community gastroenterologists and primary-care providers. Aliment Pharmacol Ther. 2009;29(8):871–81.

8. OECD United States Health Data at a Glance [Internet]. OECD; Available from: http://www.oecd.org/unitedstates/Health-at-a-Glance-2013-Press-Release-USA.pdf

9. OECD. OECD Health Data 2013: How Does the United States Compare [Internet]. Available from: http://www.oecd.org/unitedstates/Briefing-Note-USA-2013.pdf

10. OECD Infant Mortality [Internet]. OECD; Available from: http://www.oecd.org/els/family/CO1.1%20Infant%20mortality%20-%20updated%20081212.pdf

11. OECD United States Briefing Note [Internet]. OECD; 2013. Available from: http://www.oecd.org/unitedstates/Briefing-Note-USA-2013.pdf

12. MedPac. A Data Book: Health Care Spending and the Medicare Program [Internet]. MedPac; 2012 Jun. Available from: http://www.medpac.gov/documents/Jun12DataBookEntireReport.pdf

13. American Recovery and Reinvestment Act [Internet]. CMS; 2009. Available from: https://fdsys.gpo.gov/fdsys/pkg/BILLS-111hr1ENR/pdf/BILLS-111hr1ENR.pdf

14. Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 [Internet]. HITECH; 2009. Available from: https://www.healthit.gov/policy-researchers-implementers/health-it-legislation.

15. Medicare and Medicaid Electronic Health Records (EHR) Incentive Programs [Internet]. CMS; 2009. Available from: https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/index.html

16. Adler-Milstein J, DesRoches CM, Furukawa MF, Worzala C, Charles D, Kralovec P, et al. More than half of US hospitals have at least a basic EHR, but stage 2 criteria remain challenging for most. Health Aff (Millwood). 2014;33(9):1664–71.

17. Hsiao CJ, Hing E, Ashman J. Trends in electronic health record system use among office-based physicians: United States, 2007-2012. Natl Health Stat Rep. 2014;(75):1–18.

18. HRSA. What is meaningful use? [Internet]. HRSA; Available from: http://www.hrsa.gov/healthit/meaningfuluse/MU%20Stage1%20CQM/mu.html

19. NQF. HHS Performance Measurement: Consensus-based Entities Regarding Healthcare Performance Measurement [Internet]. Available from: http://www.qualityforum.org/About_NQF/HHS_Performance_Measurement.aspx

20. NQF. NQF Mission and Vision [Internet]. NQF; Available from: http://www.qualityforum.org/About_NQF/Mission_and_Vision.aspx

21. Donabedian A. Evaluating the quality of medical care. Milbank Q. 2005;83(4):691–729.

22. Iezzoni LI. Assessing quality using administrative data. Ann Intern Med. 1997;127(8 II SUPPL.):666–73.

23. Hsia DC, Krushat WM, Fagan AB, Tebbutt JA, Kusserow RP. Accuracy of Diagnostic Coding for Medicare Patients under the Prospective-Payment System. N Engl J Med. 1988 Feb 11;318(6):352–5.

24. Lawthers A, McCarthy E, Davis R, Peterson L, BSN S, Palmer R, et al. Identification of In-Hospital Complications From Claims Data: Is It Valid? Med Care. 2000.

25. Peabody JW, Luck J, Jain S, Bertenthal D, Glassman P. Assessing the accuracy of administrative data in health information systems. Med Care. 2004;42(11):1066–72.

26. Quan H, MD P, Parsons G, Ghali W, MD M. Validity of Procedure Codes in International Classification of Diseases, 9th revision, Clinical Modification Administrative Data. Med Care. 2004.

27. Romano P, MD M, Mark D, MD M. Bias in the Coding of Hospital Discharge Data and Its Implications for Quality Assessment. Med Care. 1994.

28. Scholle SH, Roski J, Dunn DL, Adams JL, Dugan DP, Pawlson LG, et al. Availability of Data for Measuring Physician Quality Performance. Am J Manag Care. 2009 Jan;15(1):67–72.

29. So L, Beck CA, Brien S, Kennedy J, Feasby TE, Ghali WA, et al. Chart documentation quality and its relationship to the validity of administrative data discharge records. Health Informatics J. 2010 Jun 1;16(2):101–13.

30. Simborg DW. DRG Creep. N Engl J Med. 1981 Jun 25;304(26):1602–4.

31. Iezzoni LI. Finally present on admission but needs attention. Med Care. 2007;45(4):280–2.

32. Fisher ES, Whaley FS, Krushat WM, Malenka DJ, Fleming C, Baron JA, et al. The accuracy of Medicare's hospital claims data: progress has been made, but problems remain. Am J Public Health. 1992 Feb 1;82(2):243–8.

33. Hebert PL, Geiss LS, Tierney EF, Engelgau MM, Yawn BP, McBean AM. Identifying Persons with Diabetes Using Medicare Claims Data. Am J Med Qual. 1999 Nov 1;14(6):270–7.

34. Iezzoni LI, Burnside S, Sickles L, Moskowitz MA, Sawitz E, Levine PA. Coding of acute myocardial infarction: Clinical and policy implications. Ann Intern Med. 1988 Nov 1;109(9):745–51.

35. Jollis JG, Ancukiewicz M, DeLong ER, Pryor DB, Muhlbaier LH, Mark DB. Discordance of databases designed for claims payment versus clinical information systems: Implications for outcomes research. Ann Intern Med. 1993;119(8):844–50.

36. Kennedy GT, Stern MP, Crawford MH. Miscoding of hospital discharges as acute myocardial infarction: Implications for surveillance programs aimed at elucidating trends in coronary artery disease. Am J Cardiol. 1984 Apr 1;53(8):1000–2.

37. Lix LM, Yogendran MS, Leslie WD, Shaw SY, Baumgartner R, Bowman C, et al. Using multiple data features improved the validity of osteoporosis case ascertainment from administrative databases. J Clin Epidemiol [Internet]. 2008;61. Available from: http://dx.doi.org/10.1016/j.jclinepi.2008.02.002

38. Stein BD, Bautista A, Schumock GT, Lee TA, Charbeneau JT, Lauderdale DS, et al. The validity of international classification of diseases, ninth revision, clinical modification: Diagnosis codes for identifying patients hospitalized for COPD exacerbations. Chest. 2012;141(1):87–93.

39. Keating NL, Landrum MB, Landon BE, Ayanian JZ, Borbas C, Guadagnoli E. Measuring the Quality of Diabetes Care Using Administrative Data: Is There Bias? Health Serv Res. 2003;38(6 I):1529–45.

40. MacLean CH, Louie R, Shekelle PG, Roth CP, Saliba D, Higashi T, et al. Comparison of Administrative Data and Medical Records to Measure the Quality of Medical Care Provided to Vulnerable Older Patients. Med Care [Internet]. 2006;44(2). Available from: http://journals.lww.com/lww-medicalcare/Fulltext/2006/02000/Comparison_of_Administrative_Data_and_Medical.7.aspx

41. McCarthy E, PhD M, Iezzoni L, MD Ms, Davis R, Palmer R, et al. Does Clinical Evidence Support ICD-9-CM Diagnosis Coding of Complications? Med Care. 2000.

42. Romano PS, Chan BK, Schembri ME, Rainwater JA. Can administrative data be used to compare postoperative complication rates across hospitals? Med Care. 2002;40(10):856–67.

43. Zhan C, MD P, Elixhauser A, Richards C, Jr MD M, Wang Y, et al. Identification of Hospital-Acquired Catheter-Associated Urinary Tract Infections From Medicare Claims: Sensitivity and Positive Predictive Value. Med Care. 2009.

44. Jensen RE, Chan KS, Weiner JP, Fowles JB, Neale SM. Implementing Electronic Health Record-Based Quality Measures for Developmental Screening. Pediatrics. 2009 Sep 28;124(4):e648.

45. Hazelhurst B, McBurnie MA, Mularski RA, Puro JE, Chauvie SL. Automating care quality measurement with health information technology. Am J Manag Care. 2012;18(6):313–9.

46. Miller DC, Litwin MS, Sanda MG, Montie JE, Dunn RL, Resh J, et al. Use of quality indicators to evaluate the care of patients with localized prostate carcinoma. Cancer. 2003;97(6):1428–35.

47. Kerr EA, Smith DM, Hogan MM, Krein SL, Pogach L, Hofer TP, et al. Comparing Clinical Automated, Medical Record, and Hybrid Data Sources for Diabetes Quality Measures. Jt Comm J Qual Improv. 2002 Oct;28(10):555–65.

48. Barkhuysen P, De Grauw W, Akkermans R, Donkers J, Schers H, Biermans M. Is the quality of data in an electronic medical record sufficient for assessing the quality of primary care? J Am Med Inform Assoc. 2014;21(4):692–8.

49. Mathias JS, Gossett D, Baker DW. Use of electronic health record data to evaluate overuse of cervical cancer screening. J Am Med Inform Assoc. 2012;19(E1):e96–101.

50. Weiner M, Stump TE, Callahan CM, Lewis JN, McDonald CJ. Pursuing integration of performance measures into electronic medical records: beta-adrenergic receptor antagonist medications. Qual Saf Health Care. 2005 Apr 1;14(2):99.

51. Goulet JL, Erdos J, Kancir S, Levin FL, Wright SM, Daniels SM, et al. Measuring performance directly using the Veterans Health Administration electronic medical record: A comparison with external peer review. Med Care. 2007;45(1):73–9.

52. Baker DW, Persell SD, Thompson JA, et al. AUtomated review of electronic health records to assess quality of care for outpatients with heart failure. Ann Intern Med. 2007 Feb 20;146(4):270–7.

53. Garrido T, Kumar S, Lekas J, Lindberg M, Kadiyala D, Whippy A, et al. e-Measures: Insight into the challenges and opportunities of automating publicly reported quality measures. J Am Med Inform Assoc. 2014;21(1):181–4.

54. CMS. Clinical Quality Measures [Internet]. 2014. Available from: http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/2014_ClinicalQualityMeasures.html

55. Kern LM, Malhotra S, Barron Y, Quaresimo J, Dhopeshwarkar R, Pichardo M, et al. Accuracy of electronically reported “meaningful use” clinical quality measures: a cross-sectional study. Ann Intern Med. 2013;158(2):77–83.

56. Dorr DA, Cohen AM, Williams MP, Hurdle J. From simply inaccurate to complex and inaccurate: complexity in standards-based quality measures. AMIA Annu Symp Proc AMIA Symp Symp. 2011;2011(Journal Article):331–8.

57. Winnenburg R, Bodenreider O. Issues in creating and maintaining value sets for clinical quality measures. AMIA Annu Symp ProceedingsAMIA Symp. 2012;2012(Journal Article):988–96.

58. Kallem C. Analyzing clinical quality measures for meaningful use. J AHIMA Am Health Inf Manag Assoc. 2010;81(11):56–9.

59. Metzger J. The complexity behind quality measures. Jane Metzger, principal researcher, emerging practices, CSC, reflects on a drill-down of the MU Quality Measures. Interview by Mark Hagland. Healthc Inform Bus Mag Inf Commun Syst. 2010;27(10):40–2.

60. Benin AL, Fenick A, Herrin J, Vitkauskas G, Chen J, Brandt C. How Good Are the Data? Feasible Approach to Validation of Metrics of Quality Derived From an Outpatient Electronic Health Record. Am J Med Qual. 2011 Sep 16;26(6):441–51.

61. Viera AJ, Garrett JM. Understanding interobserver agreement: The kappa statistic. Fam Med. 2005;37(5):360–3.

62. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol. 1993 May 1;46(5):423–9.

63. Barnett JM. Review: Benjamin Crabtree & William Miller (Eds.) (1999). Doing Qualitative Research (2nd edition). Forum Qual Sozialforschung Forum Qual Soc Res Vol 3 No 4 2002 Spec Issue FQS Rev II [Internet]. 2002 Nov 30; Available from: http://www.qualitative-research.net/index.php/fqs/article/view/778/1688

64. Meaningful Use with Epic Quality Measures Reporting Logic Technical Guide. 2012.

65. Diabetic Retinopathy: Documentation of Presence of Absence of Macula Edema and Level of Severity of Retinopathy (Measure) Version 3. United States Health Information Knowledgebase;

66. Primary Open Angle Glaucoma (POAG): Optic Nerve Evaluation (Measure) Version 3. United States Health Information Knowledgebase;

67. U.S. National Library of Medicine. Value Set Authority Center [Internet]. U.S. National Library of Medicine; 2014. Available from: U.S. National Library of Medicine

68. StataCorp. Stata Statistical Software: Release 13. College Station, TX: StataCorp LP; 2013.

69. Phipps M, Fahner J, Sager D, Coffing J, Maryfield B, Williams L. Validation of Stroke Meaningful Use Measures in a National EHR System. Stroke. 2015 Feb 9;46(Suppl 1):AWP275.

70. Jha AK, Perlin JB, Kizer KW, Dudley RA. Effect of the Transformation of the Veterans Affairs Health Care System on the Quality of Care. N Engl J Med. 2003 May 29;348(22):2218–27.