Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Ye Ye Date

AN EVALUATION PLAN FOR A PILOT

OF CDC EMR ALERTING SERVICE PROTOTYPE

BY

Ye Ye

Master of Science in Public Health

Department of Biostatistics and Bioinformatics

Public Health Informatics Program

Vicki Stover Hertzberg, Ph.D. Director of Public Health Informatics Program

Committee Chair

Nedra Y. Garrett, MS, Field Advisor

Committee Member

AN EVALUATION PLAN FOR A PILOT

OF CDC EMR ALERTING SERVICE PROTOTYPE

BY

Ye Ye

B.A., Peking University, 2006

M.Sc., Peking University, 2009

Thesis Committee Chair: Vicki Stover Hertzberg, Ph.D.

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Public Health Informatics
2011

Abstract

AN EVALUATION PLAN FOR A PILOT

OF CDC EMR ALERTING SERVICE PROTOTYPE

BY Ye Ye

Background: As a critical component of public health surveillance, public health authorities disseminate alerts to healthcare providers to increase their awareness of potential public health threats and to enable timely and effective responses. The variety of communication channels and the diversity of message formats, however, lead the receptors to a paradoxical situation: too much useless information, or too little relevant information. To address these challenges, the CDC EMR alerting team developed a Public Health Alert Repository System, and planned to conduct a real pilot to evaluate its performance. A comprehensive evaluation plan was needed to guarantee the completeness and validity of the test results.

Method: This evaluation plan was designed by taking account of the objectives of major stakeholders, selecting an evaluation framework, identifying evaluation elements, and clarifying measurement methods.

Results: Based on the logic model framework, stakeholders' objectives (actionable alerts; consumption of alert; integration with the decision support system; clinician action performed; sensitivity, specificity, and PPV of matching algorithm; and local customization) were mapped with five evaluation elements, including system quality; quality of the alerts (sensitivity, PPV, and specificity); clinician use of the alerts ("% of matched alerts that are clicked"); user perception (a user perception questionnaire); and impact (public health impact: local customization; health care impact: "% of positive specimen stool results" and "number of specimen stools per patient"). To test the validity of these indicators, the plan also suggested calculating their correlations, including the correlation between an objective indicator (% of matched alerts that are clicked) and a subjective indicator (the score of the "read alerts" question), and the correlations between many input, process, and outcome indicators.

Conclusion: In this study, stakeholders' objectives were successfully translated into a measurable evaluation matrix, where feasible measurement methods, study design, and data resources were identified. The causal relationships between input, process, and outcome, and the correlations between objective indicators and subjective indicators, were also recommended to be used for checking the validity of many indicators. The development process of this evaluation plan and many of its results may be possibly adapted for other system evaluations.

AN EVALUATION PLAN FOR A PILOT

OF CDC EMR ALERTING SERVICE PROTOTYPE

BY

Ye Ye

B.A., Peking University, 2006

M.Sc., Peking University, 2009

Thesis Committee Chair: Vicki Stover Hertzberg, Ph.D.

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Public Health Informatics
2011

Acknowledgements

I would especially like to thank my thesis advisor, Dr. Vicki Stover Hertzberg, for patiently reviewing my thesis and giving constructive feedback. Moreover, as the director of our academic program, she has helped me to build a foundation of knowledge in Public Health Informatics and continually conveyed a spirit of adventure in regard to research. I would like to thank my academic advisor, Dr. Tianwei Yu, for his persistent encouragement and guidance. Many thanks to Mr. Kirk Easley, Mr. Michael Lynn, Dr. Michael Haber, Dr. Robert Lyles, Dr. Qi Long, Dr. Lance Waller, Dr. Barbara Massoudi, Dr. Barry Rhodes, Mr. Mark Conde, Mr. Christopher Callahan, Ms. Melissa Sherrer, Ms. Tracy Wachholz, and other faculty, instructors, and staff of the Department of Biostatistics and Bioinformatics at the Rollins School of Public Health Emory University for always kindly providing guidance for my study and thesis writing.

I am also very grateful for the tremendous support that the CDC EMR alerting team has provided. I would like to thank Ms. Nedra Garrett, the director of CDC/OSELS/PHITPO/DIPPC, my field advisor, for providing the opportunity to immerse myself in the CDC culture and allowing me to realize the exciting possibilities of leveraging electronic health record systems for public health informatics. I would like to thank Mr. Ninad Mishra, the team leader of the CDC EMR alerting project, whose abundant knowledge helped me to make my practicum productive. I would like to thank Mr. Chuck Akin, the technology leader of the CDC EMR alerting project, for helping me to develop methods (correlation calculation) to test the validity of indicators, and showing me that the success of information system development does not only depend on sophisticated technologies, but also on how these technologies solve real world problems. I would like to thank Ms. Jessica Lee, the manager of the CDC EMR alerting project, for providing me with logistic and organizational support and a lot of helpful advice. I would like to thank Mr. Sanjeev Tandon, a core member of the CDC EMR alerting project, for persistently guiding and encouraging me. I would like to thank the other EMR alerting team members, Ms. Onnalee Gomez, Mr. Steve Gu, Mr. Raghu Jayachandran, and Mr. Melvin Crum, for providing me many guidance during my summer practicum. I would also like to thank Mr. Sundak Ganesan at the CDC, for providing me with suggestions about the mode of questionnaire dissemination.

I also really appreciate the great efforts of Mr. Thomas Fabisiak and Mr. Patrick Jamieson (Emory Writing Center tutors). They patiently helped me go through the thesis modification process.

Lastly, I dedicate this work to my whole family, especially my parents, Zhonghua Ye and Yanghua Cheng, my younger sister, Yunhan Ye, and my friend Diyang Xue. Their unconditional support keeps me going!

Table of Contents

List of Tables

List of Figures

The contents of this publication are solely the responsibility of the author and do not necessarily represent the official views of the CDC.

Chapter I Introduction

Problem statement

Public health surveillance

Public health surveillance is "the ongoing systematic collection, analysis, and interpretation of health-related data essential to the planning, implementation, and evaluation of public health practice, closely integrated with the timely dissemination of these data to those who need to know (Lee LM et al. 2010)." In the United States, the public health surveillance practices are mainly conducted by public health agencies with hiercharchical constructions, with the Centers for Disease Control and Prevention (CDC) performing most national public health duties as well as several international services, and state and local health departments taking charge of front line investigation and responses.

Communication with healthcare providers

To ensure adequate and timely public health responses to a disease outbreak or a public health event, public health agencies must try to engage and collaborate with all related partners (e.g. healthcare organizations, healthcare providers, public media, other government agencies, social groups and organizations, etc).

Because healthcare providers are on the front line of patient identification and treatment, increasing their awareness of potential public health threats that are affecting their patient population in a timely way is critically important. Public Health alerts are disseminated to healthcare providers through various channels. The CDC Emergency Communication System's Clinician Communication Team manages the Clinician Outreach Communication Activity

(COCA), which provides Email Updates and Reminders as well as Telephone Conference Calls to "a wide variety of clinicians, including: physicians, nurses, physician's assistants, pharmacists, paramedics, veterinarians, epidemiologists, public health practitioners, and state and local health department officials (about COCA, http://www.bt.cdc.gov/coca/about.asp)." In addition, the CDC also manages a strong national program, named the Health Alert Network (HAN), which disseminates "Health Alerts, Advisories, and Updates to over one million recipients (Health Alert Network, http://www2a.cdc.gov/HAN/Index.asp)." Similarly, state and local health authorities send emails to public health providers, or release messages through the media or some secure messaging systems. In addition to the variety of communication modes, the information is often disseminated in diverse formats. After reviewing the HAN archives, Garrett NY et al. (2011) have identified eighteen relevant data elements for communication: "agency creating alert; notification date; public health event name and description; public health event date range; disease agent name; characteristics of agent; clinical features of disease; area of the outbreak; number of cases; demographics of affected population; recommendations: diagnosis information, prevention information, treatment information, reporting information; special instruction information; patient education information; alert urgency and severity; and links to additional resources (Garrett NY et al. 2011)."

Electronic health record alerting for public health

Unfortunately, few public health alerts that were delivered to healthcare providers have been successfully integrated with existing clinical workflows. The variety of information pathways has lead healthcare providers to a paradoxical situation. On one hand, they always feel

overwhelmed after continuously receiving similar messages that contain the same guidelines and recommendations as an outbreak progresses (Staes CJ et al. 2011). On the other hand, they are busy seeking relevant public health information when a patient is suspected to have suffered from an infectious disease, but they don't know where they should look. Gesteland PH et al. (2008) found that only one third of the physicians accessed a state-based public health information website to know about respiratory pathogens, and this website may be more relevant for their patient population.

In the fall of 2008, CDC's "EHR alerting" stakeholder meeting agreed that public health alerting that leverages electronic health records (EHR) would be an important area, and "the development of a prototype to examine appropriate technologies, data formats, opportunities, and constraints was deemed a priority (Garrett NY et al. 2011)."

A prototype: Alert Knowledge Repository (AKR) service

A prototype, Alert Knowledge Repository (AKR) service, which integrated public health alerts into the clinical workflow was developed with the collaboration of the CDC, the Johns Hopkins University Applied Physics Laboratory Center of Excellence in Public Health Informatics, GE Healthcare, and the Regenstrief Institute (Lombardo JS et al. 2009). It was successfully presented in the Health Information and Management Systems Society (HIMSS) and the Public Health Information Network (PHIN) conferences in April and September 2009.

An updated prototype and test pilot

After the AKR, the CDC EMR alerting team began to develop an updated prototype, the Public Health Alert Repository System (PHARS). The PHARS defines the message transmission format and matching algorithms, and acts as a repository of many public health alerts. During a patient visit, his/her chief complaints (e.g., fever, or diarrhea) and demographic characteristics (i.e., age, gender, zip code) are typed into the EMR system. This information is parsed into a message that is then sent to the alert repository under a transmission standard. After that, this message is checked by the matching algorithms to find matched alerts in the PHARS. Those matched alerts are then sent back to the EMR system with an icon "alert available" as a notice in the user interface. Clinicians can click the icon to check the matched alerts. Furthermore, after they click the "alert available" icon, they can click many icons in the new user interface to read more detailed information (e.g., treatment guidance, or health education information).

Many system and function parameters of this newly developed PHARS will be evaluated within a real pilot. The team determined that a comprehensive evaluation plan was needed to guarantee the completeness and validity of the test results.

Purpose statement

This document will present the development process of the evaluation plan and its contents.

Objectives

- To restrict the evaluation scope
- To select an evaluation framework
- To identify evaluation elements
- To clarify measurement methods

Significance statement

Systematic evaluation is critically important for measuring system, identifying system deficiencies and potential challenges, and further system updating in the System Development Life Cycle (SDLC) (Bennatan EM, 2000).

It is very necessary to rigorously prepare an evaluation plan before the pilot. Starting it after the pilot is rather dangerous. With less organization ahead, evaluation scopes may be obscure; important information may be partially or completely missing; and results may have several avoidable biases; and so forth. These problems will greatly weaken the validity of the evaluation reports, while most of them are avoidable if an evaluation plan was rigorously prepared before the pilot.

This plan was developed through the system development process, and will assist the team in capturing data before and during the pilot and finalizing the system performance evaluation reports. Several related issues (evaluation scope, evaluation framework, evaluation elements, study design, related measures, and data resources) will be clarified in this document.

In addition, it may serve as a possible reference, when designing other evaluation plans to measure the performance of systems that are developed to serve as channels connecting public health information systems and electronic medical record systems.

Chapter II Review of literature

Evaluation frameworks

Many evaluation frameworks have been developed based on several domains, including technical, sociological, economic, human, and organizational domains (Yusof MM et al. 2008). These frameworks have been used to evaluate different characteristics of different information systems.

Combining different characteristics from these frameworks enables evaluators to understand comprehensive changes in several domains (Yusof MM et al. 2008). However, these frameworks ignore the causal relationships between these changes.

The "logic model" method has been developed to supplement these traditional frameworks. The Public Health Informatics Institute (2005) developed a nine-dimension logic model evaluation framework to systematically assess each possible dimension of many integrated systems (integrations of public health newborn screening laboratory information management systems and child health program information systems). It synthesized all main elements into a matrix, which included inputs, information quality, system quality, service quality, use, user perception, economic impact, organizational impact, individual impact, health impact, and health service impact. Moreover, its logical structure helped evaluators predict how the program would work when they made an evaluation plan on the initial stage, and it also assisted them in understanding why some results had not appeared as expected when they evaluated established systems.

Study designs

After evaluators choose a framework, they are ready to select an appropriate study design. There are three kinds of designs: the simple before-after evaluation, the controlled before-after evaluation, and the randomized controlled trial.

Simple before-after evaluation

The simple before-after design is the most commonly used non-experimental design for information system evaluation. In this design, the evaluation team will compare the current system performance with its previous performance. Of course, an appropriate amount of time should be left to allow the new system to be fully functional. Moreover, measuring indicators repeatedly at different time points may also help to show the reliability of those effects (Shojania KG et al. 2005).

Preliminary evidence for effectiveness could be obtained through the simple before-after design. However, its internal validity may be greatly influenced by many uncontrolled factors (Simon S et al. 2008), such as environmental factors, user characteristics, and measurement errors. These possibilities will add to evaluators' uncertainties about whether some outcomes are truly the results of this new system and whether those benefits will be easily transferred when implementing this system in other conditions. Moreover, evaluators may have different attitudes toward those uncertainties, depending on whether the evidence has validated their research hypothesis (Wyatt JC et al. 2003). Therefore, we may have to use some complementary information to increase our confidence about the accuracy of the evaluation results.

Controlled before-after evaluation

To compensate for possible significant biases that the simple before-after design will produce, evaluators could use an external control, or an internal control, or both.

To add an external control, evaluators need to locate a suitable organization, which is comparable with their organization with respect to many significant factors (Ray-Coquard I et al. 2002), such as organization size, management model, other existing information systems, and routine workflows. Then, they will need to request the external control site collect data twice, coinciding with the baseline and post implementation data collections at the site where their "intervention" (new information system or modification of an existing information system) takes place (Wyatt JC et al. 2003). After that, they will compare the changes that happened in the study site with those that happened in the external site.

However, what if the control site does not have similar influential factors to theirs as they assumed? Then, they may need an internal control within their organization, which would enable evaluators to make more confident assumptions about the similarity between non-specific factors influencing the subject studied and the internal control. Wyatt JC et al. (2003) provided an example. When evaluating the influence that an order communication system will make to the number of blood tests, the number of test orders on histopathology or bacteriology specimens could serve as internal controls, if they are not supposed to be influenced by the order communication system. If the number of blood tests falls while the number of bacteriology and histopathology orders increase, then it is strongly suggestive that it is the information system, not other non-specific factors, that is responsible for the changes.

Although both external and internal controls could greatly increase the certainty of the result, evaluators still cannot get a very reliable result. If they still want to make a definite conclusion, they should consider the gold standard design: the randomized controlled trial.

Randomized controlled trial

The randomized controlled trial is the most rigorous, least biased design.

Similar to "before-after evaluation", this design also has external or internal controls. However, its great advantage is randomization (Simon S et al. 2008). Evaluators will not have to assume that the two groups are as identical as possible in all non-specific characteristics that may influence the outcomes. They will randomly assign patients to two groups: patients whose visits are assisted with the information system, and patients whose visits are not assisted with the information system. If the sample sizes are large enough, these two groups will be comparable in all aspects, therefore diminishing evaluators' lingering doubt about the reliability of the evaluation result.

Although the randomized controlled trial is the gold standard design, it can test only specific hypotheses about selected aspects of computer systems (Berner ES, 2007). In addition, evaluators could consider their research objectives and feasibility, and choose an appropriate randomization unit. Usually, they choose patient unit (Meystre SM et al. 2008), but they could also take health care provider (Fiol GD et al. 2008), hospital department, clinic center (Goud R et al. 2009), hospital, or healthcare system as the randomization unit.

Although randomized controlled trials may increase the budget of evaluation because of their complexity, they can actually be carried out economically, if most data needed could be collected routinely (Eccles M et al. 2002).

Considerations for study design selection

Although the randomized controlled trial has the least bias and the simple before-after design has the most bias, there are still many situations where the simple before-after design is more feasible.

If a policymaker needs clear evidence that a well developed system can bring great results that deserve strong policy and funding support, a randomized controlled trial must be used, because it is the most rigorous design for determining the size of the results, even if this complicated design may cost a significant amount.

On the other hand, if evaluators want to test the system parameters of a newly developed prototype, get some feedback, and know the possible effects, a randomized controlled trial may be inappropriate. It is very risky to widely implement a prototype of which the system parameters are still unknown and may need to be modified. For this newly developed prototype, it is also hard to gather a large amount of partners and get enough funding supports to carry out the randomized controlled trial.

Considerations for sample size

Evaluators must get enough samples to be able to detect statistically significant effects. There are three factors that determine the sample size: "how much the measurement (e.g. number of tests ordered) varies between individuals (often assessed by the standard deviation)", "the minimum benefit needed", and "how accurately we need to estimate benefit, in terms of statistical significance (usually fixed at P = 0.05) and the power of the study to defect it (usually 0.8) (Wyatt et al. 2003)."

Although more samples enable a study to have a higher probability of detecting statistically significant effects, the larger study also means a much higher cost. In practice, evaluators always try to achieve a balance between the feasibility of the study and the sample size, using the minimum possible sample size that has enough power to validate a research hypothesis when it is true.

Evaluation methods

Many information systems are technologically complex and most organizational systems are even more complex because of complicated workflows, management, and human factors. Therefore, a full understanding of how and to what extent information systems assist the functionality of organization systems will be extremely challenging.

After carefully considering the system parameters, organization characteristics, and major outcomes, evaluators will have to restrict the evaluation scope, making a compromise between what is required (evaluation objectives) and what is affordable (resources). They will then need to identify a set of detailed measurement approaches.

Measurement methods can be simply classified into subjective methods and objective methods.

Subjective methods

Stakeholder and expert review

The major purpose of the information system implementation is to fulfill stakeholders' requirements (Elizabeth H et al. 2010). Therefore, it is necessary to gather their opinions about whether the information system has met their needs. Unfortunately, stakeholders' comments, especially those who have been involved in the development and implementation process, may have some bias because they really hope the system will be successful. Thus, the external expert review needs to be added to contribute to both the accuracy and the credibility of the evaluation results (Lund T et al. 2001).

Interview

Conversing with the stakeholders and clarifying their initial responses can result in better summarizations of their high-level requirements, low-level, more specific, requirements, and whether those requirements have been met (Jacobs D, 2004). Interviews provide a great opportunity to explore or clarify topics in more detail (Phillips JJ et al. 2002).

To conduct an interview, evaluators should firstly identify major stakeholders. It is better to divide interviewees into different groups, according to their responsibilities and schedules. Small groups require less planning and scheduling efforts than large workshops. Evaluators need to obtain a general understanding of the objectives of the evaluation, and develop relevant interview questions. After that, they should set meeting times and locations, and they should also provide a set of questions to interviewees prior to the interview. Providing interviewee the questions ahead of time is very helpful for the interviewee to clarify topics in more detail during

the question-answer session of the interview. After that session, evaluators should also leave enough time for interviewees to expand their opinions because the questions that evaluators developed may not be comprehensive. Finally, evaluators should summarize the interview, and get the interviewee's confirmation of the contents.

Questionnaire survey

Questionnaire surveys can be used to quickly gather information (Friedman C et al. 2005). Designers firstly develop a conceptual model, using both theoretical and empirical methods. After the designing and modifying processes, they need to achieve a compromise: they want to get relevant answers that are as detailed as possible, but responders may not be willing to spend the time that is necessary to provide them.

The scientific advisory committee of the Medical Outcomes Trust (Instrument Review Criteria, Medical Outcomes Trust Bulletin, 1995) has provided a guideline for instrument (questionnaire) evaluation. The criteria required enough rigorous experiments to test questionnaire's reliability, validity, responsiveness, and interpretability. Furthermore, they recommended that if evaluators wanted to adapt a well developed standard tool, they should also consider possible cultural and language adaptation issues for responders.

Objective methods

Organizational Profiling

Westbrook JI et al. (2004) has recommend that the profiling process could be used to "capture extant organizational and systems-wide data including budgets, staffing profiles and skill-mix, service profile, organizational structure, existing process indicators and current information technologies." The organization profiling process provides evaluators with logistic and organizational support, and enables them to study the organization and systems-wide effects a new system has brought.

Observation

Recording naturally occurring scenarios, observation enables evaluators to find many unexpected benefits and challenges a new information system has brought to an organization. For example, public health practices in real disease outbreak events are very important evidence when evaluators want to know how an automatic disease outbreak detection system performs (Buehler JW et al. 2004). Perhaps, they can find, as expected, that alarms provided by this system did notify public health officials of an unusual event in a near real-time way. On the other hand, evaluators may also find that so many unexpected false alarms have resulted in waste of resources and alert fatigue of public health official.

In addition, to capture all critical information, evaluators should complete recording as soon as possible after a critical event has occurred. Moreover, they must try to describe the situations as objectively as possible, separating their descriptions from their interpretations.

Simulation

Unlike observation, simulation has a limited ability to completely represent natural occurring events. However, simulation has a great ability to control many non-specific factors, and to study specific system performance across a set of common scenarios. Evaluators could also use multiple simulations to test system performance in different scenarios, so that they can generate characteristic curves to evaluate performance in various situations (Buehler JW et al. 2004).

Aligning study methods with research questions

There are several measurement methods that could be utilized for evaluations. None of them are comprehensive and perfect. Each has its pros and cons. Only a multi-method evaluation can provide sound, comprehensive results.

As would be expected, selecting correct methods for evaluation depends not only on what technology is being evaluated (e.g., whether it is a clinical decision supporting system or a case registration system) but also on the questions that the study is designed to answer, and how reliable the answers must be.

Studies

As the development of biomedical or health information system management science, the great importance of the system performance evaluation becomes to be realized by both stakeholders and developers. Stakeholders want to know whether and to what extent their requirements have been met. Developers hope to know system performance and user feedback to update their systems. In order to get continuing funding support from investors, developers also have to demonstrate the positive impacts of their systems.

Several studies have conducted performance evaluation of information system. Most of these studies focused on process measurement, while only a few of them evaluated outcomes. A review of controlled trials that assessed the effects of computerized clinical decision support systems (CDSSs) concluded that "the effects on patient outcomes remain understudied and, when studied, inconsistent. (Garg AX et al. 2005)" Similarly, after reviewing evaluation studies of Outpatient Computerized Physician Medication Order Entry Systems, Eslami S et al. (2007) found that only relatively small number of studies had assessed the effects on safety. With incomplete or obscure outcomes, the evaluation results of the information system are less comprehensive and convincing.

Most evaluations focused on subjective measurement. Questionnaire surveys have been widely used to capture customers' feedback for alerting systems. For example, Magnus D et al. (2002) sent questionnaires to general practices in four primary care trusts in the Nottingham area of the U.K. to find the reasons why computerized drug interaction alerts had not helped to decrease the number of prescriptions with potentially hazardous drug-drug combinations. Yu K et al. (2007) mailed questionnaires to measure Veterans Affairs (VA) prescribers' and

pharmacists' perceptions about computer-generated drug–drug interaction (DDI) alerts in order to obtain suggestions for improving DDI alerts.

Most questionnaires were completely designed by evaluators, while some questionnaires have cited a few items from a standard questionnaire to facilitate result comparison. Abernethy NF (2005) designed a questionnaire to assess users' perspective of the utility and usability of his Outbreak Investigator software. This questionnaire consisted of twenty questions geared towards contact investigation, network visualization, and data integration, as well as ten questions cited from the System Usability Scale (SUS). With the questionnaire, he measured many specific features of his software, as well as compared the SUS score of his software with other usability tests.

However, only a few studies have been conducted to develop a questionnaire to be a standard instrument. Cork RD et al. (1998) has conducted a validation study of their questionnaire "Computers in Medical Care", which aimed to measure attributes of computer use, self-reported computer knowledge, computer feature demand, and computer optimism of academic physicians, and found that this questionnaire had adequate reliability and positive validity. Statistical approaches were used for these reliability and validity analyses. Cronbach's alpha coefficient was used to compute the reliability of the questionnaire. Principal components factor analysis with orthogonal varimax rotation was used to determine "the dimensionality of each scale and degree of association of each item with the attribute of interest" (Cork RD et al. 1998). Factor analysis and correlation analysis were used to examine the construct validity of the questionnaire. In addition, there have been some considerations of scale step selection for a user perception questionnaire. Lewis JR (1993) has found that the increase in reliability tended to level off at about seven scale steps, so he recommended using seven scale steps to achieve an

appropriate balance between scale reliability and requirement put on respondents to distinguish between too many choices. Preston CC et al. (2000) also found that scales steps that were larger than ten tended to have lower test-retest reliability. According to Dawes J (2007), when using the five scale steps and the seven scale steps, the same means score was produced once the initial scores were rescaled. However, when they used the ten scale steps, they got slightly lower mean result that appeared to be biased.

Although many studies were well-designed to evaluate information systems, few of them checked the reliability and validity of the measurements. If the indicators themselves have less reliability (great measurement errors), or less validity (cannot indicate the right aspect), then how convincing will the evaluation results be?

Chapter III Methodology

Involving stakeholders

The development and implementation of the EMR alerting system is a rigorous, multidisciplinary process that involves a range of stakeholders, including the "Actionable public health alerts for Electronic Medical Record (EMR) systems" project team at the CDC, the Alliance of Chicago Community Health Services, General Electric Healthcare, the Chicago Department of Public Health (CDPH), and any other people and organizations who have a vested interest in the EMR alerting system.

Each stakeholder will have interests in areas specific to its mission. Therefore, stakeholders' high level cooperation is critical for project success. Evaluation of this system will be most successful when stakeholders are involved in all phases: identifying stakeholders' objectives about the project, deciding the evaluation matrix, identifying evaluation methods and tools, giving feedback for the evaluation results, and using the feedback for their performance improvement.

Methods to define the scope of the evaluation

The scope of the evaluation for the pilot is determined by stakeholders' objectives, available resources, the project schedule, and so forth. A comprehensive evaluation should be able to measure whether the newly developed prototype has met all core stakeholders' requirements. To define the final scope of those evaluations, I will carefully consider the feasibility of measuring related evaluation attributes. A balance of the priorities of the various

interest groups should be achieved, and the most important and feasible aspects should be carefully defined.

After communication with core stakeholders, a range of evaluation elements and measurement methods will be identified, authorized by decision makers, and then fixed.

Choosing an evaluation framework

The “nine-dimension logic model framework” (Public Health Informatics Institute, 2005) will be used because of its ability to summarize the program's overall mechanism of change by linking processes (e.g., providing relevant alerts to clinicians) to eventual effects (e.g., assisting clinicians to identify potential cases in a public health event).

Furthermore, our system will be a complex architecture that is built upon the collaboration of different jurisdictions. This framework will help us to define not only effects on entities, but also effects on their complex relationships, highlighting how integration changes the organization and how the organization changes the integration.

Selecting a study design

For this evaluation, both the randomized controlled trial design and the controlled before-after design may require more resources and collaborations than the simple before-after design.

This evaluation did not use the randomized controlled trial design because:

(1) The prototype was newly developed and its system parameters still needed to be tested.

(2) The randomized controlled trial may require more partners and funding support, which are rarely available at this stage. If we chose each patient as a randomization unit, the randomized controlled trial design would require us to install two systems in each participating clinician's computer as well as to get patients' consent for the randomization. For all participating clinicians, their patients would be randomly divided into two groups: patients whose visits are assisted with the EMR system that has been integrated with the public health alerting service (study group), and patients whose visits are assisted with the former EMR system (control group). Every participating clinician would use different systems according to the randomization number of each patient. If we chose another randomization unit (e.g., clinician, hospital department, or hospital), the randomized controlled trial design might require a much larger sample size to guarantee enough power to detect statistically significant effects, because we might have to consider the similarities of patient situations within each unit.

This evaluation did not use the controlled before-after trial design because:

(1) During the initial development and testing stage, it is not very easy to find other partners (i.e., hospitals) to serve as external controls.

(2) We are still not clear about what can be selected as appropriate internal controls. We just anticipated participating clinicians' possible actions when they use EMR alerting service. The actual effects still needed to be studied in this evaluation and the results

may provide some indications for us to identify possible internal controls for evaluation in the next SDLC.

Therefore, considering our study purpose and feasibility, both the randomized trial design and the controlled before-after design will not be used, and the simple before-after design will be the mainstay of our evaluation. To compensate for possible defects of the simple before-after design, we will identify approaches to check the validity of results.

Using both subjective and objective methods

Because each evaluation method has its own advantages and limitations, a systematic review combining all the results will help us to find sound evidence of project success or failure.

Therefore, both subjective and objective methods will be used. Moreover, the correlations between the objective indicators and subjective indicators, as well as the correlations between input, process, and outcome indicators will be calculated. Their significant correlations may indicate the validity of these indicators.

Chapter IV Results

Objectives for EMR alerting service

Major stakeholders (the CDC, Alliance of Chicago Community Health Services, General Electric Healthcare, and Chicago Department of Public Health) have helped to define objectives of the EMR alerting system. After considering the priority of requirements and feasibility of measurement, they identified objectives of the pilot.

CDC objectives

To develop a prototype that can return specific, actionable public health alerts that can be consumed by an existing EMR system.

1) Specific. To avoid alert fatigue and information overload, alerts that are provided to clinicians should be relevant to their patients. For example, it would be inappropriate to provide an infant Rotavirus Gastroenteritis outbreak alert when an adult has diarrhea symptoms. Another example is: if a young man has watery diarrhea and eye swelling, a food-borne disease outbreak alert (Trichinosis) may help to remind clinicians to ask where the patient ate and check whether he was a potential case of the outbreak.
2) Actionable. The specific public health alerts can lead to positive changes in the follow-up actions of a healthcare provider or entity. These actions may include process improvement (e.g., clinicians educating patients based on guideline information of the alerts), and outcome improvement (e.g., clinician identifying high risk populations (the increase of percentage of positive lab results).

3) Consumption. The EMR alerting service must be coherently integrated into existing clinical workflow because the system is designed to assist the clinician decision making instead of interrupting their routine practices.

Objectives of Alliance of Chicago Community Health Services and General Electric Healthcare

1) Process change. GE and Alliance are interested in clinicians' specific actions after they are provided with alerts. This objective aligns with the CDC "actionable" objective.
2) Clinician perception. GE and Alliance want to understand whether clinicians are satisfied with the EMR alerting service. Additionally, they are interested in clinicians' perceptions about benefits and loads it brings. Some subjective approaches, such as interview and questionnaire survey, may be used to gather clinicians' feedback.
3) Outcome change. This objective also aligns with the CDC "actionable" objective. GE and Alliance hope to get some summarization reports, which compare the health outcomes of patients whose visits have been provided with the EMR alerting service with whose visits have not been provided with the service. For example, will it have any impact on the average number of specimen stool orders per patient, as well as the results of these orders?

Chicago Department of Public Health objectives

1) Local customization. Chicago Department of Public Health wants to know the usage of localized public health alerts. Compare to national wide alerts (HAN network), localized public health alerts were assumed to be more relevant and informative for local populations and communities.
2) Expedited timeline of alert dissemination. The timeline of the alert dissemination by the automatic EMR alerting service (with high specificity) will be much shorter than routine alert dissemination approaches, e.g. email, telephone conference, or mail.

Considering all major stakeholders' interest areas, the evaluation will mainly focus on seven objectives: actionable alerts; consumption of alert; integration with decision support system; clinician action performed; sensitivity, specificity, and PPV of matching algorithm; and local customization.

Evaluation elements

These seven objectives were mapped with five evaluation elements, including system quality, quality of alerts, clinician use of the alerts, user perception, and impact, which were identified based on the logic model framework (Figure1).

Logically, these domains are in a certain order (Figure2). If the system has high quality and provides relevant alerts and guidelines, then healthcare providers will be more likely to use the EMR alerting services and feel satisfied with them. Timely disseminating relevant public

health alerts to healthcare providers to assist their clinical decision making will bring positive impacts to both public health agencies and health care authorities. When reporting the evaluation results, we can assess the causal relationship between technologies functions, alert usage, and public health and health care outcomes.

It is important to show that this system performs well – providing relevant alerts and being accepted by healthcare providers. However, it is more critical to check whether this system could bring positive impacts on both public health information dissemination and health care quality. That's because we not only want to implement a high quality system, but also hope to realize "meaningful use" of our system.

For each evaluation domain, I will provide many detailed information to guide the evaluation implementation, including what to measure (content), how to measure (design, methods, and indicators), how to collect data (data resource), and how to analyze the findings.

Figure 1: Mapping stakeholders' system objectives with evaluation elements

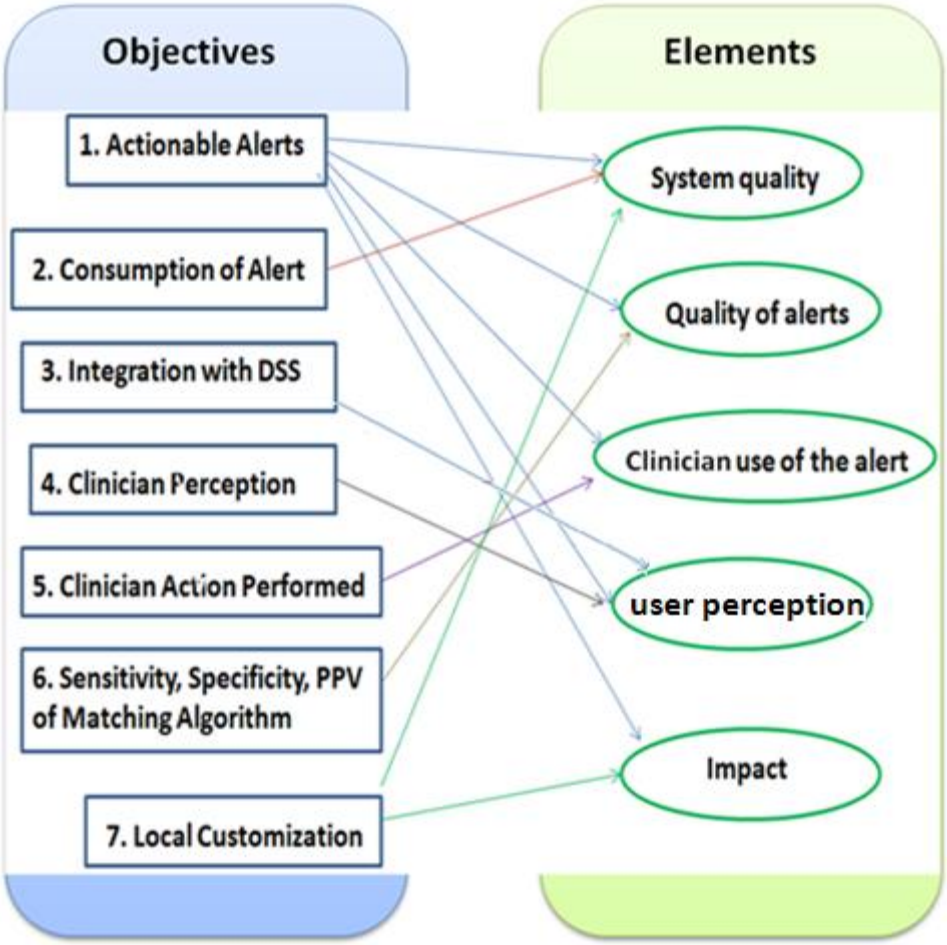
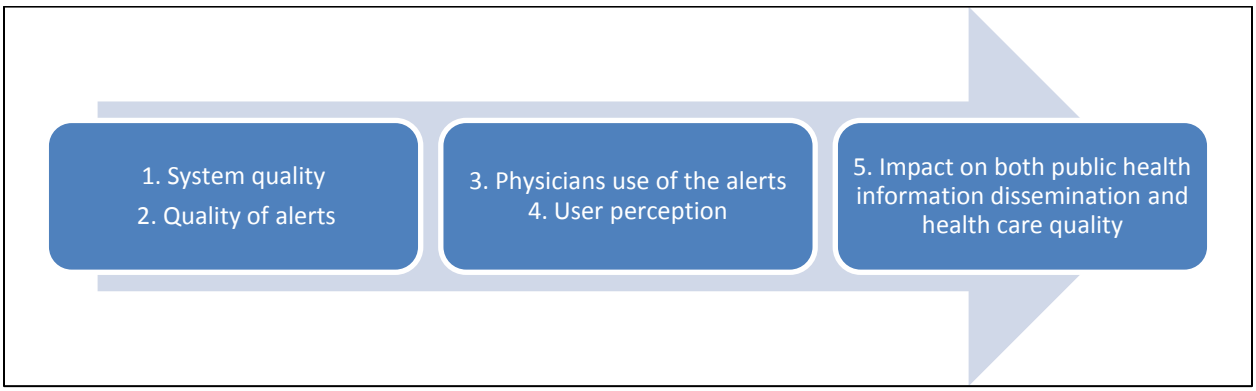


Figure 2: Logic order of evaluation elements



System quality

Description of both technological and functional dimensions

The evaluation report will provide detailed description of both technological and functional dimensions of the system, including alert repository (PHARS), matching algorithms, integration with EMR system, user interface, as well as use case.

Mock-up data scenarios for algorithm testing

Mock-up data scenarios have been used to test the technical capability of alert repository and matching algorithm during the system developing process. Test data includes demographic information (patient zip, facility zip, gender, and age) and chief complaints (their words were completely covered by the "Vocabulary for Food borne disease" spreadsheet) to evaluate the implementation of matching algorithm. Based on the results, timely modification of matching algorithms will be conducted.

System performance recording in the pilot

System performance (e.g., task, time required, and errors) will also be recorded during several real scenarios in the pilot. This will help us to evaluate integration with EMR system, user interface, as well as capture those features that might have not been studied during the mock-up use case session. A detailed, step-by-step description of how exactly the system processes the alert (from reception of alert to logging of clinician response) will be required for each scenario.

Quality of alerts

Performance of matching algorithm is one of the most important characteristics. The Gold Standard Method will be used to calculate three indicators to measure how the matching algorithm works (Table 1) (Gordis L, 2008). Although the matching algorithm is a two-step match, I will evaluate its overall performance, as it will function as a process. Expert review will be used as the gold standard. If there are several experts, we will calculate the Kappa statistics (Rosner B, 2005) to demonstrate the high agreement of experts.

Table 1. Gold Standard Method

		Gold Standard Method		
		positive	***negative***	
Another classification approach	***positive***	True Positive (TP)	False Positive (FP)	→ Positive predictive value = TP/(TP+FP)
	negative	False Negative (FN)	True Negative (TN)	
		↓ Sensitivity = TP/(TP+FN)	↓ Specificity =TN/ (FP+TN)	

As described in Table 2, with the gold standard (expert review), patients will be classified into two groups: those are potential cases of public health events and those are not potential cases of public health events. With the matching algorithm, patients will be classified into another two groups: those whose electronic medical records will be returned with alerts from the PHARS and those whose electronic medical records will not be returned with alerts from the PHARS. Three

indicators will be calculated from those four cells that will describe the extent of the overlap of the results of the two classification approaches.

Table 2. Gold Standard Method to test matching algorithm

		gold standard (expert review)		
		relate to public health events	**not relate to public health events**	
results of matching algorithm	**return with alerts from PHARS**	# of patients who "relate to public health events" and "with alerts" (TP)	# of patients who do not "relate to public health events" and "with alerts" (FP)	***# of patients whose EMR queries are returned with alerts from PHARS***
	not return with alerts from PHARS	# of patients who "relate to public health events" and "without alerts" (FN)	# of patients who do not "relate to public health events" and "without alerts" (TN)	***# of patients whose EMR queries are not returned with alerts from PHARS***
		# of patients who are potential cases of public health events	***# of patients who are not potential cases of public health events***	***total # of patients***

`

Sensitivity (recall)

Sensitivity (recall) = TP / (TP + FN) = # of patients who "relate to public health events" and "with alerts" / # of patients who are potential cases of public health events. It is the

probability that PHARS will return alerts when a patient is very likely to relate to a public health event. It indicates the system's capability to provide necessary alerts.

Specificity

Specificity = TN / (FP + TN) = # of patients who do not "relate to public health events" and "without alerts" / # of patients who are not potential cases of public health events. It measures the power to not provide alerts when alerts are unnecessary (to avoid alert fatigue).

PPV (precision)

PPV (precision) = TP / (TP + FP) = # of patients who "relate to public health events" and "with alerts" / # of patients whose EMR queries are returned with alerts from PHARS. This measures the relevance of the alerts. However, PPV will also depend on the disease prevalence (Altman DG et al. 1994). Even if the classification approach has high sensitivity and specificity, it will have low PPV if the disease prevalence is low.

$$PPV = \frac{sensitivity \times prevalence}{sensitivity \times prevalence + \ (1 - specificity)(1 - prevalence)}$$
$$= \frac{sensitivity}{sensitivity + \ (1 - specificity)(1/prevalence - 1)}$$

Clinician use of the alerts

Clinician use of alerts is a critically important evaluation element because it is a pathway that connects system quality and information quality with impact. Thus, we hope to measure whether and how the clinicians use the alerts.

"% of matched alerts that are checked" is an objective indicator for measuring how often clinicians read the alert, which is the necessary pathway for the following action changes. It could also indicate the quality of information to some extent. If only ten percent of matched alerts are reviewed, it is very likely that the matched alerts are really not relevant, or not informative.

$$\%\ \textbf{of matched alerts that are checked} = \frac{\#\ \boldsymbol{of}\ \textbf{matched alerts whoe message icons are clicked by the clinicians}}{\#\ \boldsymbol{of\ matched\ alerts}}$$

Of course, this indicator may be much higher during the pilot than during a normal situation because the clinicians who have agreed to participate in the pilot will be much more likely to actively click the "alert available" button.

Therefore, we have considered the possibility to use the "average duration of alert window opened" as another indicator, which may be more meaningful. An alert that has been reviewed for 1 minute may be much relevant than an alert whose information window is opened for just 2 seconds. However, capturing these parameters is still challenging.

Although only one objective indicator of clinician usage will be calculated, many subjective items could help to supplement the measurement of clinician use. A user questionnaire

has been designed. It has four items that specially focus on usage measure: (1) read alerts: “How often did you read the alerts? (never, rarely, sometimes, frequently)”; (2) view further information: “How often did you view extended information? (never, rarely, sometimes, frequently)”; (3) request further testing: “How often did an alert motivate you to order a specimen stool? (never, rarely, sometimes, frequently)”; (4) educate patients: “How often did an alert add value for patient education? (never, rarely, sometimes, frequently)”. Its development process and more detailed information will be described in the “user perception” section.

User perception

Questionnaire design

To measure the clinicians' perception of the EMR alerting service, a user perception questionnaire was created. It will be an important way to measure whether and to what extent the EMR alerting service prototype meet stakeholders' objectives.

The whole EMR alerting team, including team leader, technical leader, project manager, and system developer, were engaged to create and modify the content and expression of the questionnaire. In addition, suggestions and comments of persons who have experience in questionnaire design, as well as some clinicians who may have similar medical backgrounds as our intended users were received.

The design started with detailed questions or statements, measuring clinicians' perception of each project objective. Nineteen questions were firstly created. The modification process helped to update the questionnaire to be more meaningful for evaluation objectives and more relevant to potential responders. Many questions or statements were deleted because it may not relate to clinicians, such as "Alerts have helped me identify public health events" and "The alert system disseminated public health information more efficiently than regular approaches". Or, they were deleted because respondents may not be the best reviewer of the characteristic of the system and some more objective measures may be better, such as "How often was an alert relevant" and "Alerts have helped to improve health care quality". Moreover, the respondents' perceptions about the statements or questions should be measurable or answerable. For example, "How many minutes did you spend on checking the alert" may be difficult to answer. In addition, many questions or statements were modified because they may not be appropriate. For

example, “I find the public health alerts that EMR system provided were helpful,” is not appropriate because the EMR alerting system will be integrated with the clinical decision supporting system, and what a clinician will view will be an “alerts available” icon and several alerts instead of a new EMR alerting system. In this case, the statement that “I find the public health alerts that EMR system provided were helpful,” will be much better as it describes exactly the situation.

The final version (table 3) consists of two parts: (1) nine fixed-choice questions; (2) one open response question. Nine questions will be used to measure clinicians' perception in six aspects: overall usability, integration with workflow, relevance, informativeness, actionability, and impact. To guarantee a balance between questionnaire reliability and responders’ load, users’ perception will be measured with five scale steps. Five options, “strongly disagree, disagree, neutral, agree, strongly agree”, will be used to capture the extent of clinicians’ attitude. Another five options, “never, rarely, sometimes, frequently, often”, will be used to measure frequency. (In the old version, these options were supplemented with free text comments areas. To be concise, they are replaced by one open response area that is located at the bottom of the questionnaire.)

The user perception questionnaire was geared toward the evaluation of the project objectives. It included many general questions as well as some direct, specific questions. For example, “I find the public health alerts that the EMR system provided were helpful (strongly disagree, disagree, neutral, agree, strongly agree)” will capture users’ perceptions about the overall usability of the EMR alerting service. In the “actionable” section, four specific questions were respectively developed to measure four topics – “read alerts”, “view further information (click links)”, “request further testings”, and “educate patients”. The specific responses will be important evidence for measuring the extent of usage to which the EMR alerting service lead.

Quantitative methods, including a frequency table and a bar chart, will be used to analyze clinicians' responses about the first part. Qualitative analysis will be used to summarize and organize the free text response of the open response area.

Questionnaire dissemination

A free web tool was utilized to build a web version to facilitate questionnaire dissemination and response gathering. A link to the questionnaire will be sent to participating clinicians soon after the pilot.

Table 3. User Perception Questionnaire

A. Fixed-choice questions			
This questionnaire has nine questions, which will be used to measure clinicians' perception in six aspects: overall usability, integration with workflow, relevance, informativeness, actionablility, and impact. The answer options are five point scale (range from one to five).			
Category	**Topics**	**Questions**	**Answer Options (five point scale)**
Overall Usability	overall usability	I find the public health alerts that the EMR system provided were helpful.	strongly disagree, disagree, neutral, agree, strongly agree
Integration with Workflow	integrate with existing clinical workflow	The alerts provided by the EMR system didn't significantly hinder existing clinical workflow.	strongly disagree, disagree, neutral, agree, strongly agree
Relevance	relevance of the alerts	How often was an alert relevant?	never, rarely, sometimes, frequently, often
Informativeness	information quality of alerts	The alerts are informative.	strongly disagree, disagree, neutral, agree, strongly agree
Actionablility (4 questions)	read alerts	How often did you read the alerts?	never, rarely, sometimes, frequently, often
	view further	How often did you view extended information?	never, rarely, sometimes,

	information (click links)		frequently, often
	request further testing	How often did an alert motivate you to order a specimen stool?	never, rarely, sometimes, frequently, often
	educate patients	How often did an alert add value for patient education?	never, rarely, sometimes, frequently, often
Impact	decision-making capability	Alerts have added value for decision making.	strongly disagree, disagree, neutral, agree, strongly agree

B. Open Response Area

Please provide some comments and suggestions for the EMR alerting service.

_____

Impact

A. Impact on public health

Timeline shorten of public health information dissemination

The evaluation report will describe the routine procedure to disseminate public health information, and discuss the timeline shorten that results from leveraging the EMR system for automatic public health alert dissemination.

Local customization

When being entered into the PHARS, the public health alerts are tagged with the information source, including national alerts (from COCA or national HAN) and local alerts (from Chicago HAN or local Chicago DOH websites). We assume that local customization can improve the relevance of the alerts:

(1) Local alerts were assumed to be more likely to match patient information than national alerts. Statistical method will be used to test whether "% of matched local alerts" is different from "% of matched national alerts".

$$\%\ \textbf{of matched local alerts} = \frac{\textbf{No. of matched local alerts}}{\textbf{No. of local alerts}} \times \textbf{100}\%$$

$$\%\ \textbf{of matched national alerts} = \frac{\textbf{No. of matched national alerts}}{\textbf{No. of national alerts}} \times \textbf{100}\%$$

(2) Matched local alerts were assumed to be more likely to be checked by clinicians than matched national alerts. Statistical method will be used to test whether "% of matched local alerts that are checked" is different from "% of matched national alerts that are checked".

$$\%\ \textbf{of matched local alerts that are checked} = \frac{\textbf{No. of matched local alerts that are checked}}{\textbf{No. of matched local alerts}} \times \mathbf{100}\%$$

$$\%\ \textbf{of matched national alerts that are checked} = \frac{\textbf{No. of matched national alerts that are checked}}{\textbf{No. of matched national alerts}} \times \mathbf{100}\%$$

Contingency-Table approach will be used, if no expected value of four cells in the data table is less than 5 (Rosner B, 2006). Otherwise, Fisher's exact test will be used to calculate the exact level of significance (Rosner B, 2006).

B. Impact on health care

Subjective measure

The user perception questionnaire has asked clinicians' opinions on whether and to what extent alerts have added value for their decision making.

Objective measure

We hope that the EMR alerting service can help clinicians to target high risk population, and therefore help to improve health care quality.

There are two indicators:

(1) "% of positive specimen stool results" will be a major health care quality indicator of this evaluation. We assume that it will increase if clinicians can identify high risk population with the help of EMR alerting service. The measure will be compared with historical data. Fisher's Exact Test will be used for statistical significant difference testing.

$$\%\ \textbf{of positive specimen stool results} = \frac{\textbf{No. of specimen stools with positive results}}{\textbf{No. of specimen stools}} \times \mathbf{100}\%$$

(Note: the study population will only include patients of participating clinicians during the pilot period)

In addition, it is necessary to guarantee that we will get enough samples to test this change. Power Analysis and Sample Size Software was utilized to estimate sample size. Table 4 indicates that there are two factors that determine the sample size - the number of years and the effect size. If we compare the specimen stool results with those in a greater number of years, a smaller sample size may be needed in the pilot. If the effect size is larger (i.e., the proportion of positive specimen stool results increase more), a smaller sample size will be needed in the pilot.

Figure3 indicated that sample size will dramatically decrease if the proportion for positive specimen stool results during pilot is 7% or more, a 2% or larger increase compared to previous year(s).

(2) "Number of specimen stools per patient" will be another measure. We will measure whether it has significantly changed, and in what direction it changed. It is not a convincing indicator for health care quality improvement because neither the increase of unnecessary lab orders nor the decrease of necessary lab orders is meaningful. Fisher's Exact Test will be used for statistical testing.

$$\textbf{No. of specimen stools per patient} = \frac{\textbf{No. of specimen stools}}{\textbf{No. of patients}} \times \textbf{100}\%$$

(Note: the study population will only include patients of participating clinicians during the pilot period)

Table 4. Sample size estimations for different comparison strategies

comparison strategy	the proportion of positive specimen stools during pilot (%)	historical proportion of positive specimen stools (%)	# of specimen stools during the pilot	# of specimen stools in previous year(s)
Compared with previous year	6	5	11120	11120
	7	5	3061	3061
	8	5	1483	1483
	9	5	903	903
	10	5	621	621
Compared with previous two years	6	5	8289	16578
	7	5	2270	4540
	8	5	1095	2190
	9	5	664	1328
	10	5	455	910
Compared with previous three years	6	5	7345	22035
	7	5	2005	6015
	8	5	965	2895
	9	5	583	1749
	10	5	399	1197

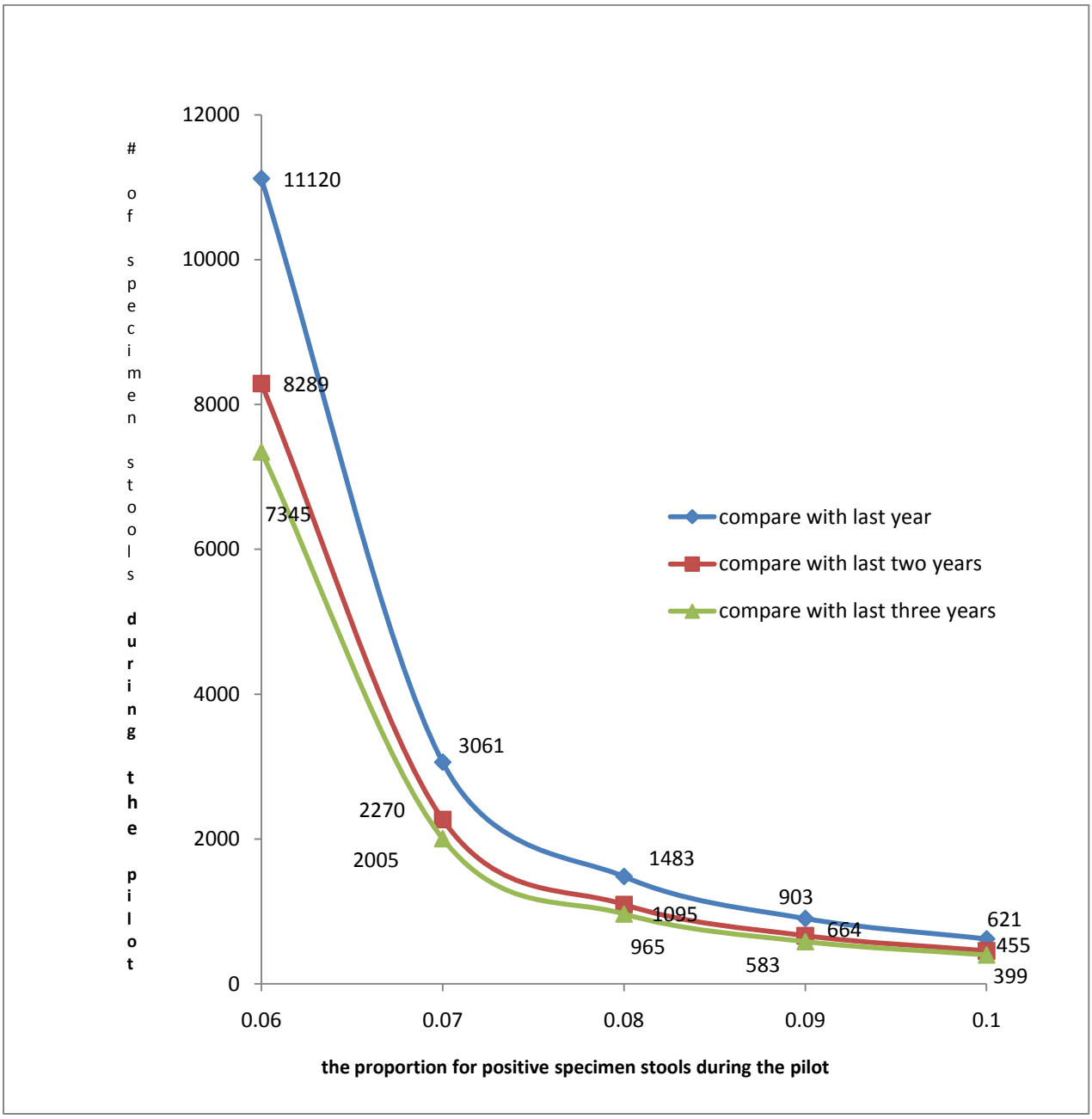


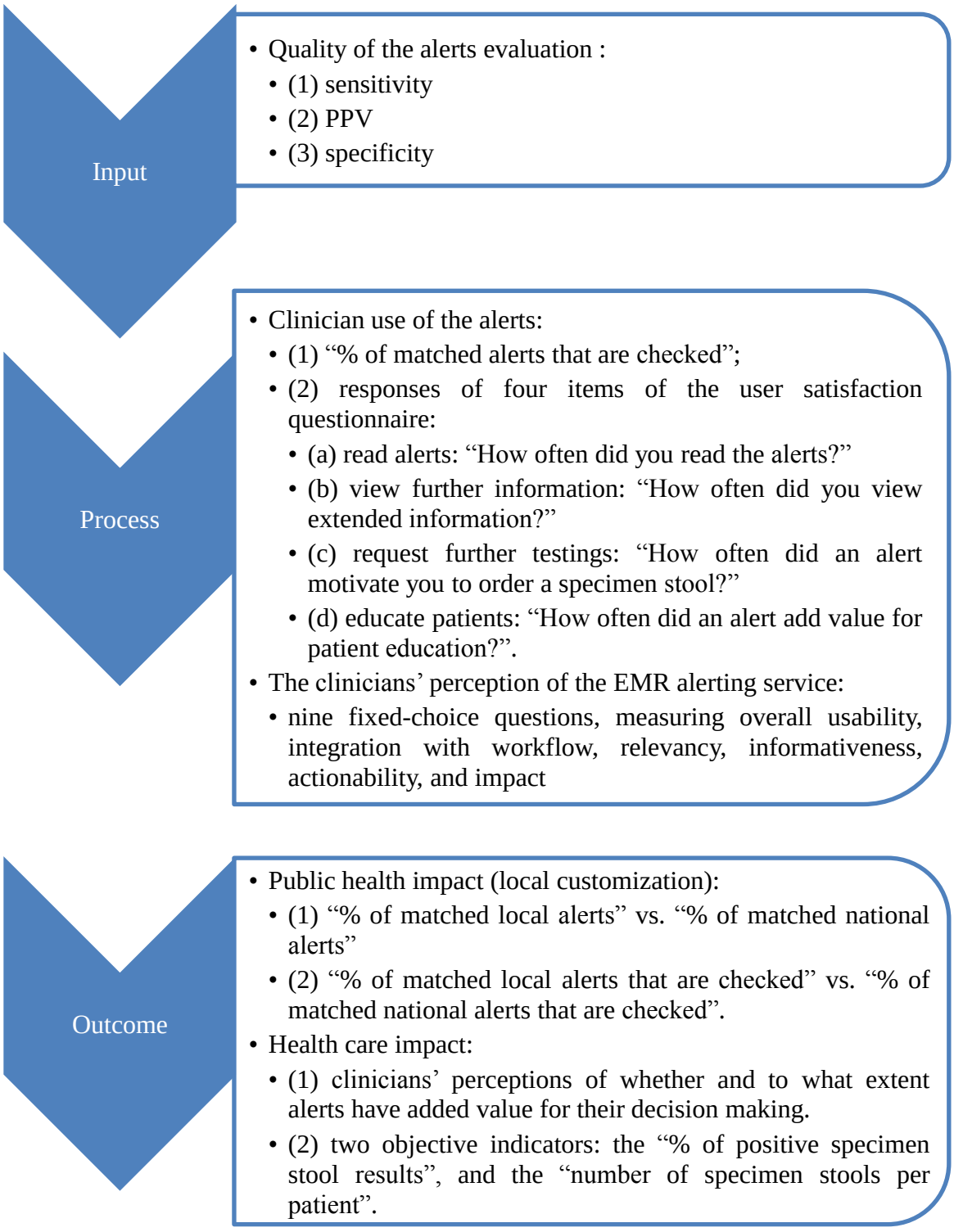
Figure 3. Sample size estimations for different comparison strategies

Correlations between indicators

As described above, several indicators will be used to evaluate the system performance. To guarantee the comprehensiveness of the evaluation, some features will be indicated by both objective and subjective indicators. If there is a significant correlation between an objective indicator and a subjective indicator that both measure a similar feature, both of these indicators are more likely to have high validity, indicating that they measure the same, right, thing.

Moreover, there is a causal relationship between the input, process, and outcome (Figure 4), which is measured by different indicators, thus these indicators may have significant correlations. If we really find that there are significant correlations between the values of these indicators, we may be more confident that the outcome results from the input and process rather than other factors.

Figure 4. Objective/subjective indicators for input, process, and outcome evaluation



Correlation between objective indicator and subjective indicator

The significant correlation between objective indicators and subjective indicators can indicate the validity of these indicators. For the "alert usage" feature in the process evaluation, two kinds of indicators were identified: the "% of matched alerts that are clicked" (an objective indicator), and the score of the "read alerts" item in user perception questionnaire "How often did you read the alerts? (never, rarely, sometimes, frequently, often)". If there is a significant correlation between these two indicators, they may both have high validity to indicate the extent of "alert usage".

Causal relationship between input, process, and outcome indicators

(1) <u>Correlation between input and process</u>

There is a casual relationship between input and process. For example, if matched alerts have high quality (e.g. return alerts are highly relevant to the patient), a clinician will be more likely to use the alerts and feel satisfied with the EMR alerting service when they meet with patients. In that case, the PPV, which indicated the relevance of the matched alerts, may correlate with the score of the "relevance item" in the user perception questionnaire "how often was an alert relevant? (never, rarely, sometimes, frequently, often)", which captures the clinician satisfaction with the relevance feature.

(2) <u>Correlation between process and outcome</u>

There is also a casual relationship between process, and outcome. High usage of high quality alert service may bring better outcomes.

For example, if the alerts motivated the clinicians to order specimen stools, the average number of specimen stools per patient may increase. This casual relationship can be indicated by the correlation between the score of "how often did an alert motivate you to order a specimen stool?" (a process measure) and the "number of specimen stools per patient" (an outcome measure).

There is another example: if the alerts add values for clinician decision making process (helping to identify high potential cases), then these populations' laboratory results may be more likely to be positive than those of low risk populations. This hypothesis can be tested by the correlation between the score of "alerts have added value for decision making" (a process measure) and the "% of positive specimen stool results" (an outcome measure).

Statistical method to analyze the correlations

Kendall's Tau-b statistic will be used to calculate these correlations. For example, to calculate the correlation between "% of matched alerts that are clicked" and the score of the read alerts item, we will first calculate the values of these two indicators for each participating clinician. Then, we will use the Kendall's Tau-b statistic to measure the strength of this correlation. The Kendall's Tau-b statistic is usually used for calculating the rank-based association between two ordinal variables (Wilcox RR, 2010.). The reason for using this non-parametric statistical method is because the number of participating clinicians is not large enough to make the assumption that the distributions of the values were Normal Distributions.

In addition, other possible correlations could also be tested. For example, quality of alerts may be correlated with alerts usage. If alerts are more relevant to patients, clinicians will be more likely to click the box to further check information.

Chapter V Conclusions, Implications, and Recommendations

Summary of study

As a critical component of public health surveillance, the CDC and state and local public health authorities disseminate public health alerts to healthcare providers to increase their awareness of potential public health threats and to enable timely and effective responses. These communications are realized through various channels in diverse formats, which have not been integrated with existing clinical workflows. This leads the audiences to a paradoxical situation - too much useless information (repeated or irrelevant messages), or too little relevant information (lack of access authority to state-based information, or hard to quickly find related messages).

To address these challenges, the CDC convened a stakeholder meeting that identified a potential solution - the integration of the public health alerts into an existing electronic health record system. A prototype, Alert Knowledge Repository, was developed. After that, the CDC EMR alerting team began to develop an updated prototype, Public Health Alert Repository System, and determined to conduct a real pilot to evaluate its performance.

Before that pilot, a comprehensive evaluation plan was needed to guarantee the completeness and validity of the evaluation results. As a major part of my practicum in the CDC EMR alerting project, I designed a comprehensive evaluation plan.

Major stakeholders are the CDC EMR alerting project team, Alliance of Chicago Community Health Services, General Electric Healthcare, and the Chicago Department of Public Health. Considering their interest areas, the evaluation will mainly focus on seven objectives: actionable alerts, consumption of alert, integration with the decision support system, clinician action performed, sensitivity, specificity, and PPV of matching algorithm, and local customization.

These objectives were mapped with five elements, including system quality, quality of the alerts, clinician use of the alerts, user perception, and impact, which were identified based on the logic model framework.

System quality will be evaluated on technological characteristics of alert repository, matching rules, integration with EMR system, user interface, as well as functional performance during both mock-up data scenarios and the real pilot.

Quality of the alerts will be measured with sensitivity (recall), PPV (precision), and specificity, using the Gold Standard method (Kappa value will be an indication of the expert agreement).

Clinician use of the alerts will be indicated by "% of matched alerts that are checked" and responses of four items of the user perception questionnaire: (1) read alerts: "How often did you read the alerts? (never, rarely, sometimes, frequently)"; (2) view further information: "How often did you view extended information? (never, rarely, sometimes, frequently)"; (3) request further testings: "How often did an alert motivate you to order a specimen stool? (never, rarely, sometimes, frequently)"; (4) educate patients: "How often did an alert add value for patient education? (never, rarely, sometimes, frequently)".

The clinicians' perception of the EMR alerting service will be measured with a user perception questionnaire, the final version of which has of two parts: (1) nine fixed-choice questions, measuring clinicians' perception in six aspects - overall usability, integration with workflow, relevance, informativeness, actionability, and impact (Quantitative methods, including a frequency table and a bar chart will be used to analyze responses); (2) one open response

question (Qualitative analysis will be used to summarize and organize the free text response of the open response area).

Both public health impact and health care impact of the EMR alerting services will be measured during the evaluation process. Public health impact could be indicated by the expedited timeline of alerts dissemination resulting from integration with existing clinician workflow, and relevance improvement because of local customization ("% of matched local alerts" was assumed to be higher than "% of matched national alerts"; and "% of matched local alerts that are checked" was assumed to be higher than "% of matched national alerts that are checked"). Health care impact could be measured with clinicians' perceptions of whether and to what extent alerts have added value for their decision making as well as two objective indicators (the "% of positive specimen stool results" and the "Number of specimen stools per patient"). The increase of the "% of positive specimen stool results" may indicate that EMR alerting services could help clinicians to identify high risk populations. The "Number of specimen stools per patient" was only an indicator for change, but not a very convincing indictor for health care quality improvement because neither the increase of unnecessary lab orders nor the decrease of necessary lab orders is meaningful. Simple before-after design was used to compare the indicators in the pilot with those that were calculated from historical data in the same months of previous years. Fisher's Exact Test will be used for statistical significant difference testing.

We will check whether there is a significant correlation between a subjective indicator and an objective indictor. The "% of matched alerts that are checked" may correlate with the score of the "read alerts" item in the user perception questionnaire, "how often did you read the alerts? (never, rarely, sometimes, frequently, often)". If the Kendall's Tau-b statistic is significant, the two indicators may both have high validity.

We will also check the possible correlations between input, process and outcome indicators. The PPV, which indicates the relevance of the matched alerts, may correlate with the score of the "relevance item" in user perception questionnaire "how often was an alert relevant? (never, rarely, sometimes, frequently, often)". Two possible correlations between process and outcome indicators will also be calculated, including the correlation between the score of "how often did an alert motivate you to order a specimen stool? (never, rarely, sometimes, frequently, often)" and "Number of specimen stools per patient", and the correlation between the score of "alerts have added value for decision making.(strongly disagree, disagree, neutral, agree, strongly agree)" and the "% of positive specimen stool results".

Conclusion and Implication

This thesis is intended to present an evaluation plan for a pilot of a developed prototype that will be a "channel" to connect a public health alert repository to an electronic medical record system. This thesis identified stakeholders' objectives, translated these objectives into measurable evaluation matrix, and clarified measurement methods, study design, and data resources. Moreover, I justified the causal relationships between input, process, and outcome, as well as the correlations between objective indicators and subjective indicators, providing approaches to measure the validation of indicators.

This process can be learned to develop other evaluation plans that aim to map stakeholders' general objectives into validated, measurable indicators. Many aspects, including evaluation framework, evaluation elements, and measurement strategies, may be possibly adapted for other evaluations for similar systems and services.

On the other hand, this plan has its limitations. It is not intended to provide tools to measure the cost-effectiveness of development and implementation of the prototype during the pilot. It is also not designed generically in nature. Necessary modification should be made to better meet the specific objectives of the stakeholders and the service population, especially when measuring the system performance within other circumstances, or with other versions.

Recommendations

Recommendations for EMR alerting service

The EMR alerting service may not necessarily replace a public health alerting system. However, it can exist contemporaneously as a means of delivering the most relevant public health alerts to healthcare providers in a near real-time way. A well-functioning public health alert system (e.g. HAN) and a well-functioning electronic medical record system are prerequisites for the adoption of this EMR alerting service.

There may still be several other challenges for the EMR alerting service: lack of wide agreement and regulations for "public health agencies – health care facilities" collaboration, incompleteness of the disease conditional library, lack of agility for application in different EMR systems, lack of strong and convincing evidence for system success, and so forth.

In my opinion, effective public-private collaboration, well-developed standards, high levels of system agility, and rigorous evaluation with scientific methods will greatly help us to overcome these challenges. Moreover, sophiscated artificial intelligence technologies will be extremely helpful. For example, using the natural processing technology may help to parse public health messages and automatically transfer them into the format of public health alerts in the PHARS, which is a manual task in this pilot. Another example is using artificial intelligence technologies to update the matching algorithm (currently the rule based algorithm). Some machine learning algorithms, such as Bayesian Network, could estimate the uncertainties of some clinical phenomena by learning from historical data. This algorithm may predict clinical diagnosis more accurately than completely defined rules.

However, the success of developing and implementing the EMR alerting service will not be solely indicated by its technological attributes. It will also be determined by how successful this system integrates with existing clinical workflow as well as by the cost-effectiveness for both health care facilities and public health agencies.

Discussion on PPV

As the proportion of relevant alerts out of all alerts, the PPV is a critical indicator of the relevance of the EMR alerting service. However, its value also depends on disease prevalence. Low disease prevalence will lead to low PPV even if the matching algorithm works well (has high sensitivity and specificity). It seems to be unfair to give a low "grade" for the performance of the matching algorithm when both PPV and disease prevalence are low, because low PPV may result from low disease prevalence rather than bad performance. On the other hand, when disease prevalence is high, low PPV may indicate bad performance.

Although PPV may not be a fair indicator for system performance, PPV can indicate how necessary the EMR alerting service is. If only a small portion of alerts are relevant, stakeholders will undoubtedly question the benefits the EMR alerting service can bring, and they will also have concerns about the possible alert fatigue clinicians might feel.

Moreover, the positive correlation between PPV and disease prevalence indicates the priority to provide EMR alerting service for high prevalence diseases or conditions, or for the high prevalence periods of the diseases.

Recommendations for evaluation design

The simple before-after comparison will be adapted for the evaluation for our newly developed prototype. This design is more feasible but less convincing because we have to ignore the influence of several other factors, such as updates of lab procedure, or changes of population structure. It is possible that the health care quality improvement is the result of utilizing advanced specimen stool testing procedure that is more precise, rather than targeting a higher risk population with the help of EMR alerting service. It is also possible that the disease prevalence of the population have increased, making the proportion of positive lab results increase. All in all, if we do not take account of these possibilities, our evaluation results may have some biases.

On the other hand, the possible correlations between input, process, and outcome will be calculated. If there are significant correlations between them, we may be more confident that the outcome results from the input and process rather than other factors.

Recommendations for user perception questionnaire design

Although the questionnaire will be our major tool to measure user perception, it is still necessary to conduct some personal interviews that will help us to clarify interviewers' responses in the "open response area". This information, especially negative feedback, will be helpful for updating the prototype to better meet users' needs.

The responses to the questionnaire may not be comparable to those of other systems because our questionnaire did not adapt items that are cited from other questionnaires. For instance, the answer “strongly agree” for one item in our questionnaire cannot be simply regarded as evidence of better satisfaction with our system than other system whose evaluation questionnaire has an “agree” response, even if the two items are designed to evaluate similar features.

Repeated measurements will help to test the reliability of the questionnaire. Moreover, if a larger sample size is available, the principal factor analysis method could be utilized to find how many factors this questionnaire contains and the correlations between items. The factors that are calculated from the responses can be compared with the initial conceptual model of the questionnaire, indicating the validity of the designed questions.

References

Abernethy NF. Automatics social network models for tuberculosis contact investigation. Biomedical informatics PhD Dissertation of Stanford University. 2005.

Altman DG, Bland JM. Diagnostic tests 2: predictive values. BMJ. 1994;309:102.

Bennatan EM. On time within budget, Software Project Management Practices and Techniques, third edition. 2000.

Berner ES. Clinical Decision Support Systems: theory and practice, second edition. 2007.

Buehler JW, Hopkins RS, Overhage JM, Sosin DM, Tong V. Framework for Evaluating Public Health Surveillance Systems for Early Detection of Outbreaks, recommendations from the CDC Working Group. MMWR Recommendations and Reports. May 7, 2004/53(RR05):1-11. http://www.cdc.gov/mmwr/preview/mmwrhtml/rr5305a1.htm (accessed March 14th, 2011).

Centers for Disease Control and Prevention (CDC), Department of Health and Human Services. Health Alert Network. http://www2a.cdc.gov/HAN/Index.asp (accessed March 14th, 2011).

Centers for Disease Control and Prevention (CDC), Emergency Risk Communication Branch (ERCB), Division of Emergency Operations (DEO), Office of Public Health Preparedness and Response (OPHPR). About COCA. http://www.bt.cdc.gov/coca/about.asp (accessed March 14th, 2011).

Cork RD, Detmer WM, Friedman CP. Development and Initial Validation of an Instrument to Measure Physicians' Use of, Knowledge about, and Attitudes Toward Computers. JAMIA. 1998;5:164-176.

Dawes J. Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. International Journal of Market Research. 2007;50(1):61–77.

Eccles M, McColl E, Steen N, Rousseau N, Grimshaw J, Parkin D, Purves L. Effect of computerised evidence based guidelines on management of asthma and angina in adults in primary care: cluster randomised controlled trial. BMJ. 2002;325:941-944.

Elizabeth H, Ken J, Jeremy D. Requirements Engineering, third edition. 2010.

Eslami S, Abu-Hanna A, de Keizer NF. Evaluation of Outpatient Computerized Physician Medication Order Entry Systems: A Systematic Review. JAMIA. 2007;14:400-406.

Fiol GD, Haug PJ, Cimino JJ, Narus SP, Norlin C, Mitchell JA. Effectiveness of Topic-specific Infobuttons: A Randomized Controlled Trial. JAMIA. 2008;15:752-759.

Friedman C, Wyatt J. Evaluation methods in medical informatics, second edition. 2005.

Garrett NY, Mishra N, Nichols B, Staes CJ, Akin C, Safran C. Characterization of Public Health Alerts and Their Suitability for Alerting in Electronic Health Record Systems. Journal of Public Health Management & Practice. January/February 2011;17(1):77–83.

Garg AX, Adhikari NK, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, Sam J, Haynes RB. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. The Journal of the American Medical Association. 2005; 293(10):1223-1238.

Gesteland PH, Allison MA, Staes CJ, Samore MH, Rubin MA, Carte ME, Wuthrich A, Kinney AY, Mottice S, Byington CL. Clinician use and acceptance of population-based data about

respiratory pathogens: implications for enhancing population-based clinical practice. AMIA Annu Symp Proc. 2008:232-236.

Gordis L. Epidemiology, fourth edition. 2008.

Goud R, de Keizer NF, ter Riet G, Wyatt JC, Hasman A, Hellemans IM, Peek N. Effect of guideline based computerized decision support on decision making of multidisciplinary teams: cluster randomized trial in cardiac rehabilitation. BMJ. 2009;338: b1440.

Instrument Review Criteria. Medical Outcomes Trust Bulletin. 1995.

Jacobs D. Requirements engineering so things don't get ugly. Software Engineering Technology. 2004,19-25.

Lee LM, Teutsch SM, Thacker SB, Louis ME. Principles & Practice of Public Health Surveillance, third edition. 2010.

Lewis JR, Human Factors Group, Boca Raton, FL. IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use. IBM Corporation Technical Report 54.786. 1993.

Lombardo JS, Garrett N, Loschen W, Seagraves R, Nichols B, Babin S. An Informatics Solution for Informing Care Delivery of Immediate Public Health Risks to Their Patients. Online Journal of Public Health Informatics. 2009; 1(1). http://ojphi.org (accessed March 14th, 2011).

Lund T, Barksdale S. Rapid Evaluation (ASTD Learning and Performance Workbook). 2001.

Magnus D, Rodgers S, Avery AJ. GPs' views on computerized drug interaction alerts: questionnaire survey. Journal of Clinical Pharmacy and Therapeutics. 2002;27(5):377–382.

Meystre SM, Haug PJ. Randomized controlled trial of an automated problem list with improved sensitivity. International Journal of Medical Informatics. 2008;77(9): 602-612.

Phillips JJ, Bothell TW, Snead GL. The Project Management Scorecard: Measuring the Success of Project Management Solutions. Butterworth-Heinemann. 2002.

Preston CC, Colman AM. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. Acta Psychologica. 2000;104:1–15.

Public Health Information Institute. Towards Measuring Value: An Evaluation Framework for Public Health Information System. 2005.

Ray-Coquard I, Philip T, De Laroche G, Froger X, Suchaud J-P, Voloch A, Mathieu-Daudé H, Fervers B, Farsi F, Browman GP, Chauvin F. A controlled 'before-after' study: impact of a clinical guidelines programme and regional cancer network organization on medical practice. Br J Cancer. 2002;86(3):313–321.

Rosner B. Fundamentals of Biostatistics, Sixth edition. 2005.

Staes CJ, Wuthrich A, Gesteland P, Allison MA, Leecaster M, Shakib JH, Carter ME, Mallin BM, Mottice S, Rolfs R, Pavia AT, Wallace B, Gundlapalli AV, Samore M, Byington CL. Public Health Communication with Frontline Clinicians During the First Wave of the 2009 Influenza Pandemic. Journal of Public Health Management and Practice. 2011;17(1):36 – 44.

Shojania KG, Grimshaw JM. Evidence-Based Quality Improvement: The State Of The Science. Health Affairs. 2005;24(1):138-150.

Simon S, Higginson IJ. Evaluation of hospital palliative care teams: strengths and weaknesses of the before-after study design and strategies to improve it. Palliative Medicine. 2009;23:23-28.

Westbrook JI, Braithwaite J, Iedema R, Coiera EW. Evaluating the impact of information communication technologies on complex organizational systems: a multi-disciplinary, multi-method framework. Medinfo. 2004;1323-1327.

Wilcox RR. Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy, second edition. 2010.

Wyatt JC, Wyatt SM. When and how to evaluate health information systems? International Journal of Medical Informatics. 2003;69 (2):251-259.

Yu K, Jacob A, Daniel CM, Donna CD, Doug G, Antoun H, William NJ, W. Paul N, Gregory PS, Michelle W. Practitioners' Views on Computerized Drug–Drug Interaction Alerts in the VA System. Journal of the American Medical Informatics Association. 2007; 14(1):56-64.

Yusof MM, Papazafeiropoulou A, Paul RJ, Stergioulas LK. Investigating evaluation frameworks for health information systems. International journal of medical informatics. 2008 (77): 377–385.