THE DEVELOPMENT AND EVALUATION OF A LEARNING ELECTRONIC MEDICAL RECORD SYSTEM

by

Andrew Joseph King

B.S., University of Pittsburgh, 2013

M.S., University of Pittsburgh, 2015

Submitted to the Graduate Faculty of

the School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Andrew J. King

It was defended on

July 13, 2018

and approved by

Dr. Shyam Visweswaran, Associate Professor, Biomedical Informatics

Dr. Harry Hochheiser, Associate Professor, Biomedical Informatics

Dr. Gilles Clermont, Professor, Critical Care Medicine

Dissertation Director: Dr. Gregory F. Cooper, Professor, Biomedical Informatics

THE DEVELOPMENT AND EVALUATION OF A LEARNING ELECTRONIC MEDICAL RECORD SYSTEM

Andrew J. King, PhD

University of Pittsburgh, 2018

Electronic medical record (EMR) systems are capturing increasing amounts of data per patient. For clinicians to efficiently and accurately understand a patient's clinical state, better ways are needed to determine when and how to display patient data. The American Medical Association envisions EMR systems that manage information flow and adjust for context, environment, and user preferences. We developed, implemented, and evaluated a prototype Learning EMR (LEMR) system with the aim of helping make this vision a reality.

A LEMR system, as we employ the term, observes clinician information seeking behavior and applies it to direct the future display of patient data.

The development of this system was divided into five phases. First, we developed a prototype LEMR interface that served as a testing bed for LEMR experimentation. The LEMR interface was evaluated in two studies: a think aloud study and a usability study. The results from these studies were used to iteratively improve the interface. Second, we tested the accuracy of an inexpensive eye-tracking device and developed an automatic method for mapping eye gaze to patient data displayed in the LEMR interface. In two studies we showed that an inexpensive eye-tracking device can perform as well as a costlier device intended for research and that the automatic mapping method accurately captures the patient information a user is viewing. Third, we collected observations of clinician information seeking behavior in the LEMR system. In three studies we evaluated different observation methods and applied those methods to collect training data. Fourth, we used machine learning on the training data to model clinician information seeking behavior. The models predict information that clinicians will seek in a given clinical context. Fifth, we applied the models to direct the display of patient data in a prospective evaluation of the LEMR system. The evaluation found that the system reduced the amount of time it takes for clinicians to prepare for morning rounds and highlighted about half of the patient data that clinicians seek.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

You can learn something from everyone.

I am grateful to the many people who have influenced who I am and the direction of my life.

Thanks to Mom (Jackie King) whose loving heart never wavers and to Dad (Gary King) who offered the continuous reminder that I did not yet have a job.

Thanks to my Grandparents (Howard and Joan Kern) who have fed me nearly every Wednesday for the last nine years.

Thanks to my brothers because the competitive spirit among the four of us encouraged me to work harder, longer, and be a better person. Thanks in particular to my oldest brother Gary (and Rebecca King) for trusting me to be the Godparent to his son Paxton; to my older brother Howie (and Carolyn King; daughters Evianna and Alaina) who occasionally helped me buy beverages in college; to my younger brother Bobby (and Jess King) with whom in adulthood I enjoy a peaceful camaraderie.

Thanks to my advisors and committee members: Dr. Shyam Visweswaran for his research guidance and funding support; Dr. Harry Hochheiser for leading the training department and providing helpful research advice; Dr. Gilles Clermont for his clinical guidance and for helping me recruit intensivists for over 150 hours of user studies; and, a special thanks to Dr. Gregory F. Cooper for being a terrific advisor.

Thanks also go to Howard Su for extracting data, Louisa Zhang for helping with data mapping, John Levander for jumpstarting the interface, Victoria Khersonsky for solving IT challenges, Dr. Gerry Douglas for the global health informatics experience in Malawi (Africa), John Galt for being who he is, and Dr. David Boone and Dr. Mike Becich for the opportunities to grow as an advisor and an educator.

Thanks to the many administrators who have helped me along the way: Lucy Cafeo and Nova Smith for processing study participant compensation; Brandon McLaughlin for reserving office space to conduct my studies; and, a special thanks to Toni Porterfield for insuring that the many students within the department are registered, paid, and insured.

Thanks to the National Institutes of Health and National Library of Medicine for support (under the award numbers T15LM007059, R01LM012095, and R01GM088224), the Jewish Healthcare Foundation for the Health Innovators Fellowship experiences, and the American Medical Informatics Association for being my professional home.

Thanks to my friends from my hometown (Uniontown, PA), from college, and from graduate school for the support and companionship.

And, thank you to my very tall and beautiful girlfriend, Louisa Zhang, for helping me through the rough patches of the home stretch.

Abbreviations:

AKF: Acute Kidney (Renal) Failure

ARF: Acute Respiratory Failure

AUC: Area Under the Curve

AUROC: Area Under the Receiver Operator Characteristic (curve)

D2K: Data to Knowledge

EMR: Electronic Medical Record

HID: Highlighted Information Display

HIDENIC: HIgh DENsity Intensive Care (data)

ICU: Intensive Care Unit

K2P: Knowledge to Practice

LEMR: Learning Electronic Medical Record

LHS: Learning Health System

P2D: Performance to Data

PR: Precision Recall

ROC: Receiver Operator Characteristic (curve)

UPMC: University of Pittsburgh Medical Center

1.0 INTRODUCTION

Going beyond serving as a repository of patient data, electronic medical record (EMR) systems should assist clinicians in decision making by intelligently integrating and presenting patient data [1]. Current EMR systems are capturing increasing amounts of such data and have few mechanisms, if any, to prioritize and present the data in clinically meaningful ways. As a result, a clinician's comprehension of a patient's condition may be incomplete or inaccurate because critical data in the EMR, such as an abnormal laboratory test result, may be overlooked [2]. Subsequent clinical action or inaction may be inappropriate and result in adverse medical events [3–5]. Thus, there is an urgent need for EMRs that better organize and display patient data, help identify patterns in the data, and aid more effectively in clinical assessment and management.

The problem of presenting data that are helpful to the clinician can be addressed in different ways. One approach is to use pre-attentive features, such as color, to bring a clinician's attention to specific information [6]. Another approach is to alert clinicians with a notification or a popup [7]. A third approach is to adapt the presentation of data in the EMR according to the context of use, such as the use condition-specific templates and user-specific profiles [8–10]. These context-aware views focus attention on data that a clinician is predicted to use while reducing the prominence of less useful data [11]. This dissertation concentrates on developing and evaluating a data-driven, learning electronic medical record (LEMR) system that observes clinician information seeking behavior and applies it to direct the display of patient data.

More specifically, clinician information seeking behavior involves patient data (e.g., glucose levels, insulin dosing regimen) that are recorded in the EMR for a particular patient and are sought by a clinician in that patient for a specific task. For example, a clinician who is preparing to present at morning rounds a patient who has diabetes mellitus and is on insulin may seek glucose levels and the insulin dosing regimen. In a different patient who has kidney failure, glucose levels may be measured, but may not be sought by the clinician. Clinician information seeking behavior may vary by context. Context includes (1) EMR user type — a clinician, a nurse and a pharmacist may have different information seeking behaviors [12]; (2) clinical task — a clinician has different information seeking behavior when performing differential diagnosis than when performing medication reconciliation; and (3) patient case — the same clinician when performing the same clinical task for different patients may have different information seeking behavior that are driven by differences in diagnoses and stage of disease.

The work presented in this dissertation is EMR-centric. However, the basic concepts and methods could be applied in a wide range of other domains, such as operations control centers and online education.

1.1 CURRENT METHODS

Computerized clinical decision support provides clinicians and other healthcare workers with knowledge, information, and recommended actions in a range of settings for a variety of tasks [13]. Examples of clinical decision support include info buttons [14], alerting systems [7], reminder systems [15], and recommender systems [16]. To be truly effective, decision support must address user needs, deliver support in a timely manner, fit into the users' workflow, and maintain an effective knowledge base [17]. Sittig et al.[13] has outlined a list of grand challenges in clinical decision support. Three of the challenges are especially relevant to this dissertation research. They are to (1) summarize patient-level information, (2) prioritize and filter recommendations to the user, and (3) combine recommendations for patients with comorbidities. Patient-level summaries allow clinicians to gain rapid understanding of a patient's state when a large amount of data may be present [18–20]. Prioritizing and filtering data is a strategy to avoid clinician information overload. Comorbidities are the simultaneous presence of two or more diseases or conditions in a patient case. To provide a centralized and complete picture of a case, decision support systems should combine recommendations based on all of a patient's comorbidities rather than considering each condition in isolation. Considering combinations is especially important when the management for one of a patient's comorbidities conflicts with the management of another. Each of these issues becomes increasingly important as more patient data are captured by EMR systems and becomes available to health care providers [21]. Addressing all these challenges is important in building effective clinical decision support systems.

To meet these challenges, an effective clinical decision support system would likely need to be integrated closely with the EMR[*] system. EMR systems are increasingly common [22] and offer the potential to improve patient safety [23]. However, due to clutter [24,25], information overload [26,27], a mismatch between clinician workflow and EMR workflow [28,29], and additional issues, preventable adverse events are still prevalent [30]. The American Medical Association has listed reducing the cognitive workload on health care providers as a top priority in improving EMR usability [31]. A top priority states that EMRs "should support medical-decision making by providing

[*] In this dissertation, EMRs and EHRs (electronic health records) are considered synonymous.

concise, context sensitive, and real-time data uncluttered by extraneous information." A concise and context sensitive system ideally will present only the data that is necessary to deliver optimal care (now and into the future). To increase usability further, EMRs "should manage information flow and adjust for context, environment, and user preference" so that the display of data matches the clinician's workflow [31].

A 2007 Institute of Medicine report, The Learning Healthcare System: Workshop Summary, envisions EMR systems that provide "an intelligent integration of information about the individual with evidence related to that individual, presented in a way that lets the provider and the patient make the right decisions" [1]. The need for this integration stems from the challenge clinicians face in aggregating, synthesizing, and identifying increasing amounts of data that are displayed by the EMR system [13,18,32]. An EMR system that focuses the clinician's attention on the patient data that she is likely to use could help reduce the time she needs to assess a patient's condition [33], as well as improve decision making and reduce medical errors. This dissertation describes progress towards making such an EMR system a reality.

1.2 PROPOSED APPROACH

Let a *Learning EMR* (LEMR; pronounced lemur) designate a system that observes clinician information seeking behavior and applies it to direct the future display of patient data[*]. The LEMR system dynamically adapts the interface to highlight context-relevant patient data. Highlights are any presentation of data that guides the clinician to focus more on one subset of patient data relative

[*] Other types of EMR-related learning tasks are possible as well, but we do not pursue them in this dissertation.

to the remainder. Highlighted information should be personalized to the clinician who is using the EMR system, the purpose for which she is using it, and the clinical condition of the patient whose data is being viewed. Ideally, the system would function as if there was a team of clinical experts that behind the scene is able to efficiently decide for the current patient which data to highlight for a given clinician. Instead of having a team of experts, the LEMR system uses statistical models to identify the data to highlight.

To learn clinician information seeking behavior, we build models that use patient data that are recorded in the EMR to predict which data items would be sought by clinicians (the *target data*) and, therefore, should be highlighted in a future patient case. The predictors are all the data items in the patient record and the targets are items that might be highlighted, with one distinct model for each item. Target data are not readily available in the EMR, so we collected them from clinicians in a laboratory setting with the LEMR system. Methods for inferring and collecting target data are one of this dissertation's contributions. The LEMR system is intended to improve the efficiency of using the EMR and help reduce the risk of missing important patient data, due for example to information overload [26,27,34]. The LEMR system accomplishes this *by observing clinician information seeking behavior and applying it to direct the future display of patient data.* In any context, a clinician uses a subset of the available patient data in the EMR [35,36]; the LEMR system seeks to identity the right subset of data and highlight it at the right time [37,38].

1.3 HYPOTHESIS AND SPECIFIC AIMS

Our hypothesis is that a LEMR system that highlights patient data that are likely to be sought by clinicians will yield the following results on a set of test cases: (1) on average clinicians will use less time in preparing for a specified task (e.g., summarize a patient case at morning rounds), and (2) clinicians will judge that the system highlights all the patient data that they would seek in each case for the specified task.

To evaluate the hypothesis, this dissertation performed the following specific aims:

1. Develop a LEMR interface that is sufficient for the planned experiments.

2. Develop automatic eye-tracking for the LEMR interface: Develop an eye-tracking system for the LEMR system that will automatically identify data that a clinician views in the LEMR interface.

3. Observe clinician information seeking behaviors: In a set of patient cases, observe and record data in each case that clinicians seek as relevant when performing a given clinical task. Use clinical information seeking data in conjunction with patient data to create a training data set for applying machine learning.

4. Model clinician information seeking behavior and evaluate the models: Apply machine learning methods to derive statistical predictive models from the training data set. Evaluate the performance of the models using precision, recall and area under the Receiver Operating Characteristic curve (AUROC).

5. Apply models to direct the future display of patient data: On a separate set of evaluation cases, apply high-performing models obtained in the previous Aim, and evaluate their performance with clinicians. In particular, measure clinician time in preparing for a specific

task (e.g., summarize a patient case for morning rounds) and clinician judgement that the system highlights data that they would use in each case for the specified task.

The five specific aims are depicted in Figure 1.

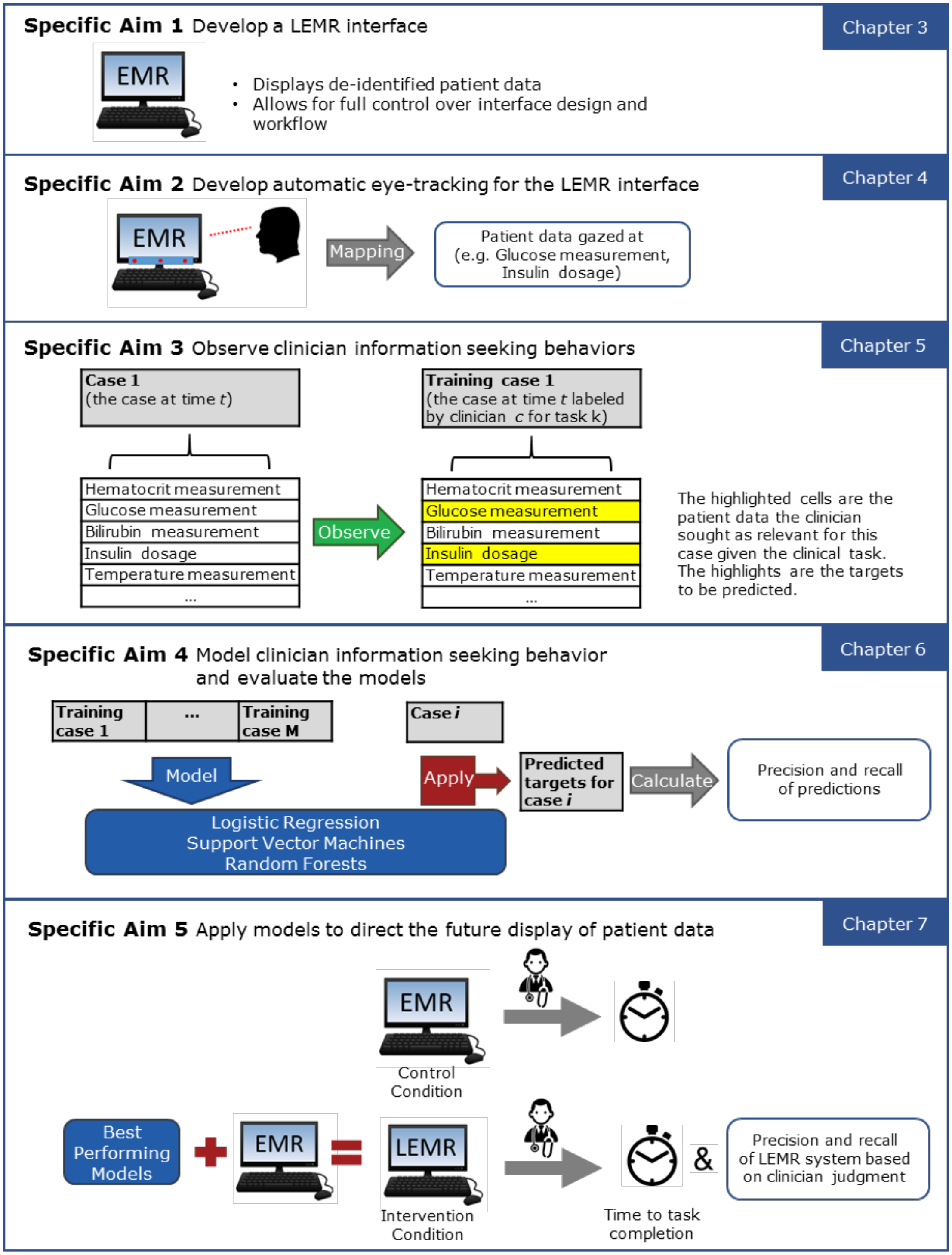

Specific Aim 1 Develop a LEMR interface
Chapter 3
EMR
• Displays de-identified patient data
• Allows for full control over interface design and workflow
Specific Aim 2 Develop automatic eye-tracking for the LEMR interface
Chapter 4
EMR
Mapping
Patient data gazed at (e.g. Glucose measurement, Insulin dosage)
Specific Aim 3 Observe clinician information seeking behaviors
Chapter 5
Case 1 (the case at time t)
Hematocrit measurement
Glucose measurement
Bilirubin measurement
Insulin dosage
Temperature measurement
...
Observe
Training case 1 (the case at time t labeled by clinician c for task k)
Hematocrit measurement
Glucose measurement
Bilirubin measurement
Insulin dosage
Temperature measurement
...
The highlighted cells are the patient data the clinician sought as relevant for this case given the clinical task. The highlights are the targets to be predicted.
Specific Aim 4 Model clinician information seeking behavior and evaluate the models
Chapter 6
Training case 1
...
Training case M
Case i
Model
Apply
Predicted targets for case i
Calculate
Precision and recall of predictions
Logistic Regression
Support Vector Machines
Random Forests
Specific Aim 5 Apply models to direct the future display of patient data
Chapter 7
EMR
Control Condition
Best Performing Models
EMR
LEMR
Intervention Condition
&
Time to task completion
Precision and recall of LEMR system based on clinician judgment

Figure 1. Five specific aims of the LEMR system. A LEMR system observes clinician information seeking behavior and applies it to direct the future display of patient data. This figure maps the five specific aims of LEMR system development to chapters of this dissertation.

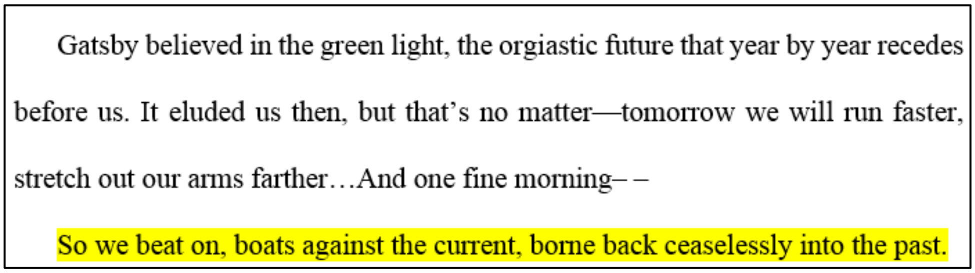
1.4 EXAMPLE APPLICATIONS OF LEMR SYSTEMS

To demonstrate the board applicability of LEMR system methods, this section presents five example applications of LEMR systems. The first two examples are of LEMR systems that highlight data, which are of special interest here because this dissertation presents the development and evaluation of such a system. We briefly describe how such a LEMR highlighting system might function in the intensive care unit and in an outpatient clinic. The next three examples demonstrate additional types of LEMR systems, including systems that provide clinical alerts, diagnostic suggestions, and clinical order suggestions. These examples illustrate how LEMR systems might help address alert fatigue [39], cognitive errors [40], and ordering appropriate medical tests [41].

1.4.1 A LEMR system with in-place highlighting

Clinicians in the intensive care unit analyze large amounts of patient data every day. When doing so, they must be careful to discern trends in a patient's laboratory test results and vital sign measurements. Clinicians sometimes overlook new trends or test results because of cognitive limitations that result in information overload [27] and change blindness [42]. Clinicians who analyze large amounts of patient data could benefit from an EMR that helps them focus appropriately.

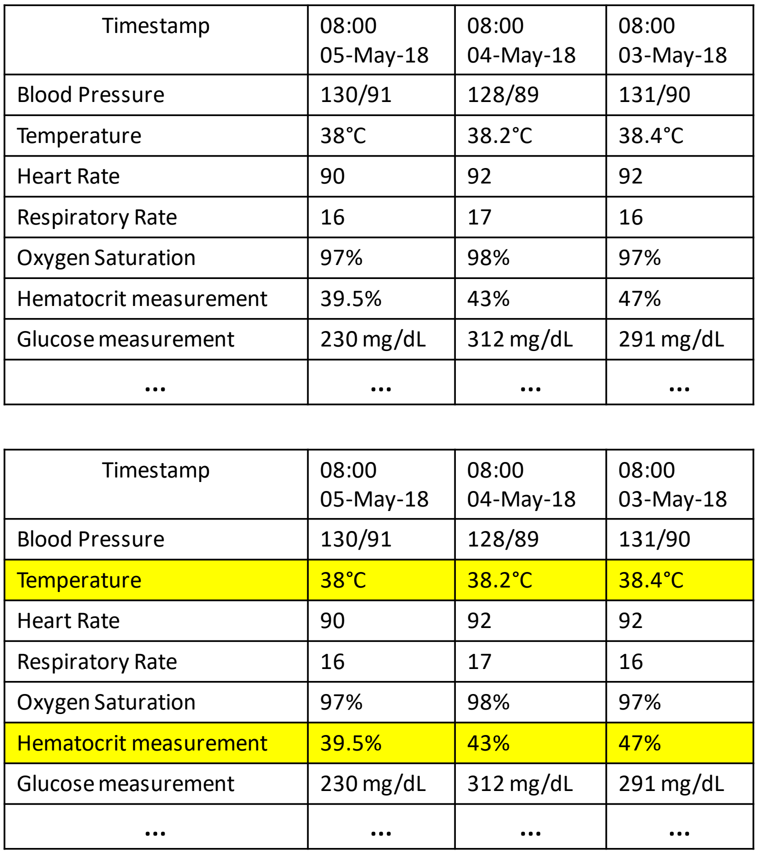
Different approaches can be used to focus user attention [43]. For example, in reading literature a reader might use a yellow highlighter to highlight lines of text that they want to find again (see Figure 2; figure text from Fitzgerald (1991) [44]). Highlighting done by one person can be used by another person to focus on particular sections of a document. A LEMR system uses highlighting to focus a clinician's attention on particular test results, vital sign measurements, and other patient data, as shown in Figure 3.



Gatsby believed in the green light, the orgiastic future that year by year recedes before us. It eluded us then, but that's no matter—tomorrow we will run faster, stretch out our arms farther…And one fine morning– –

So we beat on, boats against the current, borne back ceaselessly into the past.

Figure 2. Highlighting text is common when reading literature.



Timestamp	08:00 05-May-18	08:00 04-May-18	08:00 03-May-18
Blood Pressure	130/91	128/89	131/90
Temperature	38°C	38.2°C	38.4°C
Heart Rate	90	92	92
Respiratory Rate	16	17	16
Oxygen Saturation	97%	98%	97%
Hematocrit measurement	39.5%	43%	47%
Glucose measurement	230 mg/dL	312 mg/dL	291 mg/dL
…	…	…	…

Timestamp	08:00 05-May-18	08:00 04-May-18	08:00 03-May-18
Blood Pressure	130/91	128/89	131/90
Temperature	38°C	38.2°C	38.4°C
Heart Rate	90	92	92
Respiratory Rate	16	17	16
Oxygen Saturation	97%	98%	97%
Hematocrit measurement	39.5%	43%	47%
Glucose measurement	230 mg/dL	312 mg/dL	291 mg/dL
…	…	…	…

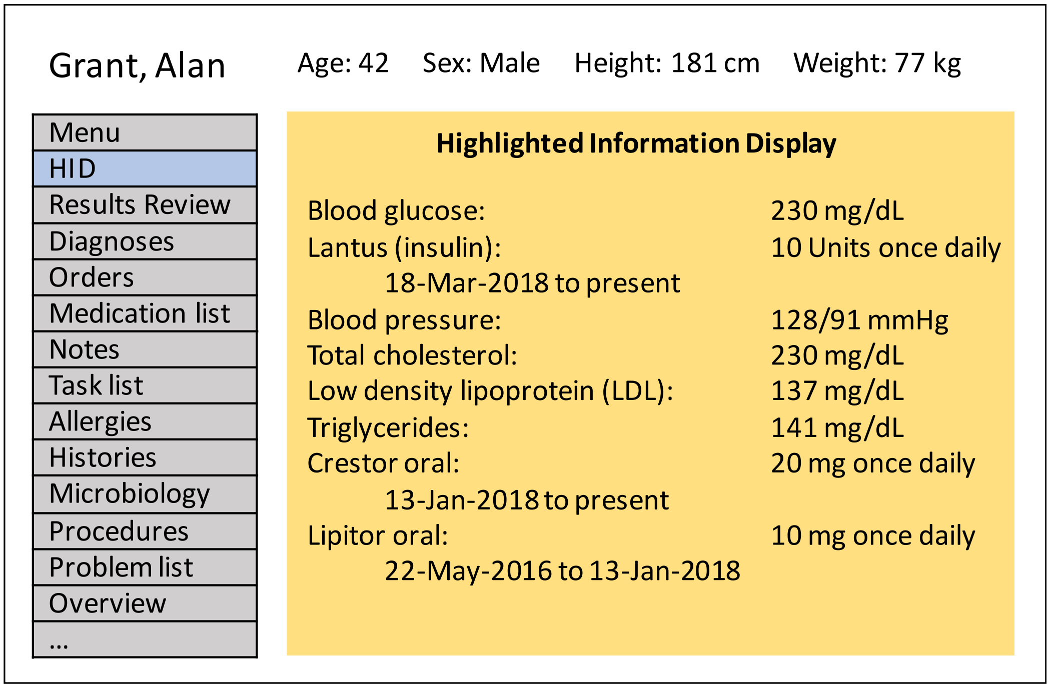
Figure 3. Example of highlighting patient data to focus a clinician's attention. This figure demonstrates how highlighting works using a fictitious list of laboratory tests results and vital sign measurements both without (top) and with (bottom) highlighting. In current EMR systems, these and other patient data are distributed across multiple tables, tabs, and screens, which increase the need for an aid to focus a user's attention. Current EMR systems highlight abnormal patient data; the LEMR system is different because it highlights any data the clinician seeks, regardless of whether its value is normal or abnormal.

1.4.2 A LEMR system with a highlighted information display

Clinicians in outpatient care must deal with a variety of patient conditions with histories sometimes dating back decades and medical records containing data collected in multiple care locations. Finding desired patient data among the large set of available data is difficult. Even when the clinician knows what patient data they desire, each page they visit and note they read still has a time cost. If desired data are spread among multiple screens, then the clinician must remember, write down, or revisit the data.

The LEMR system described in Section 1.4.1 highlights patient data in place. Alternatively, context-relevant patient data could be highlighted through a dynamically populated highlighted information display (HID) in the LEMR interface. The HID could contain any type of patient data, rather than being data source-oriented. For example, in an EMR system that is source-oriented, glucose levels are usually displayed with other laboratory test results and insulin dosing regimens are usually displayed with other medication orders. Clinicians using a source-oriented EMR system may have to switch between the laboratory test results screen and the medication order screen to appropriately adjust the patient's insulin dosing regimen. A LEMR system that predicts a clinician will seek both glucose levels and insulin dosing regimens can place them together in the HID.

Clinicians seek some patient data for nearly all patients (e.g., patient name, age, weight). These data items will always be displayed in the LEMR interface. There are other patient data that a clinician will seek for some patients but not for others (e.g., glucose levels, insulin dosing regimen, cholesterol measurements, cholesterol drug regimens). These data could be displayed in the dynamically populated HID, when a model of clinician information seeking behavior predicts that they will be sought as relevant for the current patient. For patient data a clinician seeks but is not highlighted in the LEMR interface, a clinician could find them using either traditional means of EMR navigation or an EMR search engine that adds its results to the HID. Figure 4 shows such a LEMR design.



Grant, Alan   Age: 42   Sex: Male   Height: 181 cm   Weight: 77 kg

Highlighted Information Display

Blood glucose:	230 mg/dL
Lantus (insulin): 18-Mar-2018 to present	10 Units once daily
Blood pressure:	128/91 mmHg
Total cholesterol:	230 mg/dL
Low density lipoprotein (LDL):	137 mg/dL
Triglycerides:	141 mg/dL
Crestor oral: 13-Jan-2018 to present	20 mg once daily
Lipitor oral: 22-May-2016 to 13-Jan-2018	10 mg once daily

Figure 4. Example of a highlighted information display (HID). The HID, shown with yellow background, displays patient data a clinician will seek for this case, as predicted by a statistical model of clinician information seeking behavior. The data shown in the HID demonstrates that blood glucose levels may be highlighted with the insulin dosing regimen and that cholesterol laboratory tests may be highlighted with cholesterol drug regimens. The grey panels on the left are for traditional EMR navigation.

1.4.3 A LEMR alerting system

Clinicians override up to 90% of the alerts they receive [45]. To reduce alert override rates and alert fatigue, alerts should be raised judiciously. To be useful, alert triggering criteria should be clinician-specific because different users have different past experience with patient cases and clinical alerts.

A LEMR system is suited for learning which alerts are useful for different EMR users. Such a system can track the alerts a user has seen and overridden in the past, and observe the actions of similar users after they see the same alert. If a user has overridden an alert in the past, then the system could silence the alert and monitor the user's behavior to verify their actions are consistent with the actions of other users who have seen the same alert. If the actions are consistent, then the user is likely already aware of the information contained in the 'silenced' alert. If the actions are inconsistent, then the user may not be aware of the information contained in the 'silenced' alert, so the alert should trigger. For an explicit example, rifampin and isoniazid have a moderate drug-drug interaction that leads to an increase in the incidence of hepatotoxicity. These drugs are still used in combination along with frequent monitoring of liver enzymes [46]. If an alert for this interaction were silenced for an ordering clinician who has overridden the alert in the past, and they frequently view the results of liver enzyme tests, then no further action is required because their actions are consistent with the observed actions of similar clinicians who have seen this alert (Figure 5, Scenario C). However, if an alert for this interaction were silenced for an ordering clinician who does not frequently view the results of liver enzyme tests, then the alert might trigger due to lack of follow up because their actions are inconsistent with the observed actions of similar clinicians who have seen this alert (Figure 5, Scenario B).

Scenario A	Scenario B	Scenario C
A clinician who has **never** overridden an alert for combining rifampin and isoniazid.	A clinician who has overridden an alert for combining rifampin and isoniazid.	A clinician who has overridden an alert for combining rifampin and isoniazid.
The clinician orders rifampin and isoniazid for the current patient.	The clinician orders rifampin and isoniazid for the current patient.	The clinician orders rifampin and isoniazid for the current patient.
Drug-Interaction Alert Combining rifampin and isoniazid leads to increased incidence of hepatotoxicity. If ordered, monitor liver enzymes frequently.	The moderate drug-interaction alert is silenced because a personalized model predicts the ordering clinician is already aware of risks.	The moderate drug-interaction alert is silenced because a personalized model predicts the ordering clinician is already aware of risks.
	The ordering clinician **does not** frequently monitor liver enzymes, as similar clinicians do.	The ordering clinician frequently monitors liver enzymes, as similar clinicians do.
	A 'silenced' alert is now triggered: **Drug-Interaction Alert** Combining rifampin and isoniazid leads to increased incidence of hepatotoxicity. If ordered, monitor liver enzymes frequently.	No alerts triggered.

Figure 5. Scenarios for reducing alerts with a LEMR system.

1.4.4 A LEMR diagnostic system

All humans face cognitive biases, and clinicians are no exception. Some cognitive biases most prevalent in health care are anchoring, confirmation, and diagnostic momentum [40]. Anchoring is the tendency to hold onto an initial impression, even after additional information has become available; confirmation bias is a tendency to look for evidence that conforms to one's belief about a patient's diagnosis while ignoring evidence that refutes it; and diagnostic momentum is the tendency for a particular diagnosis to become reinforced by subsequent health care providers after an initial provider attached the diagnoses to a patient (e.g., once a person is diagnosed with mild traumatic brain injury, subsequent visits to other health care providers may quickly yield the same diagnosis without the clinician fully considering all symptoms and the possibility of a more serious neurologic condition).

One strategy for reducing errors due to cognitive biases is crowd wisdom [47]. Under crowd wisdom, the biases of any one individual are offset by the opposing beliefs of other individuals within a large sample of people. Therefore, the aggregate answer of the large sample of people (the crowd) can be better than the expected answer of any one individual from within the crowd. A LEMR system could use crowd sourcing to lessen the effects of cognitive biases. To do so for patient diagnostics, the system would learn from the diagnoses assigned to many different patient cases. If for a patient case, an assigned diagnosis is not highly predicted by the LEMR system, then the system might alert the clinician to other diagnoses that are more highly predicted. To support the clinician's diagnostic decision, the LEMR system may highlight patient data other clinicians would seek when managing patients with the more highly predicted diagnoses. Highlights could be in-place (Figure 3 in Section 1.4.1) or in a HID (Figure 4 in Section 1.4.2). The highlights might focus the clinician's attention on a laboratory test result that refutes the patient's current diagnosis in favor of an alternative diagnosis that was included in the alert. In other words, alerts and highlights predicted using a model of the crowd could free a clinician from the patient data that initially suggested one condition (anchoring), and help them see other patient data that they were ignoring because those data were not relevant to the assumed diagnosis (confirmation bias and diagnostic momentum).

1.4.5 A LEMR system to support order selection

Providing good care includes ordering appropriate tests. Unnecessary testing causes patient discomfort or worse, increases health care costs, and can lead to false positive results [48]. Clinicians may order unnecessary tests because they do not know current effectiveness data, because patients ask for them, or because of the practice of defensive medicine [49,50]. These actions result in as many as 88% of patients receiving at least one unnecessary test during their first 24 hours of emergency department care [48]. Choosing Wisely® is a campaign to reduce unnecessary medical tests, treatments, and procedures [41].

A LEMR system might help realize the goals of Choosing Wisely®. A system that observes clinician information seeking behavior could keep track of which laboratory, imaging, and microbiology test results are viewed (sought) and which are ignored (not sought). During computerized physician order entry, clinicians often are shown a list of available and relevant laboratory tests to order. Using a model of test result viewing (seeking), those tests that are unlikely to be viewed (sought) might not be included in a dynamically generated order set. They could still be ordered through the EMR by explicit entry, but doing so would involve the clinician explicitly deciding that those labs are worthwhile to order, even if the results are unlikely to be viewed (sought). Besides viewing, other parameters, such as time since a test was last ordered, may be useful when predicting which tests to include in an order set.

1.5 A LEMR SYSTEM AS AN INSTANCE OF A LEARNING HEALTH SYSTEM

A learning health system (LHS) aims to "generate new knowledge as an ongoing, natural by-product of the care experience, and seamlessly refine and deliver best practices for continuous improvement in health and healthcare" [51]. LHSs are seen as an essential step in reducing the 17-year delay between scientific discovery and their use in routine clinical practice [52]. They are also a tool for achieving part of the most recent strategic plan for the National Institutes of Health: "timely dissemination and implementation of evidence-based practices" [53].

A LHS consists of a three-step learning cycle (see Figure 6) [54]. To initiate a LHS cycle, investigators or health care providers form a learning community that focuses on a health problem of interest. Once this is done, they aggregate their *data* and extract new clinical *knowledge* from it. The *knowledge* is used to influence *performance* of the clinical practices within the learning community. Finally, from clinical *performance* they generate new *data* to feed the next iteration of the learning cycle. This three-step cycle of data to knowledge, knowledge to performance, and performance to data is sometimes abbreviated as D2K, K2P, and P2D, respectively. The goal for subsequent learning cycles is continuous, rapid improvement addressing the health problem of interest.

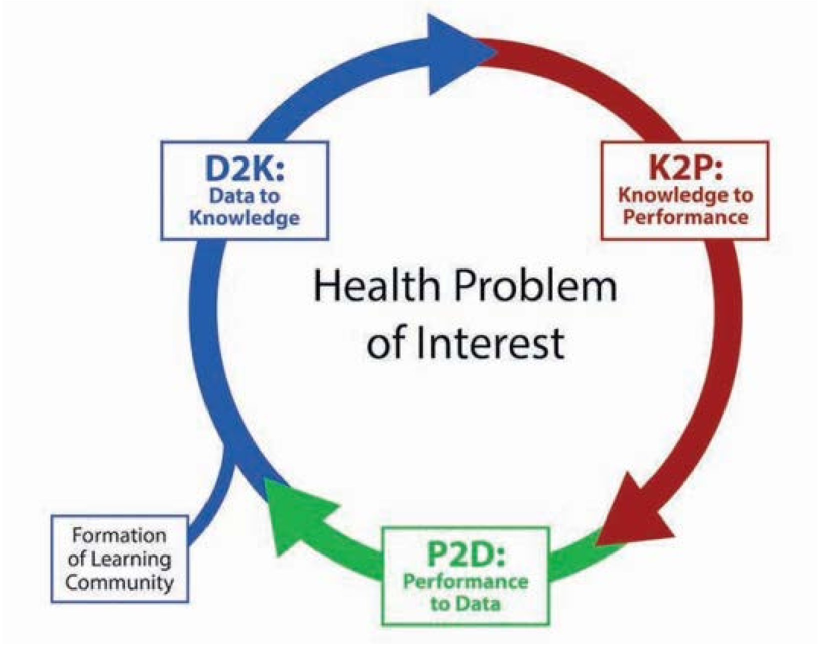


Figure 6. The learning health system (LHS) learning cycle from Friedman et al. [54]

This health problem-centric view is the typical framing of the vision of a LHS. However, learning cycles need not focus exclusively on health problems and their management. They could also focus on developing and improving computerized clinical decision support. Clinical decision support, such as a readmission risk calculator, or the LEMR system presented in this dissertation, could provide clinicians, staff, and patients with information to enhance health care and health [55]. These systems could be specific to a particular health problem or broadly applicable. An impediment to the development of clinical decision support is the need for data.

Many clinical decision support systems model local data to provide clinicians with important information about the care of a patient. The data source is important because a model trained on one patient population may have subpar performance on a different patient population. Furthermore, model performance degrades over time due to calibration drift [56]. To prevent performance loss, models must be periodically recalibrated on recently collected local data.

Therefore, to train a clinical decision support model and keep it calibrated, the system needs to continuously collect local performance *data*, use this data to update its *knowledge* base, and apply the new knowledge in practice to improve *performance*. These three steps match those of a LHS.

A LEMR system is an instantiation of a LHS. It observes clinician information seeking behavior and applies it to direct the future display of patient data. A three-step learning cycle can be used to train and calibrate a LEMR system in a clinical setting: (1) the system continuously collects local performance data (i.e., it observes clinician information seeking behavior), (2) this data is used to generate knowledge (i.e., it models clinician information seeking behavior), and (3) the new knowledge is applied in practice with the goal of improving performance (i.e., it applies a model to direct the future display of patient data). These three steps (data, knowledge, and performance) are shown in Figure 7.

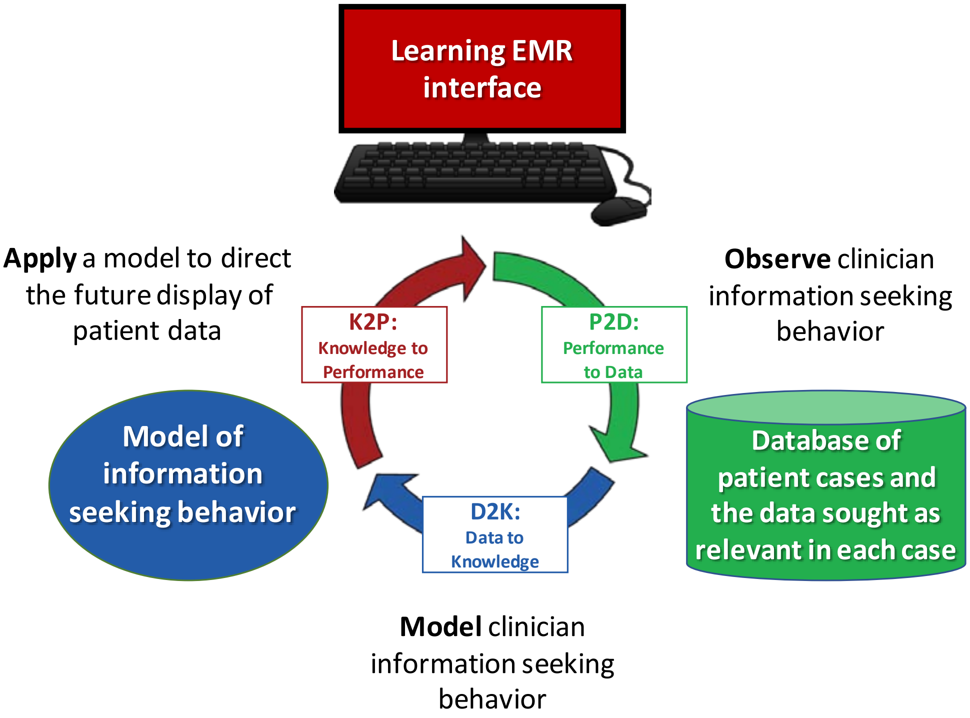

Learning EMR interface
Apply a model to direct the future display of patient data
Observe clinician information seeking behavior
K2P: Knowledge to Performance
P2D: Performance to Data
Model of information seeking behavior
Database of patient cases and the data sought as relevant in each case
D2K: Data to Knowledge
Model clinician information seeking behavior

Figure 7. LEMR as an instantiation of the learning health system.

1.6 DISSERTATION ROADMAP

This dissertation is divided into eight chapters. **Chapter 1** introduced the concept of a LEMR system with special emphasis on LEMR systems that highlight patient data. **Chapter 2** describes relevant background information, reviews related prior work, and summarizes the contributions of this dissertation. Development of a LEMR system is divided into five parts that correspond to the next five chapters. **Chapter 3** describes a prototype LEMR interface that we developed. It served as a test bed for LEMR experimentation. **Chapter 4** describes the methods that we developed for using eye-tracking with the LEMR interface to automatically observe and capture clinician information seeking behavior. **Chapter 5** explains how we enlisted the help of clinicians to manually indicate their information seeking behavior when performing a task with the LEMR interface, as well as an automatic observation method using eye-tracking. These data were applied for training the LEMR system. **Chapter 6** describes how we applied machine learning methods to model clinician information seeking behavior. **Chapter 7** presents an evaluation of the LEMR system. **Chapter 8** summarizes this dissertation and discusses limitations, future work, and insights about LEMR systems.

Figure 8 maps the chapters of this dissertation to the steps in the cycle of a LHS.

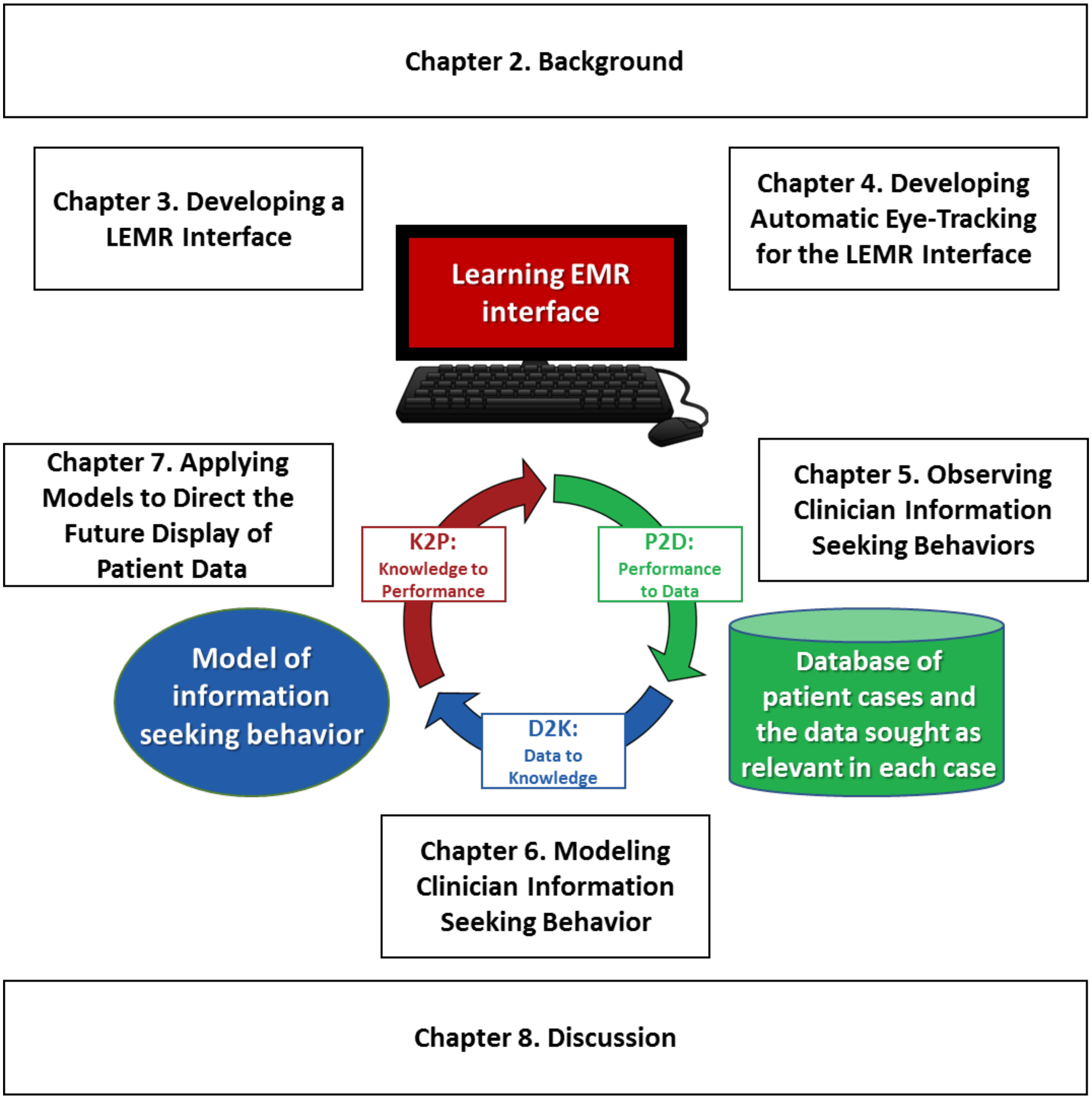

Chapter 2. Background
Chapter 3. Developing a LEMR Interface
Chapter 4. Developing Automatic Eye-Tracking for the LEMR Interface
Learning EMR interface
Chapter 7. Applying Models to Direct the Future Display of Patient Data
Chapter 5. Observing Clinician Information Seeking Behaviors
K2P: Knowledge to Performance
P2D: Performance to Data
Model of information seeking behavior
Database of patient cases and the data sought as relevant in each case
D2K: Data to Knowledge
Chapter 6. Modeling Clinician Information Seeking Behavior
Chapter 8. Discussion

Figure 8. An overview of the chapters viewed in the context of a LHS.

2.0 BACKGROUND

This chapter provides the background relevant to this dissertation, including an introduction to pertinent EMR topics; the application of eye-tracking technologies in computerized clinical decision support; the intensive care unit (ICU) environment, which is the clinical setting of this this dissertation work; the patient cases used in the experiments; and a brief overview of the supervised machine learning methods that were used. The chapter concludes with a summary of gaps in prior work and a synopsis of this dissertation's scientific contributions.

2.1 ELECTRONIC MEDICAL RECORDS

In this era of Meaningful Use [57], EMRs — with at least basic functionality — have become pervasive throughout the United States [22]. EMR systems are used to collect clinical data, to integrate data from multiple sources, and to support medical decision making [58]. While these functionalities are important, more sophisticated EMR capabilities are needed in order to further realize the promise of improved quality of care [59].

The switch to EMRs from paper-based patient reports has clear advantages in terms of information exchange, legibility, and accessibility [60,61]. Some studies have found that EMRs improve quality of care [62], while other studies offer mixed opinions [61,63]. Implementation of new information technology can result in unintended consequences [34], such as new risks for medication errors [64] and increases in mortality [65]. Some studies have concluded that clinicians feel that current EMR systems reduce their ability to stay aware and informed, resulting in reduced performance [63]. Reasons for these views include missing information, over-reliance on potentially erroneous information, and orders not being seen [63]. EMRs take clinician gaze away from the patient and onto the computer monitor [66], resulting in reduced patient centeredness [61]. Furthermore, current EMR systems often require the review of multiple screens to achieve a clinical task, due to a mismatch between clinical workflow and the way in which information is displayed [67].

Some of the issues surrounding EMR system design can be attributed to their focus on billing [68]. This focus has been driven by United States government regulations and insurance company requirements [68]. To address persisting issues, hospitals need to consider usability concerns, such as tradeoffs between unique needs of different users and system consistency, starting at implementation [69]. To classify the functionality of EMR systems, Ball et al. [70] turned to the Gartner generations model for computer-based patient record systems. Under this model, the “collector” EMR system is the first of four generations of EMR systems. In this generation, the healthcare data that have traditionally been in paper format are now electronic. The second generation is the “documenter” EMR system, where structured data can be processed for basic clinical decision support — like alerts — and for generating reports. The third generation is the “helper” system, where data are structured and standardized — with the application of standard terminologies — and most healthcare operations and task management are done through the system. Finally, a fourth generation EMR system is called the “partner.” In this generation, the EMR provides contextual support to clinicians — for example, providing decision support and workflow capabilities that are specific to the clinician and to the current clinical task. These context-aware systems are anticipated to lead to substantial improvements in healthcare [71].

2.1.1 EMR interfaces

Research into the display of patient data in EMRs has progressed in various directions including graphical summaries [72], methods to summarize and display temporal data [73], and the context specific integration of data using either systems-based [74] or disease-based [75–77] approaches. Three general strategies have been used when determining how an EMR system will group and display data [78]. The first and most common approach is to group data based on its source [79]. For example, laboratory tests results will be displayed together in a table because their source is a laboratory information system. The second approach is to use a graphical timeline [80]. While looking at a timeline, a healthcare provider can more easily understand the course of events. The third strategy is to group information by context. A review of context-aware EMR systems is provided in the next section.

2.1.2 Context-aware EMR systems

In this dissertation, *clinical context* refers to the situation surrounding an interaction between a clinician and the EMR system. It includes the type of clinician (e.g., physician, pharmacist, nurse), that clinician's specialty (e.g., cardiology, radiology, pediatrics), the clinician's role (e.g., attending, trainee, consulting), and the current channel of care (e.g., intensive care unit, general ward, outpatient). It also incorporates the patient case that is being accessed, including all the electronically available information about that patient case (e.g., history, demographics, past and current laboratory test results). Finally, it includes the purpose for which the clinician is accessing the patient's record (e.g., new admission, daily rounding review, patient handoff).

For this review of context-aware EMR systems, we only consider adaptive systems that use patient-specific details to adapt the display in a context-aware manner. Therefore, template screens, user customizable views, time-based views, and patient summary systems are not included. A summary of representative context-aware EMR systems is shown in Table 1.

Table 1. Summary of EMR systems that adapt the display of a patient case in a context-aware manner.

Author, Year	System name	Knowledge base	Knowledge source	Adaptive focus
Pickering et al., 2010 [81]	AWARE	Rules	Expert knowledge	Available data
Suermondt et al., 1993 [76]	PWS	Bayesian belief network	Medical literature, Expert knowledge	User query
Zeng et al., 1999 [82]	QCIS	Bayesian belief network	Medical literature, Expert knowledge	User query
Hsu et al., 2012 [75]	AdaptEHR	Bayesian belief network	Biomedical ontologies, Graphical disease models	Concepts extracted from clinical notes

Ambient Warning and Response Evaluation (AWARE) is an ICU system that organizes patient information in organ systems-based information packages. AWARE uses a rule base to search for pre-identified high value information. The rules were developed from expert knowledge [81,83]. When evaluated, AWARE reduced time to task completion and medical error in the assessment of ICU patients who are thought to be experiencing acute bleeding [81,84].

Physician Workstation (PWS) is an early attempt at creating patient-specific, context-aware EMR displays. PWS represents a patient state as a physiological Bayesian network. To use the system, a clinician would first select a patient problem or medication to view. Next, a program called Radarserver queries the Bayesian network to identify the patient data that influence or are influenced by the selected item. Finally, the returned patient data are displayed to the user. In addition to querying functionality, Radarserver also functions as an alert system. If a new patient event causes a worrisome change to the model, clinician users are notified via an alert message. PWS requires a manually created global physiological model to function. This model includes parameters, arcs, and relationships [76,85].

Querying Clinical Information System (QCIS) is a query-based system in which a user selects concepts of interest and relevant coded patient information is retrieved and displayed. Relevance is determined through a rule-based traversal of a semantic network. The network and rules were created from existing knowledge bases, on-line information sources, domain experts, and medical literature [77,78,82].

Adaptive EHR (AdaptEHR) aggregates and extracts findings and attributes from free-text clinical reports, maps findings to concepts in available knowledge sources, and generates a tailored presentation of the record based on the connectedness of different patient data. The available knowledge sources are biomedical ontologies and graphical disease models [75].

AWARE and the other existing integrated systems use rules to identify which of the potentially thousands of available data items are relevant in specific clinical contexts [78]. Rules are usually manually constructed from disease models, ontologies, and expert opinion. Such rule-based systems have several advantages. They are likely to be clinically informative and appropriate, since they are based on clinical knowledge, and they can be readily programmed and applied to patient data that are available in electronic form. However, construction of rules is tedious and time-consuming. Moreover, rules have limited coverage of the large space of clinical conditions, and a rule-based display may not adequately portray the context of a patient whose condition presents in an unusual way or a patient who has multiple clinical problems [86].

To our knowledge, AWARE is only one of these systems to currently be commercially available. This lack of translation is a cautionary tale for the difficulty of developing a context-aware EMR system. We do not know what became of three of the four systems. If these systems were discontinued because of the difficulty of adapting expert-driven rule bases, then the LEMR system's data-driven approach may be able to overcome this limitation.

2.2 EYE-TRACKING

Eye-tracking technologies use cameras that monitor a participant's eyes in order to determine where he or she is looking. There are two common types of eye-tracking equipment. The first is a head mounted unit that resembles a pair of glasses. These are not ordinary glasses though. They contain a pair of cameras. One camera records pupil and corneal reflection position while the other camera records the scene in front of the wearer. These two recordings are integrated in order to determine gaze location. The second type of eye tracker is a remote, external device. Eye-tracking devices of this type usually take the form of a sensor bar mounted on the bottom of a computer monitor. Thus, they are easier to setup, but can only be used in a fixed location, which typically means on a computer monitor. These sensors are able to map eye gaze onto positions (coordinates) on the computer monitor.

Eye-tracking has a long history of use in usability studies [87] and consumer sciences [88]. The use of eye-tracking for evaluating health information technology has been limited until very recently. The next section provides a review of representative work in this field.

2.2.1 Application of eye-tracking in health information technology research

During the past decade, health information technology research has incorporated the use of eye gaze data with increasing frequency. Eye-tracking devices are used to better understand clinical reasoning [89] and to evaluate usability [90]. Table 2 provides a summary of some studies that apply eye-tracking for various purposes, including patient safety, understanding workflow, and system evaluation.

Table 2. Studies that utilize eye-tracking technology in health information technology. Each study listed is described by using either a head-mounted or remote eye-tracking device. Head-mounted devices are worn by the study participant, whereas remote devices are usually mounted below a computer monitor.

Author, Year	Title	Eye-tracking device	Objective	Results
Henneman et al., 2008 [91]	Providers do not verify patient identity during computer order entry	Head-mounted, ASL Mobile Eye	Determine frequency of verifying patient ID during computerized provider order entry (CPOE)	Medical providers often miss ID errors and infrequently verify patient ID with two identifiers during CPOE.
Eghdam et al., 2011 [130]	Combining usability testing with eye-tracking technology: evaluation of a visualization support for antibiotic use in intensive care	Remote, *unknown*	Observe the visual attention and scan patterns of system users	Navigation paths were close to expected. Eye-tracking is a useful addition to usability studies.
Forsman et al., 2013 [131]	Integrated information visualization to support decision making for use of antibiotics in intensive care: design and usability evaluation	Remote, Tobii X120	Evaluate a prototype visualization tool that aids decision making of antibiotic use in the ICU	Visual attention when completing the tasks differs between specialists and residents, who focus on the tables and on exploring the GUI, respectively.
Nielson et al., 2013 [132]	In-situ eye-tracking of emergency physician result review	Remote, Tobii T60	Determine the time spent by physicians looking at lab results and fixating on specific values in a live clinical setting	Average time viewing individual lab result screen was 13.9 seconds, with 9.9 seconds fixated on particular lab values.

Table 2 (continued).

Author, Year	Title	Eye-tracking device	Objective	Results
Wright et al., 2013 [100]	Eye-tracking and retrospective verbal protocol to support information systems design	Head-mounted, ASL Mobile Eye	Development of principals to support better organization and prioritization in the presentation of electronic health data	Preliminary results described basic usability concerns, importance of laboratory information displays, and a desire to see information in big picture format.
Barkana et al., 2014 [133]	Improvement of design of a surgical interface using an eye-tracking device	Remote, SMI 500	Evaluate a proposed surgical interface in terms of gaze fixations	Fixation counts showed that displaying 8 CT scans was redundant, so they reduced the number to 2. This reduced time to task completion.
Brown et al., 2014 [96]	What do physicians read (and ignore) in electronic progress notes?	Head-mounted, ASL Mobile Eye	Identify how physicians distribute their visual attention while reading electronic notes	Physicians directed very little attention to medication lists, vital signs, or laboratory results compared with the impression and plan section of electronic notes.
Doberne et al., 2015 [94]	Using high-fidelity simulation and eye-tracking to characterize EHR workflow patterns among hospital physicians	Remote, Tobii X2 60	Characterize typical EMR usage by hospital clinicians as they encounter a new patient	Found two different information gathering and documentation workflows among participants.

Table 2 (continued).

Author, Year	Title	Eye-tracking device	Objective	Results
Gold et al., 2015 [93]	Feasibility of utilizing a commercial eye tracker to assess electronic health record use during patient simulation	Remote, Tobii X1 Light	Understand factors associated with poor error recognition during an ICU based EMR simulation.	Improved performance was associated with a pattern of rapid scanning of data manifested by increased number of screens visited, mouse clicks, and saccades.
Moacdieh & Sarter, 2015 [24]	Clutter in electronic medical records: examining its performance and attentional costs using eye-tracking	Remote, ASL D-6	Assess the effects of clutter, in combination with stress and task difficulty, on visual search and noticing performance.	Clutter degraded performance in terms of response time and case awareness, especially for high stress and difficult tasks.
Rick et al., 2015 [92]	Eyes on the clinic: accelerating meaningful interface analysis through unobtrusive eye-tracking	Remote, SMI RED-m	Observe and report clinician experiences using their EMRs.	Clinician time was predominated by searching behavior indicating that the organization of the EMR system was not conducive to clinician workflow.
Weibel et al., 2015 [134]	Lab-in-a-box: semi-automatic tracking of activity in the medical office	Remote, SMI RED-m	Create a portable multimodal data collection system to help characterize clinical workflow in the medical exam room.	Created Lab-in-a-Box and ChronoSense, which semi-automatically annotates hours of clinician-patient-EMR interaction.

Table 2 (continued).

Author, Year	Title	Eye-tracking device	Objective	Results
Fong & Hoffman, 2016 [95]	Identifying visual search patterns in eye gaze data; gaining insights into physicians visual workflow	Head-mounted, *unknown*	Propose and test an algorithmic approach to identifying search patterns from eye gaze data.	The search patterns found provide more insight into duration and directionality of area of interest transitions, than first-order results.
Mazur & Mosaly, 2016 [135]	Toward a better understanding of task demands, workload, and performance during physician-computer interactions	Head-mounted, ISCAN Vision Trak & Remote, Tobii T60XL	Assess relationships between task demands, workload, and performance for physicians using EMRs.	No significant relationship between task demands and pupillary dilations.

With respect to patient safety, Henneman et al. [91] looked at the frequency that providers verify patient identification during computerized physician order entry and found that they rarely follow the recommended two identifier verification. Rick et al. [92] found that physician time using the EMR is predominantly spent in searching and concluded that the system that they were evaluating was not conducive to physician workflow. Regarding workflow, Gold et al. [93] found that experienced physicians exhibit a pattern of rapid data scanning that increased the likelihood that they will recognize errors in simulated ICU patient cases.

Better understanding of clinical workflow is frequently the objective of eye-tracking studies. Doberne et al. [94] found that different physicians utilize different EMR information gathering strategies. Fong & Hoffman [95] developed a method for identifying visual search patterns and used it to understand physician visual workflow. Brown et al. [96] analyzed how physicians read progress notes and found that they spend most of their time reading the impression and plan section, even when the other sections contain more content. The observation by Brown et al. that different parts of a patient's record receive different levels of physician attention, suggest that the display of a patient's record should reflect these differences and highlight the patient data a clinician is going to dedicate his or her attention to.

2.2.2 Barriers to adoption

Various barriers prevent large scale eye-tracking adoption. Two of the largest barriers are cost and resource intensive gaze mapping. The combined cost of an eye-tracking device and a license for data analysis software can be over ten thousand dollars. This high cost of entry limits the number of researchers who can afford the devices. Eye-tracking data analysis (gaze mapping) can also be costly because some amount of manual annotation is usually required. Depending on the desired granularity of the results, five minutes of eye-tracking recordings can take as much as three hours to annotate [97]. Analysis software provides an automated means to do these annotations. The software also provides additional analysis options for studies that have fixed information on the computer screen. For example, if a participant is asked to look at an image on the computer screen, the software can generate a heat map that shows the duration users spent looking at different areas of the image. A heat map is good for situations where the researcher is interested in seeing what grabs a participant's visual attention and what is ignored. At a higher level of detail, a researcher can use the analysis software to outline areas of interest in an image. Once outlined, the eye gaze path between areas of interest can be automatically coded. However, this only works well for up to a few static images. If the participant is not viewing static images (e.g., they are scrolling through a webpage), then manually marking areas of interest becomes exceedingly time consuming because new markings must be made every time the image changes.

Fortunately, progress has been made in addressing these barriers. In terms of eye-tracking device cost, new lower cost technologies have been developed. The lower costs open new markets, such as applications in video games. These devices do not offer all the features of a research-grade eye-tracking device. For instance, the Tobii EyeX (cost: $139) does not have a fixed sampling rate and is not compatible with the Tobii Pro software suite. Nevertheless, it may be useful for some studies. The next section describes an automatic approach that uses an eye-tracking device to determine which patient data has been viewed by a user in an EMR interface.

2.2.3 Automatic eye gaze point-to-graphical element mapping

As stated above, mapping or annotating the data from an eye-tracking study is a big barrier to its use. Two studies have addressed this issue for studies of web-based systems by developing methods for automatically mapping the eye gaze data from an eye-tracking device to graphical elements of a website [98,99]. The WebEyeMapper and WebLogger system [98] records both the eye gaze data from a remote eye-tracking device and a detailed event log of a participant's web browsing session (which is called browser instrumentation). After the recording session, the eye gaze data are converted into fixations and mapped to the graphical elements that were present at each time point throughout the session. The WebGazeAnalyzer system [99] functions in a similar manner, but is also able to map eye gaze onto individual lines of text.

Besides automatic mapping, both the WebEyeMapper and WebLogger system and the WebGazeAnalyzer system provide exact playback of each study session. Exact playback is useful when the research team is interested in replaying the study session (e.g., Wright et al. [100]); however, if playback is not required, then a less detailed browsing log will suffice. The less detailed log stores the names and locations of each onscreen element at every page refresh, but not images or video of the elements. Gaze data are then mapped to the stored element locations to determine the names of elements being viewed. Additional automatic gaze mapping details are provided in Section 4.2.1.

2.3 CLINICAL ENVIRONMENT

We intend for the LEMR system to be applicable in virtually any clinical environment. However, to focus the early development and evaluation, this dissertation concentrates on a single clinical environment and two common clinical conditions in that environment, namely, acute respiratory failure and acute kidney failure.

2.3.1 Intensive Care Unit

Patient care in the ICU is complex, has large amounts of data per patient, and involves time-pressured decision-making. About six million adults are admitted to ICUs each year in the United States and one in five Americans who die, do so while in the ICU [101]. Information overload is a problem in this environment. A Canadian study estimated that the care of critically ill patients in the ICU generates a median of 1,348 individual data points per day [21]. Such an environment is the ideal location to investigate a LEMR system [33].

2.3.2 Acute respiratory failure

Acute respiratory failure (ARF) accounts for 25–40 % of ICU admissions and carries a mortality rate of 30% or more [102]. Common types of ARF include those caused by disorders of the airways (e.g. chronic obstructive pulmonary disease) and those caused by disorders of the alveoli (e.g. pneumonia). Mechanical ventilation is used to support patients with ARF.

2.3.3 Acute kidney failure

There are more than 200,000 cases of acute kidney failure (AKF) in the United States each year. It is a common complication in critically ill patients [103] and has a mortality rate of approximately 50% [104]. AKF usually develops due to kidney injury caused by toxins or reduced blood flow [105]. Treatment of patients with AKF includes a limited diet, diuretics, and dialysis.

2.4 PATIENT DATA

This section describes the patient data used in this dissertation. It includes a description of (1) the full data set of de-identified patient cases from which we extracted ARF and AKF cases and (2) the representation of that data in the LEMR system.

2.4.1 HIDENIC data set

The HIgh DENsity Intensive Care (HIDENIC) data set is a comprehensive collection of EMR data on thousands of patients who were hospitalized in ICUs at the University of Pittsburgh Medical

Center (UPMC) from July 2000 through December 2014 [106]. HIDENIC contains structured data including demographics, physiological measurements collected at the bedside such as vital signs, laboratory tests, and medication and fluid administration records. These data are combined with unstructured data, from a clinical data warehouse [107], in the form of a variety of clinical text reports such as history and physicals, progress notes, operative and procedure notes, and radiology, EKG, and EEG reports.

HIDENIC data was extracted from an ICU EMR system. It is a limited data set where actual calendar dates were retained and all other patient identifying information were removed. The data have been prepared for research use, including the mapping of medications and laboratory tests to standard terminologies.

2.4.2 Representing a patient state

As defined in Section 2.1.2, a clinical context includes all the electronically available information about a patient case. HIDENIC contains patient data from admission until discharge. In the LEMR system, patient cases are time sliced into successive days during the patient's stay in the ICU. Each successive day includes selected patient data that was available from the patient's day of ICU admission until a selected day *t* (see Figure 9).

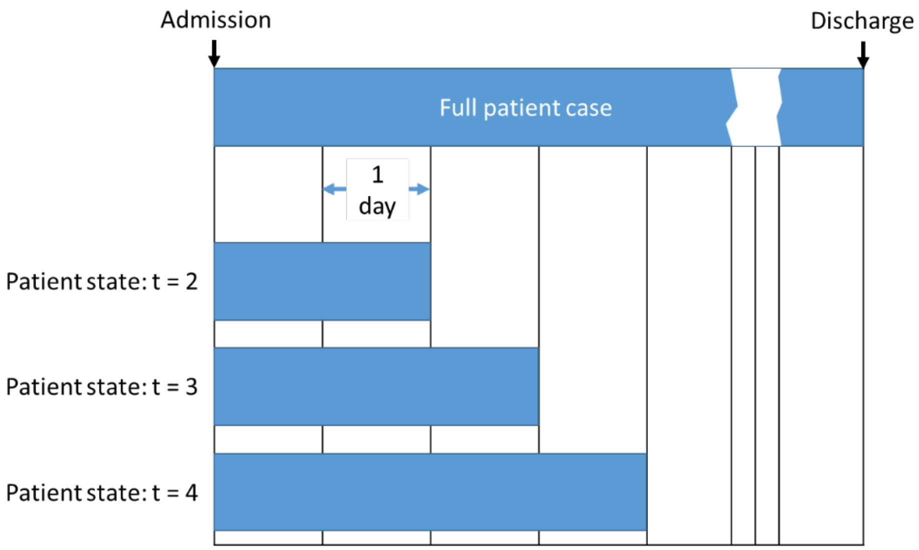


Figure 9. Successive days of a patient case are time sliced into patient states. A patient state includes all the available patient data from day of admission until selected day *t*.

The data available in HIDENIC includes both atemporal variables (e.g., demographics and co-existing medical conditions) and temporal variables (e.g., time-stamped laboratory results, medication administrations, and procedures). To create a uniform representation of each patient state, we use a vector space representation of these variables. Atemporal variables, such as gender, are included as their scaler values. Temporal variables are summarized using a combination of features. For instance, each scalar laboratory test is represented with up to 36 features including maximum value, most recent value, and slope between the two most recent values. Medications are represented with up to four variables that describe details such as whether the medication is currently active and how long it has been active. Procedures are represented in a similar manner to medications. The full list of features is available in Section 6.2.1.

2.5 SUPERVISED MACHINE LEARNING METHODS

This section provides a brief, selective introduction to machine learning methods that are relevant for this dissertation. One type of machine learning is called supervised learning. In supervised learning, a model $P(\mathbf{y} \mid \mathbf{x})$ captures the relationships between predictor variables in **x** and target variables in **y**. To learn the relationships, a learning algorithm is supplied with a data set that consists of training samples that contain both predictor variables (**x**) and target variables (**y**) for each sample. After training, the model can estimate the probability of the target variables given the values of the predictor variables. For this dissertation, the predictor variables are a vector space representation of a clinical context (including data in the EMR of the current patient case being viewed) and the target variables are the data items (e.g., blood glucose levels) that a clinician is predicted to seek as relevant while viewing the current case. We apply a machine-learned model to predict and highlight the patient data that a clinician is expected to seek in a current clinical case.

In developing a LEMR system, we collected a large training data set that includes many patient cases with target values assigned according to clinician information seeking behavior. We use this data set to train and test the performance of three supervised machine learning algorithms: logistic regression, support vector machine, and random forest.

Logistic regression is a regression model that uses a logistic function to predict the probability of a discrete target variable (traditionally, a binary variable) [108]. Regression models use an update function called gradient decent to update model parameters and reduce error when modeling the training cases. As described in Section 2.4.2, patient states (which are part of a clinical context) are represented with many variables. For logistic regression, we address the issue of high dimensionality using Lasso. The Lasso technique drives the weights of variables with little or no predictive value toward zero.

Support vector machine is a technique that maximizes the separation of two classes by a hyperplane and is usually well suited for classification problems with many predictor variables — high dimensionality [109]. The probability of the target can be derived using the distance of predictive features (in a high dimensional space) from the hyperplane.

Finally, random forest classifiers combine the output of multiple decision (classification) trees to predict the probability of the target variable. Decision trees use a tree-like structure to model a relationship between predictor variables and a target variable [110].

2.6 GAPS IN PRIOR WORK

This section summarizes the gaps in prior work that we see as impeding the development of LEMR systems.

Gap A. Clinical coverage of context-aware EMR systems will be limited if they use rule bases as their only mean of determining interface adaptions. To increase coverage of context-aware EMR systems across the wide range of clinical contexts, we need to supplement or replace expert-driven, rule-based systems with data-driven systems.

Gap B. To switch from expert-driven to data-driven context-aware EMR systems, we need a method for large scale observation of clinician information seeking behavior during EMR system use. It is not known how best to observe and collect this type of data, but a set of this data is needed for LEMR system development.

Gap C. Eye-tracking is one method for observing clinician information seeking behavior; however, eye-tracking devices used for EMR system observation studies in the past are expensive and mapping eye gaze to elements of the EMR interface is time consuming. It is not known how well inexpensive devices and automatic eye-gaze-to-graphical-element mapping will work with an EMR interface.

Gap D. It is not known which machine learning methods work best for modeling clinician information seeking behavior.

Gap E. Since data-driven context-aware EMR systems — like the LEMR system — are novel, it is not known how well models can predict clinician information seeking behavior, nor what the impact will be of clinicians using EMR systems that highlight relevant patient data.

2.7 CONTRIBUTIONS

This dissertation investigates the following questions:

A. **How do we develop context-aware EMR systems that are data-driven rather than expert-driven?**

   A data-driven approach to developing context-aware EMR systems requires data on clinician information seeking behaviors for many different patient cases, these data need to be modeled, and the models applied to direct the future display of patient data. This dissertation explores each of these tasks.

B. **What approaches are best able to observe clinician information seeking behavior?**

   The proposed LEMR system uses supervised machine learning methods. Supervised machine learning requires a training data set consisting of predictor variables and target variables.

Predictor variables are constructed from patient data. Target variables are assigned values based on observations of clinician information seeking behavior. There are several possible methods to obtain clinician information seeking behavior. For each method, we must consider the tradeoffs between observation accuracy and obtrusiveness on the clinician. One method is to ask a clinician (during training sessions) to manually select the patient data they seek in a clinical case. Another approach is to use eye-tracking technology. Eye-tracking devices can be used to estimate the patient data a clinician is viewing. If we assume that clinicians dwell longer in viewing information that they seek, then eye-tracking could be applied to automatically estimate clinician information seeking behavior.

C. Will an inexpensive eye-tracking device and an automatic gaze mapping method work with an EMR interface?

The use of inexpensive eye-tracking devices and automatic gaze mapping methods in conjunction with the LEMR interface are developed and evaluated in this dissertation. This advancement is a step towards allowing eye-tracking for clinical decision support, which will enable a plethora of new capabilities.

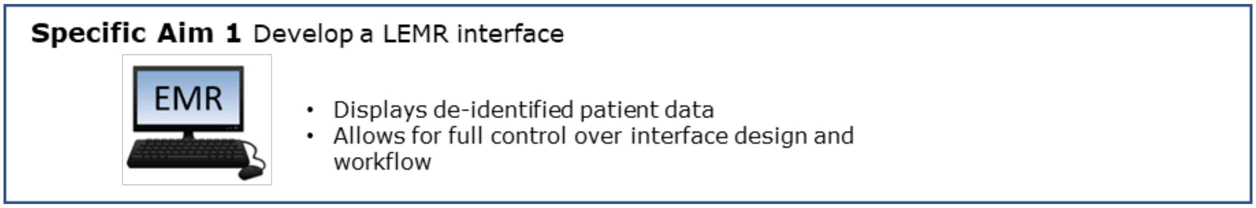
D. Among a set of state-of-the-art machine-learning methods, which is the best method to apply to model clinician information seeking behavior when preparing for morning rounds?

Little existing work has been done to understand which supervised machine learning models and predictor variables work well when predicting clinician information seeking behavior. Research that helps answer this question is important for advancing LEMR system development.

E. Will the LEMR system reduce the time it takes for clinicians to review patient information, without reducing their awareness of the case relevant information (in a given clinical context, such as preparing to present a patient case at morning rounds)?

Ideally, the LEMR system will reduce cognitive load on clinicians and result in reduced occurrence of preventable medical errors. Since cognitive load is difficult to directly measure, we instead focus on measuring the time it takes for a clinician to perform a given clinical task, namely, time to task completion. *Tasks*, such as preparing to present a patient case at morning rounds, are common clinical activities and reducing time to task completion frees up more clinician time for other care activities (patient interaction, consideration of how best to treat the patient, etc.). It may also result in greater clinician satisfaction with the EMR system. In addition to measuring time to task completion, we will also measure the extent to which the patient data highlighted in the LEMR system for each case includes all the data that clinicians self-report as seeking in those cases in preparing for morning rounds.

3.0 DEVELOPING A LEMR INTERFACE



Specific Aim 1 Develop a LEMR interface

- Displays de-identified patient data
- Allows for full control over interface design and workflow

This chapter describes the design, development, and implementation of the LEMR interface that was used in the experiments presented in this dissertation. It includes the rationale for developing a LEMR interface, initial requirements, preliminary design and evaluation, and the primary interface design used for the primary studies (Sections 5.3 and 7.1) presented in this dissertation.

3.1 RATIONALE FOR DEVELOPING A LEMR INTERFACE

To develop a LEMR system, we needed a way of displaying patient data, that is, the display capability of an EMR system. We also needed the ability for a clinician user to select the patient data that he or she sought in each clinical context. It would have been ideal to use the EMR system that study participants used for clinical activities (i.e., the local EMR system). First, using the local system would eliminate a burn in period, which is the time a participant spends learning and becoming familiar with the system within the experimental setup. The downside to a burn in period is that it is time spent without collecting the principal data. Second, using the local system would increase the external validity of the results. Unfortunately, the local system was not available for adaptation.

We considered modifying open source EMR software for study purposes. The two systems we reviewed were the VistA system, which is used in all Veteran Affairs Hospitals [111], and the OpenMRS system, which is used in many low resource settings [112]. VistA has the benefit of being a fully functional EMR system; unfortunately, it is implemented in MUMPS, which is a programing language that has fallen out of popular use. Furthermore, the VA is in the process of replacing Vista. OpenMRS, on the other hand, is growing in usage [113]. However, it lacks much of the functionality of commercial EMR systems that are in current use. Ultimately, we decided that it was simpler to build an in-house system that has sufficient EMR display functionality to support our proposed research. An advantage of doing so is that we have deep understanding and complete control of the system and user interface, allowing for rapid prototyping and addition of functionalities as needed.

3.2 LEMR INTERFACE REQUIREMENTS

Once the decision was made to build an in-house LEMR interface, we developed a minimum set of software requirements for the LEMR system. These requirements were simply (1) screens to display patient data, (2) ability for clinicians to manually record their information seeking behavior, and (3) capability to highlight patient data that clinicians are predicted to seek as relevant.

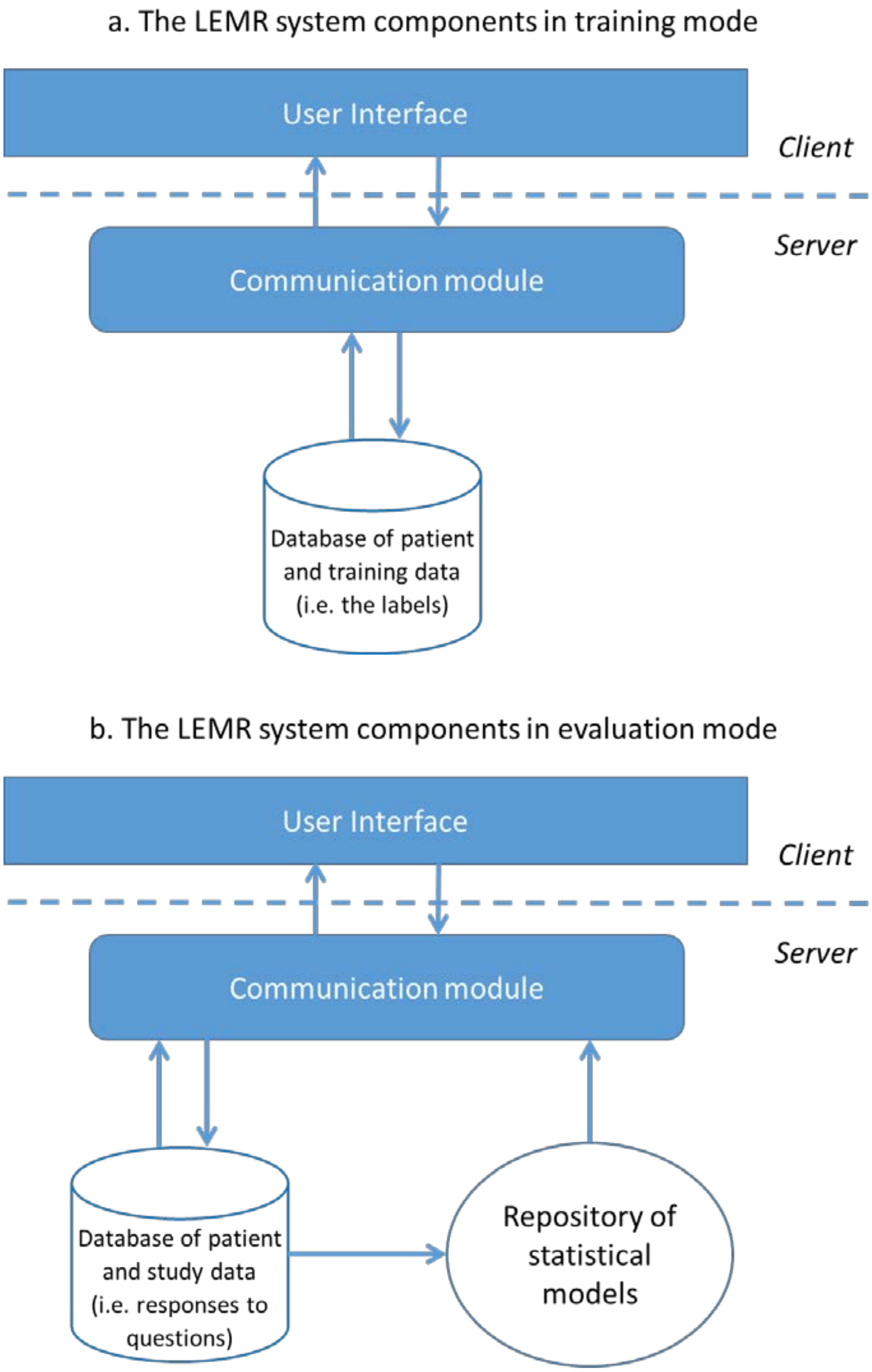
3.3 PRELIMINARY LEMR INTERFACE DESIGN

To address the first requirement, we built the graphical user interface as a web browser application. Applications that use the web browser have gained popularity in recent years because of various advantages including platform independence, the ability to access the system from anywhere, adaptability to mobile applications, and relative ease of development and maintenance.

The second requirement follows from the need to have training data to train a LEMR system. A target variable (e.g., glucose target) is assigned the value *yes* when that variable's corresponding data item (e.g., blood glucose level) is recorded for a patient case and a clinician seeks it. To collect a large training data set, the method for observing clinician information seeking behavior must not be overly burdensome for the clinicians providing the target values. In this preliminary LEMR interface, the simplest observation method was for clinicians to manually select (by clicking on) the patient data that they sought in each patient case.

For the third requirement, the LEMR interface needed a clear and intuitive way to indicate which patient data are highlighted in each patient case. The use of web components and styling provide multiple easy options for highlighting.

The LEMR system is implemented as client-server software and consists of four components. The client consists of a user interface and the server consists of three components that include a database for storing patient and other data, a repository of statistical models, and a communication module that connects the client with the server. The LEMR system can function in two modes. In the *training* mode, the system enables a user to select data items that are pertinent to a task for the current patient case (see Figure 10a). In the *evaluation* mode, the system highlights data items that are predicted to be sought by the user and collects user responses to study questions (see Figure 10b).


a. The LEMR system components in training mode

b. The LEMR system components in evaluation mode

User Interface
Client
Server
Communication module
Database of patient and training data (i.e. the labels)
User Interface
Client
Server
Communication module
Database of patient and study data (i.e. responses to questions)
Repository of statistical models

Figure 10. Components of the LEMR system.

On the server side, patient cases are stored in a MySQL *database*, which is queried to provide data shown in the user interface. The same database stores information seeking activities that are captured during the training mode and responses to study questions during the evaluation mode. The *repository* of statistical models contains predictive models that are derived offline using Scikit-learn [114]. During the evaluation mode they are applied to a current patient case to predict data items that are likely to be sought by the user. The communication module is implemented in Django Web Framework [115] and enables the flow of data from the database to the user interface and vice-versa.

The LEMR *user interface* displays patient cases in a compact manner and is implemented in a web browser using HTML, CSS and JavaScript. A screenshot of the preliminary LEMR interface with highlighting is shown in Figure 11. Panel A, the patient demographics toolbar, allows the user to move between patients and gives a summary of the current patient's demographic information and admitting diagnosis. Panel B contains quick access tabs for navigating among the various types of patient information, including laboratory test results, medication orders, and clinical text reports (e.g., history & physical examination (H&P) notes, progress notes, and operative procedure notes). Currently the "Labs/Vitals/Meds" tab is selected. This EMR interface uses times-series plots to display this structured clinical information (laboratory test results, vital signs, medication orders, and intake and output data). Panel C, the time range selector, is used to define time ranges of data to display. Below the time-range selector is the procedures axis, which is labeled with the defined times. Black diamonds on this axis represent procedures (surgeries, biopsies, etc.) that the current patient has had. Hovering over a diamond gives more details on that procedure. Panel D, the highlighted information display (HID), shows detailed time-series plots of the highlighted patient data. These plots have a labeled y-axis and blue bands to indicate the normal range. Panel E displays all available results, including those found in the HID, using plots with condensed y-axes. These plots give a notion of trends over time and are arranged by group type (basic chemistry, cardiac chemistry, etc.). The buttons across the top of this panel list all the different group types and can be used to jump to a specific type. For both Panel D and Panel E, different colors are used to indicate when a value is within or outside of the normal range (blue = below; green = within; red = above; black = no defined normal range).

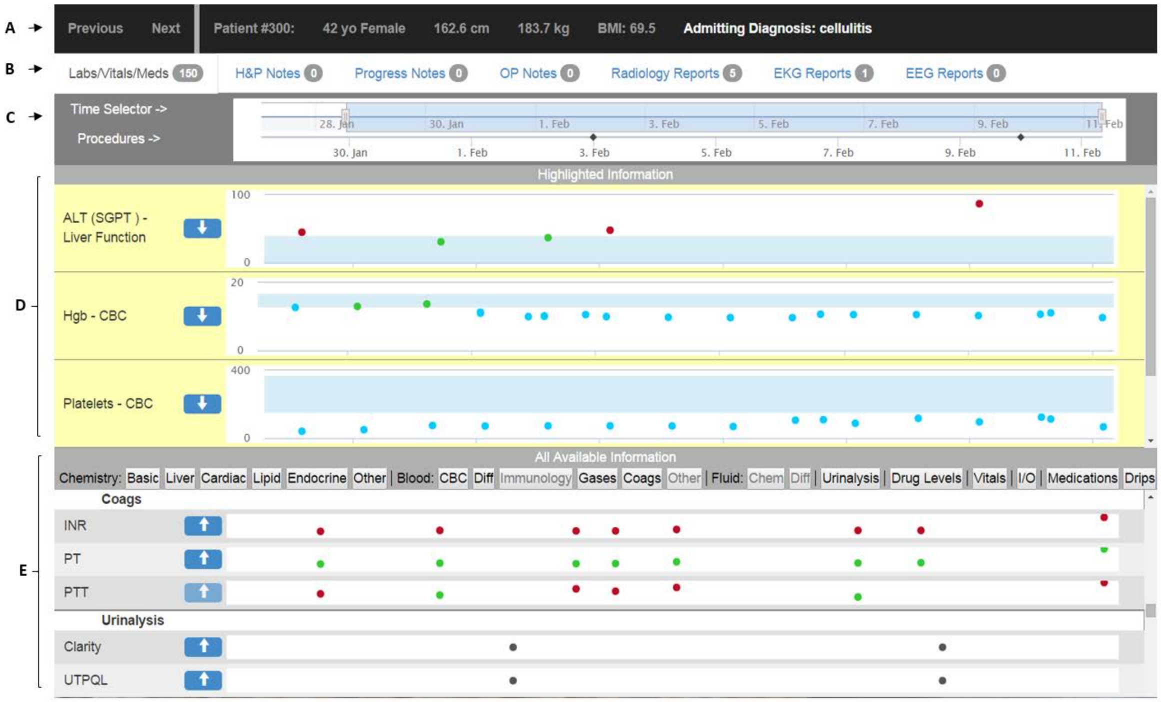


Figure 11. Screenshot of the preliminary LEMR interface. A) demographics toolbar; B) quick access tabs; C) time range selector; D) Highlighted Information Display (HID); E) all data display. Both D and E are scrollable.

For a patient case, the patient data highlighted in the HID are the patient data that a clinician is predicted to seek for a given clinical case. The HID can be populated with items by both automatic and manual means. For automatic population, the LEMR system uses stored statistical models to predict the probability that each data item (e.g. blood glucose levels) is going to be sought. The patient data that have predicted probabilities above a set threshold are placed in the HID. To make manual changes to the contents of the HID, a user clicks on the blue buttons (with white arrows) next to the name of each data item. There are buttons to move patient data into the HID and buttons to remove patient data from there. When a user adds new patient data to the HID, the change is captured in the training data set where a target value of *yes* is given to a target variable added to the HID and a target value of *no* is given to a target variable removed from the HID. Each target variable is assigned a maximum of one target value per patient case.

3.4 PRELIMINARY LEMR INTERFACE THINK ALOUD STUDY

During LEMR interface development, we met with ICU clinicians to elicit feedback on the interface and proposed LEMR system modeling strategy. One method we used to elicit feedback was a think aloud protocol [116]. In a think aloud study, an evaluator uses the interface to work through a series of tasks and, while doing so, they speak their thoughts out loud. Feedback from the think aloud study was used to refine the interface design.

3.4.1 Methods

We conducted a think aloud protocol with an ICU clinician — denoted by C1 — with the goal of identifying new ways to improve the LEMR system. This study took place in C1's office on 11/7/2014. The prototype LEMR interface was loaded with three de-identified ICU patient cases from the HIDENIC data set [106]. Patient data were highlighted based on feedback from a clinician who reviewed the same cases while being tasked to identifying changes in clinical condition or emergence of a new clinical problem (see Section 5.1). These highlighted patient data would appear in the HID as if the models predicted them. C1 talked out load about his thoughts and actions while using the prototype to assess each patient's clinical condition.

3.4.2 Results

C1 expressed several concerns with the display of highlighted patient data. First, he did not like the mix of both colors and symbols when representing a test result value as abnormally low (blue squares), normal (green circles), or abnormally high (red triangles). Instead, he suggested the use of different colors (as was shown in Section 3.3, Figure 11). C1 also noted that the differing y-axes between highlighted and non-highlighted patient data was confusing. He suggested that they be made consistent and be free for adjustment by the user. Third, C1 believed that the data on drip medication concentrations should be displayed as histograms rather than scatter plots, as was done for other laboratory tests and medications. Finally, C1 discovered a software bug that caused the ranges of the y-axis in the graphs of some laboratory tests and medications to be inconsistent, with higher positions on the y-axis not necessarily corresponding to higher values.

Regarding other interface aspects, C1 expressed concern regarding how a clinician might interpret highlighted patient data. Specifically, he believed that clinicians almost always consider only one to three time-series of laboratory tests and medications at a time. As a result, C1 suggested that the LEMR should have an option whereby tests and medications are grouped and then displayed in separate tabs. Moreover, he suggested that another window be incorporated into the LEMR system design, whereby clinicians can explore correlations between test results and medications without explicitly moving them into the HID.

Two comments were made about the proposed LEMR system modeling. C1 wanted to ensure that the models were not going to be weighted too strongly towards one's own information seeking behavior. Doing so could reinforce the potential biases of a clinician.

Second, C1 suggested that the LEMR should incorporate time series outlier detection into its models. Specifically, the system should be able to classify whether a time series is normal or abnormal within the last n (user specified) hours. Outlier detection could be used as a standalone method for identifying data to highlight or could be incorporated as predictor variables in training data sets for a LEMR system.

This study provided important insights into issues an ICU clinician may have when using this prototype. We updated the LEMR interface based on the suggested interface changes and clarifications. The modeling suggestions are tabled for future consideration and study. Next, we performed a multi-participant usability study on the updated LEMR interface.

3.5 PRELIMINARY LEMR INTERFACE USABILITY STUDY

We performed a usability study to gather usability data in the LEMR interface from multiple ICU clinicians, while they performed a simulated patient review task. In addition to identifying strengths and weaknesses of the LEMR interface, we asked each participant for their thoughts on the concept of a LEMR system.

3.5.1 Methods

Four medical fellows were recruited from UPMC's Department of Critical Care Medicine to participate in the study. The study was approved by the University of Pittsburgh Institutional Review Board (ID PRO14020588). It took place in meeting rooms at the University of Pittsburgh School of Medicine in February of 2014. Each participant used the prototype to review three to five selected patient cases. For each case, participants were shown the patient's EMR data from ICU admission to a selected day during that patient's ICU hospitalization. The participants were asked to familiarize themselves with the case as if they were the attending clinician. No data were displayed in the HID for the initial examination of each patient case. Next, the clinicians were shown an additional day of the patient's data that was meant to simulate rounding on the subsequent day. For each case's additional day, the HID was populated with laboratory tests that a clinician on the research team predetermined to be useful when identifying a change in the current patient's clinical condition or emergence of a new clinical problem. The study participants were asked to use the features of the prototype to add and remove items from the HID until the highlighted items represented the patient data that they thought another clinician who was looking at the same case would want to use when assessing the last 24 hours of that case, given that they had been following the patient since ICU admission.

During the review of each patient case, screen tracking software recorded all the on-screen actions and an audio recording captured each participant's think aloud comments. After a participant reviewed the allotted patient cases, additional time was allocated for a semi-structured interview. We asked the participants about their perceptions regarding the LEMR system concept in general and the LEMR interface specifically. The questions are listed in Table 3. The interviews were coded independently by two researchers before meeting to create a consensus. The coding was used to identify general themes in the responses. Each participant also completed the System Usability Scale [117] based on his or her interactions with the prototype.

Table 3. Interview questions used during a usability study of the LEMR interface.

1) Please share your thoughts on the premise behind this work, independent of the specific system that you just used.

    i) Please describe the applicability of this idea to your work, particularly in terms of clinical utility.

    ii) What would your reaction likely be if you were told that this technique was going to be integrated into your clinical work? Would you be enthusiastic? Worried? Why?

    iii) Do you have any concerns with this sort of approach, and if so, what are they?

    iv) What impact, if any, do you think such a system would have on quality of care?

    v) What impact, if any, do you think such a system would have on the amount of time that you spend rounding on a given day?

    vi) What factors might contribute to the success or failure of such a system?

2) The system that you used is an early prototype implementation of this approach. I'd like to get your impressions of it:

    i) What do you like about the prototype?

    ii) What do you not like about the prototype?

    iii) How could the prototype be improved?

    iv) Would you use such a system, if it were available to you?

3) Is there anything else that you might like to tell me?

3.5.2 Results

The participants identified many benefits and a few concerns of the LEMR system concept and LEMR interface. They were enthusiastic about the concept of a LEMR system and thought that designing something that utilizes current behaviors is important. They thought that a fully developed LEMR system would probably improve the quality of care. Three participants identified as positive the LEMR's potential to adapt to different specialists and thought it was applicable because not all types of clinicians look at the same type of information. Reduction of information burden was also mentioned as important. One participant would like to use any system that is able to highlight the most relevant information without slowing him down. Three of the participants liked the timeline approach to displaying information.

One concern that participants had with the system was about feasibility. They thought that the ICU setting would be difficult because clinicians there must address every organ system and abnormality. Another concern was implications of integration into workflow. They warned that a system that focuses too much on commonly sought patient data could cause an over reliant clinician to miss out on rare things that happen. Finally, there were a few design concerns including the color scheme. Three different colors were used to represent low, normal, and high test results. One of the participants said that abnormal results should be the same color regardless of whether they are abnormally high or abnormally low (i.e., red for both). Also, two participants mentioned that they did not like having to hover over a data point to get an exact test result value.

The System Usability Scale composite score for the four participants was 79. The scale ranges from 0 to 100 and any score above 68 is generally considered to be above average usability [118].

This study provided important feedback on the usability of the LEMR interface and on the concept of a LEMR system. Design suggestions from the participating ICU clinicians, such as always showing exact test result values, were included in future versions of the LEMR interface. The positive interview responses highlighted the potential benefits of this line of work.

3.6 PRIMARY LEMR INTERFACE DESIGN

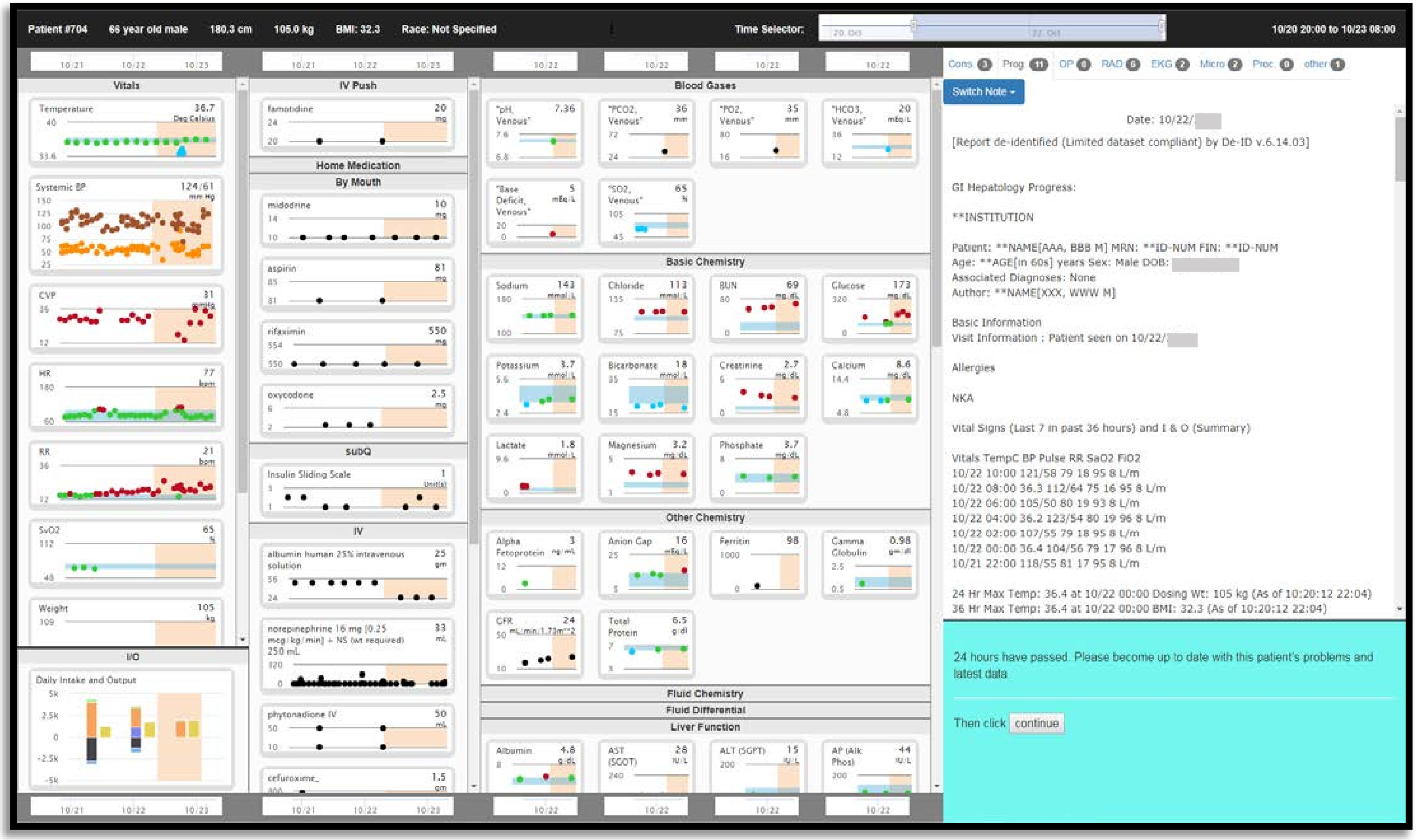
The LEMR interface was continuously improved in response to user feedback and experimental needs. This section describes the LEMR interface used during the primary collection of training data (Section 5.3) and during the primary LEMR evaluation study (Section 7.1).

Figure 12 is a screenshot of the LEMR interface that was used in the primary studies described in this dissertation. The single column of laboratory tests results, vital sign measurements, and medication orders from the preliminary LEMR interface was replaced with six columns: one column for vital sign measurements, ventilator settings, and input and output measurements; one column for medications; and four columns for the different laboratory test results, which are still organized by laboratory group. The right side of the screen is reserved for the different types of free text notes, reports, and a procedure list. On the lower right side of the screen, a blue box provides instructions to the participant on the current experimental task.

For the LEMR system evaluation study (Section 7.1), three different versions of the prototype interface were used: the *control* interface (shown in Figure 12) was used in all three arms of the evaluation study, the *highlights* interface (shown in Figure 13, top) was used in Arm 2 and Arm 3 of the evaluation study, and the *highlights-only* interface (shown in Figure 13, bottom) was used in Arm 3 of the evaluation study. On the highlights interface, patient data predicted to be sought as relevant by models of clinician information seeking behavior are highlighted (in-place) with a yellow background. On the highlights-only interface, patient data predicted to be sought as relevant are highlighted and patient data not predicted to be sought as relevant are hidden. The rational for these three versions of the interface is provided in the methods section of the evaluation study (Section 7.1.1).

Additional interface screenshots and feature descriptions are provided in Appendix D, which contains the slides of the presentation that introduced participants of the evaluation study to the study objectives and the LEMR interface.

In conclusion, we designed and developed a LEMR interface to use in a research setting. Beyond the research reported in this dissertation, this LEMR interface is readily adaptable to various areas of research, including clinical decision making, information needs, and human computer interaction. Open source software for the LEMR interface is available at (https://github.com/ajk77/LEMRinterface) or see Appendix B.



Date: 10/22/[redacted]

[Report de-identified (Limited dataset compliant) by De-ID v.6.14.03]

GI Hepatology Progress:

**INSTITUTION

Patient: **NAME[AAA, BBB M] MRN: **ID-NUM FIN: **ID-NUM
Age: **AGE[in 60s] years Sex: Male DOB: [redacted]
Associated Diagnoses: None
Author: **NAME[XXX, WWW M]

Basic Information
Visit Information : Patient seen on 10/22/[redacted]

Allergies

NKA

Vital Signs (Last 7 in past 36 hours) and I & O (Summary)

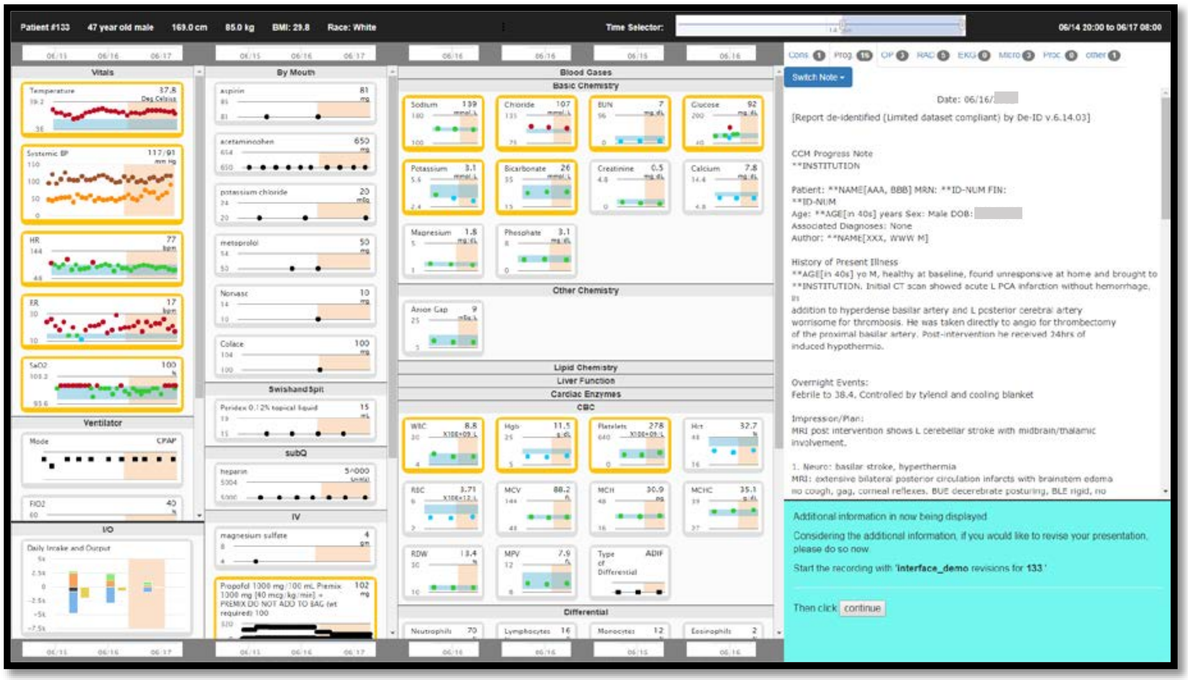
```
Vitals TempC BP Pulse RR SaO2 FiO2
10/22 10:00 121/58 79 18 95 8 L/m
10/22 08:00 36.3 112/64 75 16 95 8 L/m
10/22 06:00 105/50 80 19 93 8 L/m
10/22 04:00 36.2 123/54 80 19 96 8 L/m
10/22 02:00 107/55 79 18 95 8 L/m
10/22 00:00 36.4 104/56 79 17 96 8 L/m
10/21 22:00 118/55 81 17 95 8 L/m
```

24 Hr Max Temp: 36.4 at 10/22 00:00 Dosing Wt: 105 kg (As of 10:20:12 22:04)
36 Hr Max Temp: 36.4 at 10/22 00:00 BMI: 32.3 (As of 10:20:12 22:04)

24 hours have passed. Please become up to date with this patient's problems and latest data

Then click continue

Figure 12. Primary LEMR interface.



Date: 06/16/[redacted]

[Report de-identified (Limited dataset compliant) by De-ID v.6.14.03]

CCM Progress Note
**INSTITUTION

Patient: **NAME[AAA, BBB] MRN: **ID-NUM FIN:
**ID-NUM
Age: **AGE[in 40s] years Sex: Male DOB: [redacted]
Associated Diagnoses: None
Author: **NAME[XXX, WWW M]

History of Present Illness
**AGE[in 40s] yo M, healthy at baseline, found unresponsive at home and brought to
**INSTITUTION. Initial CT scan showed acute L PCA infarction without hemorrhage,
in
addition to hyperdense basilar artery and L posterior cerebral artery
worrisome for thrombosis. He was taken directly to angio for thrombectomy
of the proximal basilar artery. Post-intervention he received 24hrs of
induced hypothermia.

Overnight Events:
Febrile to 38.4. Controlled by tylenol and cooling blanket

Impression/Plan:
MRI post intervention shows L cerebellar stroke with midbrain/thalamic
involvement.

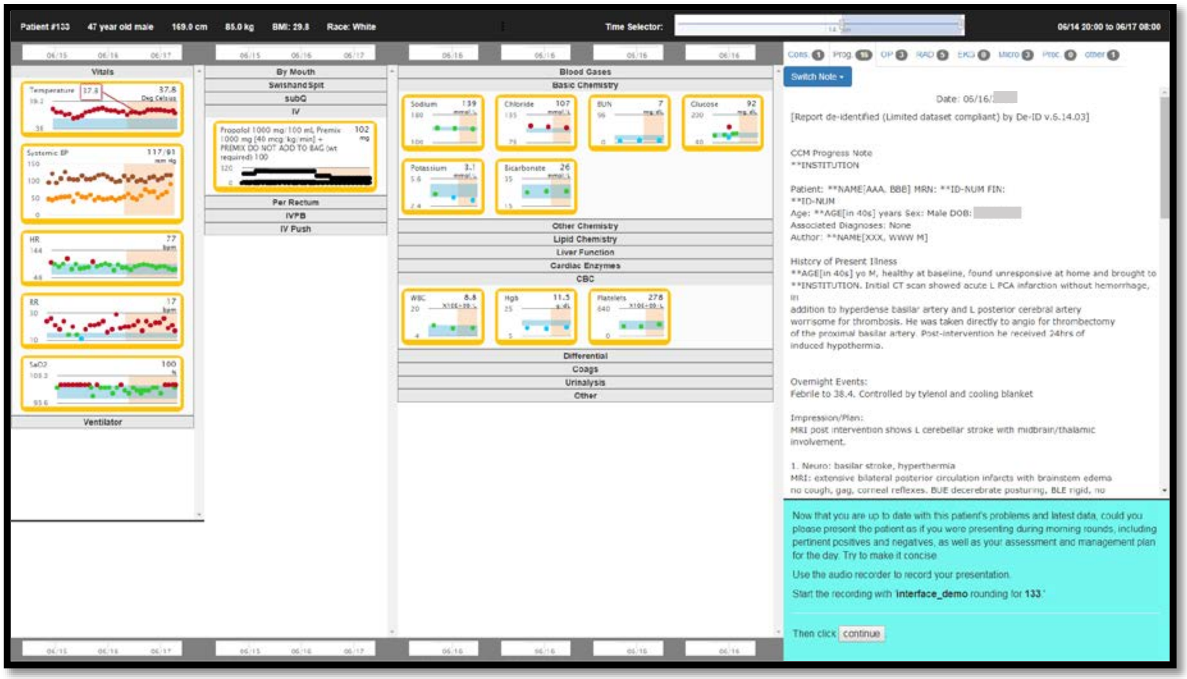
1. Neuro: basilar stroke, hyperthermia
MRI: extensive bilateral posterior circulation infarcts with brainstem edema
no cough, gag, corneal reflexes. BUE decerebrate posturing, BLE rigid, no

Additonal information in now being displayed

Considering the additional information, if you would like to revise your presentation, please do so now.

Start the recording with '**interface_demo** revisions for **133**.'

Then click continue



Date: 06/16/[redacted]

[Report de-identified (Limited dataset compliant) by De-ID v.6.14.03]

CCM Progress Note
**INSTITUTION

Patient: **NAME[AAA, BBB] MRN: **ID-NUM FIN:
**ID-NUM
Age: **AGE[in 40s] years Sex: Male DOB: [redacted]
Associated Diagnoses: None
Author: **NAME[XXX, WWW M]

History of Present Illness
**AGE[in 40s] yo M, healthy at baseline, found unresponsive at home and brought to
**INSTITUTION. Initial CT scan showed acute L PCA infarction without hemorrhage,
in
addition to hyperdense basilar artery and L posterior cerebral artery
worrisome for thrombosis. He was taken directly to angio for thrombectomy
of the proximal basilar artery. Post-intervention he received 24hrs of
induced hypothermia.

Overnight Events:
Febrile to 38.4. Controlled by tylenol and cooling blanket

Impression/Plan:
MRI post intervention shows L cerebellar stroke with midbrain/thalamic
involvement.

1. Neuro: basilar stroke, hyperthermia
MRI: extensive bilateral posterior circulation infarcts with brainstem edema
no cough, gag, corneal reflexes. BUE decerebrate posturing, BLE rigid, no

Now that you are up to date with this patient's problems and latest data, could you please present the patient as if you were presenting during morning rounds, including pertinent positives and negatives, as well as your assessment and management plan for the day. Try to make it concise.

Use the audio recorder to record your presentation.

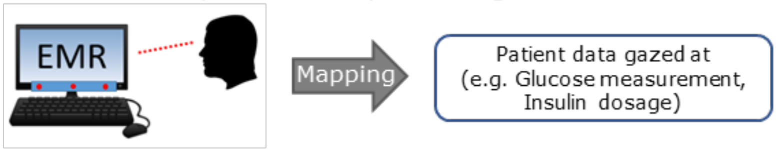
Start the recording with '**interface_demo** rounding for **133**.'

Then click continue

Figure 13. Primary LEMR interface with model-based highlighting. The top screenshot shows the *highlights* version of the LEMR interface, which has in-place, yellow highlighting of patient data. The bottom screenshot shows the *highlights only* version of the interface, in which only the highlighted patient data appears.

4.0 DEVELOPING AUTOMATIC EYE-TRACKING FOR THE LEMR INTERFACE

Specific Aim 2 Develop automatic eye-tracking for the LEMR interface



This chapter presents two studies that explore the use of eye-tracking technologies in clinical decision support [119]. The first study evaluates the use of an inexpensive eye-tracking device; most prior work has used expensive devices. The second study evaluates an automatic method for analyzing eye-tracking data for use in machine learning; manual mapping of gaze points to graphical elements in an interface is time consuming and often impractical to perform. The results from these studies provide support for the use of eye-tracking technologies in the clinical setting for observing how clinicians use the EMR and for recording their information seeking behavior.

4.1 EVALUATING AN INEXPENSIVE EYE-TRACKING DEVICE

As discussed in Section 2.2.2, eye-tracking devices designed for research are expensive. Newer devices designed for video games are far less expensive than older eye-tracking systems. In the context of the LEMR system, if eye-tracking proves to be an acceptable approach for observing clinician information seeking behavior, a less expensive device makes widespread adoption much more likely. We conducted a study to determine if an inexpensive eye-tracking device has acceptably similar accuracy when compared to a more expensive device designed for researchers. More specifically, our hypothesis was that the accuracy of the inexpensive Tobii EyeX device ($139 in March 2016) is not inferior to the moderately priced Tobii X2-30 eye-tracking device ($4,900 in March 2016).

4.1.1 Methods

We recruited one undergraduate student, seven graduate students and two post-doctoral researcher associates, to participate in a study that took place in March of 2016. Four of the participants wore corrective lenses (glasses), five had uncorrected eyesight, and one had corrective eye surgery. One participant who wore corrective lenses was excluded from the study due to difficulty in calibrating the eye-tracking device.

Each participant took part in two experiments where an experiment used one of two eye-tracking devices. For each experiment, the participant was instructed to sit in front of a computer monitor that was equipped with one of the eye-tracking devices. The monitor was adjusted to ensure that the participant was comfortable and the eye-tracking device had a clear view of the participant’s eyes. Once situated, the participant used the six-point Tobii EyeX Engine calibration program to calibrate the device. Next, the participant was instructed to stare at a small (7x7 pixel) red box as it appeared for one-second durations in 50 random onscreen locations. Then, we switched to the other eye-tracking device for the second experiment. The order of the devices varied by participant, where half was tracked by the inexpensive EyeX first and half was tracked by the more expensive X2-30 first.

Data collected during the study included the gaze points measured by the eye-tracking devices and the onscreen coordinates of each randomly generated box. We used this data to calculate the absolute error between the median location of all the gaze points measured while a box was onscreen and the coordinates of that box. Absolute error is used because an average of non-absolute (positive and negate) errors would underestimate the true error. We report the average error of each trial run and compare the errors of the two eye-tracking devices using a paired sample t-test.

The inexpensive device is considered to be not inferior to the moderate-cost device if the upper bound of the 95% confidence interval of the difference in error (inexpensive device minus moderate-cost device) is no greater than one percent of screen height, which is approximately 11 pixels. A difference of this magnitude could be accounted for with a slight increase in the size of each graphical element displayed in the LEMR interface (see Chapter 3). If the 95% confidence interval includes values greater than 11 pixels, then each graphical element would need to be increased by a larger amount, resulting in a loss of information display density that could compromise the utility of the interface. In this situation, we would conclude that the inexpensive device was not as accurate.

4.1.2 Results

For each participant, the average error was calculated both in two dimensions (diagonal error) and in single dimensions (horizontal error and vertical error). Results are shown in Table 4 and in Figure 14. Using a two-sided paired sample t-test, there was not a statistically significant difference between the error of the two eye-tracking devices in either the vertical or the diagonal directions (p-values: 0.313 and 0.768, respectively). The upper bounds of the 95% confidence intervals for the difference show that the average error for the lower cost device is likely no more than 9 pixels greater in the vertical direction and 5 pixels greater diagonally — magnitudes that are less than one percent of screen height. There was a statistically significant difference in the horizontal error; however, it was the inexpensive device that had less error than the more expensive device (p-value: 0.008).

These results support the claim that the inexpensive EyeX device is not inferior to the more expensive X2-30 device. Other than accuracy, the more expensive device has features that may be desirable to other investigators. It has a consistent refresh rate of 30 Hz, as opposed to the Tobii EyeX's inconsistent refresh rate of about ~58 Hz. The X2-30 is also compatible with Tobii Pro Studio, which offers a wide range of data analyses. Nevertheless, for the experiments described in this dissertation, the EyeX device is a good choice.

Table 4. Average errors of two eye-tracking devices. Each error cell is the average of absolute median errors across fifty gaze points for each participant.

	Horizontal error (in pixels)		**Vertical error (in pixels)**		**Diagonal error (in pixels)**	
Participant	**EyeX**	**X2-30**	**EyeX**	**X2-30**	**EyeX**	**X2-30**
1	8	9	21	10	23	15
2	13	17	32	17	36	27
3	9	10	16	17	19	22
4	10	21	21	19	24	30
5	16	15	22	12	30	21
6	5	10	14	11	16	16
7	9	14	12	14	16	22
8	8	14	16	22	19	29
9	11	16	14	21	20	28
Average	**9.9**	**13.9**	**18.6**	**15.7**	**22.6**	**23.3**
Difference (95% CI)	-4 (-6.8, -1.4)		2.9 (-3.2, 8.7)		-0.7 (-6.6, 5.1)	

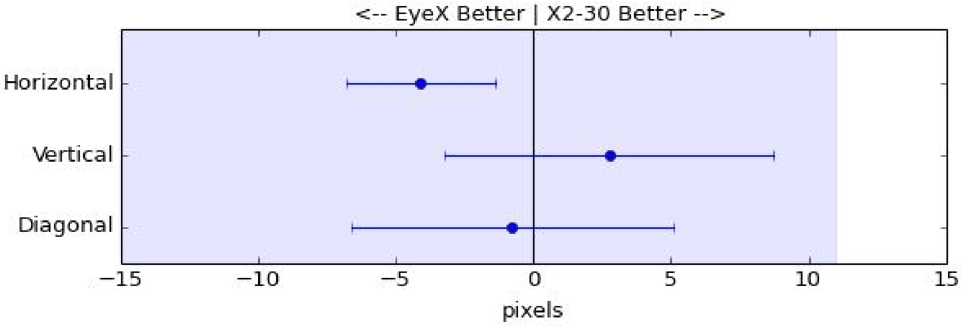


Figure 14. Difference in error of two eye-tracking devices (EyeX minus X2-30). Error bars indicate two-sided 95% confidence intervals. The shaded area indicates error values below the non-inferiority margin (11 pixels). Since, the upper limit of each error bar is below the non-inferiority margin, the data support that the EyeX device is not inferior to the X2-30 device.

4.2 EYE-TRACKING FIDELITY TEST

To use the data produced by the Tobii EyeX for machine leaning, we developed and implemented a method to automatically map eye gaze coordinates produced by the device to graphical elements displayed in the LEMR interface. Section 2.2.2 describes mapping methods that utilized web browser implementations. Since our intended use of the eye-tracking data does not require video playback of study sessions, we implemented a limited web browser method that stores less information. It uses a mapping algorithm that ranks patient data by the amount of gaze that they have received from the user. The assumption is that the longer a graphical element is cumulatively viewed in the interface, the more likely that element was sought by the clinician. Our hypothesis was that when a study participant uses the LEMR interface to answer a question about a patient case, the patient data the participant had to look at to correctly answer the question will be ranked as most gazed upon by the mapping algorithm.

4.2.1 Methods

In this section, we describe an automatic gaze point-to-graphical element mapping method and its evaluation.

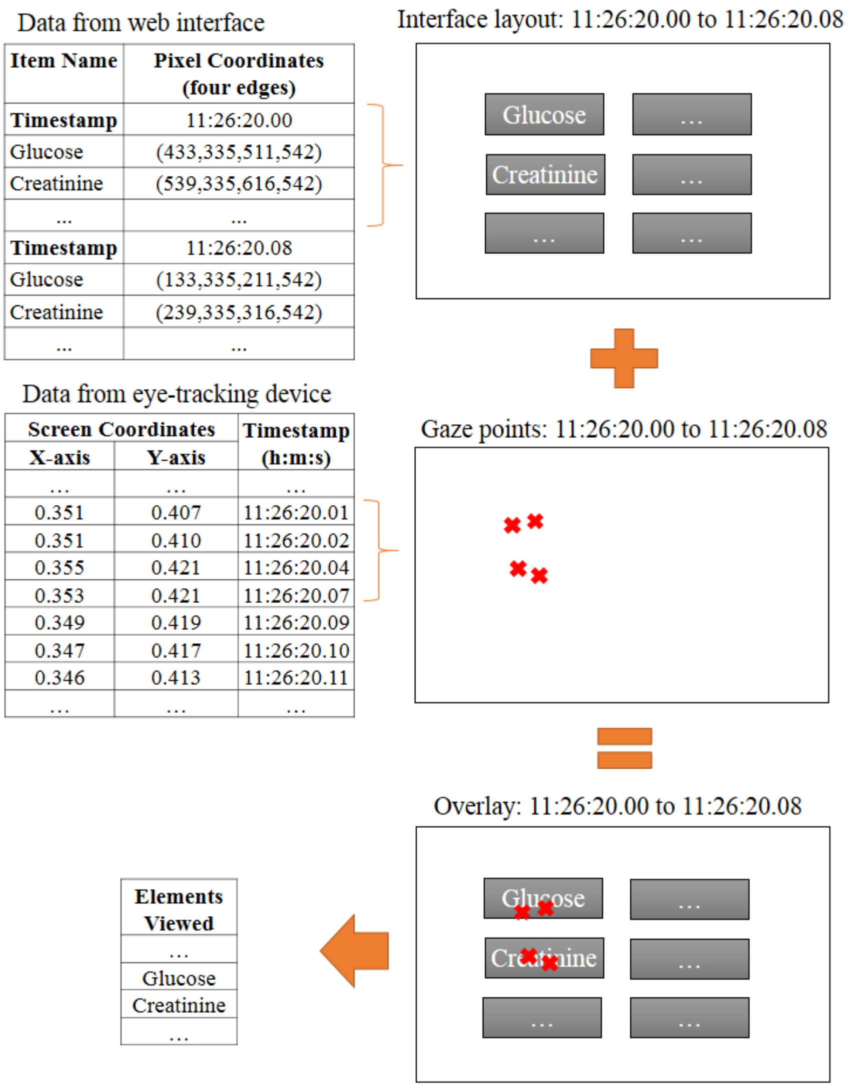
Automatic gaze point-to-element mapping

We developed an easy-to-use automatic eye gaze point-to-graphical element mapping method that stores minimal information about the onscreen interface layout (i.e., it stores the names and locations of each graphical element that depicts patient data). On each page refresh, the method uses the JavaScript function getBoundingClientRect() to determine the location of each element;

element names, locations, and a timestamp are recorded in a text file. Simultaneously, the data streaming from the eye-tracking device (x-coordinate, y-coordinate, and timestamp) are recorded in a second text file. Next, these files are overlaid using the timestamp information (as shown in Figure 15). Once this overlay is made, we calculate the mapping by counting the number of gaze points that fall within each graphical element across time. We call this the Gaze Point (GP) method.

Eye-gaze-to-graphical-element mapping via the GP method does not account for the error of the eye-tracking device. To account for this error, we developed a probability distribution-based approach, the Distributed Gaze Point (DGP) method, which allocates a portion of each gaze point to each of the elements that lie within the surrounding 100x100 pixel area. Allocations are made based on a bivariate normal distribution that was fit to the error of the eye-tracking device. Therefore, the portion of a gaze point that is allocated to an element is an estimate of the probability the participant was actually viewing that element. We rank the viewed elements by the sum of the gaze probabilities allocated to the element across an interaction.

Open source eye gaze tracking and analysis software is available online at (https://github.com/ajk77/EyeBrowserPy) or see Appendix B.



Data from web interface

Item Name	Pixel Coordinates (four edges)
Timestamp	11:26:20.00
Glucose	(433,335,511,542)
Creatinine	(539,335,616,542)
...	...
Timestamp	11:26:20.08
Glucose	(133,335,211,542)
Creatinine	(239,335,316,542)
...	...

Data from eye-tracking device

Screen Coordinates		Timestamp
X-axis	Y-axis	(h:m:s)
…	…	…
0.351	0.407	11:26:20.01
0.351	0.410	11:26:20.02
0.355	0.421	11:26:20.04
0.353	0.421	11:26:20.07
0.349	0.419	11:26:20.09
0.347	0.417	11:26:20.10
0.346	0.413	11:26:20.11
…	…	…

Elements Viewed
…
Glucose
Creatinine
…

Interface layout: 11:26:20.00 to 11:26:20.08
Glucose
…
Creatinine
…
…
…
Gaze points: 11:26:20.00 to 11:26:20.08
Overlay: 11:26:20.00 to 11:26:20.08
Glucose
…
Creatinine
…
…
…

Figure 15. Overlay of graphical elements and eye gaze data.

Experimental evaluation

We recruited five graduate students and one post-doctoral researcher to participate in a study that was conduction in May of 2016. Each participant was instructed to perform a data retrieval task for twelve different patient cases. The cases were displayed in the LEMR interface (see Section 3.6). The interface was instrumented to store element names and locations, as described previously. Figure 16 shows what the graphical element for creatinine measurements looked like. To account for some of the eye-tracking device error, each element had a 15-pixel margin.

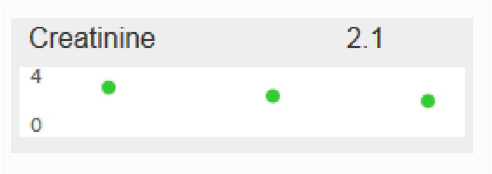


Figure 16. The display of a laboratory test. Each green point represents a creatinine test result, arranged in chronological order from left to right. Hovering over a point creates a tooltip that provides more data about that result. The value listed in the top row (2.1) corresponds to the value of the most recent test result.

During the study, the participant was asked to sit in front of a computer monitor that had the EyeX device attached. We adjusted the monitor to ensure that the participant was comfortable and the eye-tracking device had a clear view of the participant’s eyes. Once the participant was situated, they used the six-point Tobii EyeX Engine calibration program to calibrate the eye-tracking device to the computer monitor. After calibration, the participant was asked to begin performing the case tasks.

The tasks for the first four cases were to find the most recent value of specified laboratory tests. As shown in Figure 16, the most recent value is the value listed to the right of the test name. The tasks for the next four cases were to find the date of the most recent value of specified laboratory tests. To find the date, a participant must hover over the data point with the curser and read the details in the resulting pop-up tool tip. The tasks for the final four cases were to determine the trend in the values of specified laboratory tests. Figure 16 shows creatinine trending downward. The exact wording of each task is listed in Table 5. Note that there are two tasks for cases 4, 8, and 12.

Table 5. The participant tasks for each patient case in eye-tracking fidelity tests.

Case	Participant Tasks
1	Last Value of Glucose (Basic Chemistry)
2	Last Value of Platelets (CBC)
3	Last Value of WBC (CBC)
4	Last Value of Glucose (Basic Chemistry) & of pHa (Blood Gas)
5	Date of Last Glucose (Basic Chemistry)
6	Date of Last Platelets (CBC)
7	Date of Last Bands ABS (Diff)
8	Date of Last Glucose (Basic Chemistry) & Lymphs ABS (Diff)
9	Trend of Hct (CBC)
10	Trend of WBC (CBC)
11	Trend of RBC (CBC)
12	Trend of Platelets (CBC) & of Phosphate (Basic Chemistry)

We applied the DGP automatic gaze point-to-element mapping method to each patient case. The output from the method is a ranked list of the graphical elements that the participant viewed the most. The mapping method was considered accurate for a patient case if the top ranked elements were the elements that needed to be viewed to correctly answer the case task. For example, the element containing glucose levels needed to be viewed for Case 1, so the automatic mapping would be accurate only if glucose was ranked as most viewed.

4.2.2 Results

Across the twelve patient cases, the automatic gaze point-to-element mapping was 88% accurate. Table 6 shows case by case results, where there are six participants and each case requires participants to look at either one or two elements (laboratory tests). Results are summed across the six participants. *Correct elements* refers to the number of times that the top ranked element (based on the automatic eye gaze-to-graphical element mapping) was the element needed to perform case tasks. For example, Case 1 had one element was needed for the task and the top ranked element was the correct element for 3 of the 6 participants, resulting in an accuracy of 0.50.

Table 6. Performance of an automatic eye-tracking system across six participants.

Case	**Elements Needed**	**Correct Elements**	**Total Elements**	**Accuracy**
1	1	3	6	0.50
2	1	4	6	0.67
3	1	5	6	0.83
4	2	9	12	0.75
5	1	6	6	1.00
6	1	6	6	1.00
7	1	6	6	1.00
8	2	11	12	0.92
9	1	6	6	1.00
10	1	6	6	1.00
11	1	6	6	1.00
12	2	11	12	0.92
Totals		79	90	0.88

If manually performed, mapping individual gaze points to graphical elements would be a tedious process and an impractical process for large scale or real-time clinical applications. the automatic mapping method had good accuracy (88%). Accuracy seemed to improve as participants became familiar with the LEMR interface, from 69% on the first four cases to 98% on the last eight. The first four cases might also have been more difficult, because the lab value field is at the edge of the box containing it, and thus, more easily confused with elements outside of the box. This hypothesis could be evaluated in the future by repeating the experiment with a randomized ordering of the cases and case tasks. Even at current accuracy, the automatic mapping method has potential to save time and resources on eye-tracking data analysis and opens the possibility of large scale and real-time application of eye-tracking in clinical settings.

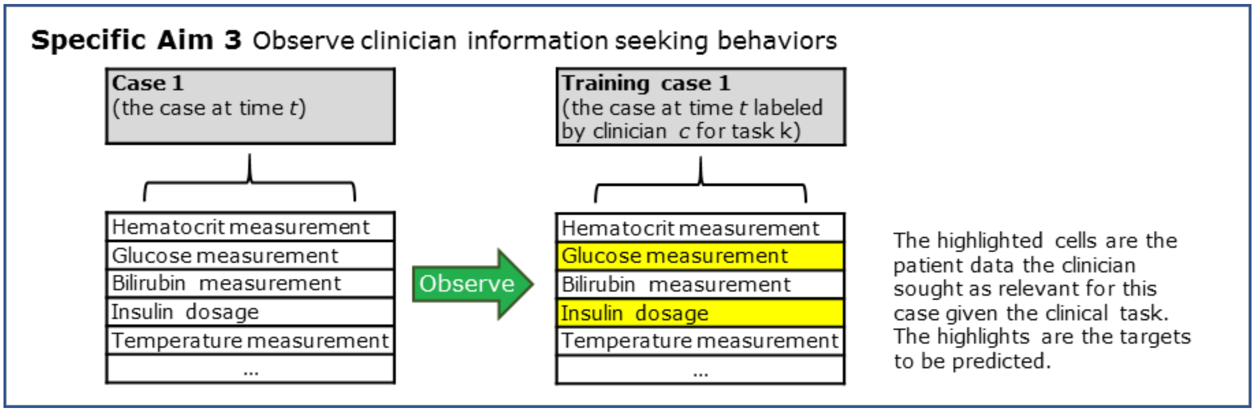
4.3 A DISSCUSSION ON EYE-TRACKING IN THE EMR

We found an inexpensive eye-tracking device to have non-inferior accuracy when compared to a more expensive device. The decreasing cost of eye-tracking devices is ushering in a new era of eye-tracking-based research and human computer interaction. As costs continue to decline, it seems likely that eye-tracking is included as a standard device in computer monitors, just like a camera is a standard device in smart phones.

Efficient processing of eye gaze data is as important as cost when considering the use of an eye-tracking device. This chapter described how a user's eye gaze can be automatically mapped to different patient data displayed in the LEMR interface. The same mapping methods would work for any instrumented interface.

Eye-tracking is described in two more sections of this dissertation. Section 5.2 evaluates using eye-tracking to observe clinician information seeking behavior when they are preparing for morning rounds (pre-rounding). Section 5.3 describes the use of eye-tracking to observe clinician information seeking behavior for a set of patient cases for training the LEMR system.

5.0 OBSERVING CLINICIAN INFORMATION SEEKING BEHAVIORS



The highlighted cells are the patient data the clinician sought as relevant for this case given the clinical task. The highlights are the targets to be predicted.

This chapter describes methods for observing clinician information seeking behavior that are needed for data-driven, context-aware LEMR systems. In the context of a LHS, the methods described in this chapter constitute the first part of a LHS loop: practice to data. In other words, this chapter presents methods of converting practice (a clinician using the EMR for a patient case) to data.

A LEMR system uses observations of clinician information seeking behavior as training data for constructing statistical models that predict clinician information seeking behavior. In a training data set, a *target variable* (or simply *target*) is any patient data item that a clinician can potentially seek as relevant for a specific task in a specific patient. Thus, any observation, measurement, action, or other information that is related to a patient, recorded in the EMR, and sought by a clinician becomes a target. A target, in a given context, takes only two values; it is assigned the value yes if it appeared in the EMR and a clinician sought it for the given task, and it is assigned the value no if it appeared in the EMR but a clinician did not seek it. It is not defined if it was not measured for the patient. For example, for a patient with diabetes mellitus, glucose **target** = *yes* denotes that the target variable **glucose target** was assigned the value *yes* because glucose was sought by a clinician. Consider a different patient who has kidney failure and glucose levels are recorded but are not sought by a clinician. Then, **glucose target** = *no* denotes that the target variable **glucose target** was assigned the value *no* because glucose was not sought by a clinician. Finally, **glucose target** = *undefined* denotes that glucose levels were not recorded for a patient and, therefore, do not appear in the EMR.

This chapter presents both manual and automatic methods of observing clinician information seeking behavior. Manual observation methods (i.e., manual selection) are used during LEMR system development and for observing any fine-tune adjustments users make to LEMR interfaces. The automatic observation method (e.g., eye-tracking) tested in this chapter is illustrative of possibilities for automatic training of LEMR systems.

5.1 PRELIMINARY MANUAL COLLECTION OF A TRAINING DATA SET

We developed and implemented a manual method for observing clinician information seeking behavior and evaluated it with a single clinician participant who used it to assign target values on a set of patient cases. These cases are used as a training data set in a preliminary modeling study (Section 6.1).

5.1.1 Methods

A clinician participant used the preliminary LEMR interface (Section 3.3 Figure 11) to view the EMR records of 59 patient cases in April 2014. The cases were selected randomly from the HIDENIC data set (described in Section 2.4.1). For each case, the participant imagined that he was the attending who was taking care of the patient. He read the clinical reports and examined the test results to determine the patient's clinical course since admission to the ICU. Then, for the last day for which patient data was displayed, he used features of the prototype to populate the HID with laboratory tests that were useful in providing evidence about (1) possible changes in the clinical condition of the patient, and (2) the emergence of a new clinical problem. Any test that he moved into the HID was considered relevant and was given a target value *yes*. Any test that the participant did not move into the HID was considered not relevant and was given the target value *no*. These labs with assigned target values are the targets in this training data set of 59 patient cases.

5.1.2 Results

Across the 59 patient cases, 36 distinct laboratory tests were identified as relevant for at least one patient case and 21 distinct tests were identified as relevant for more than one patient case. These relevant tests tended to be from basic chemistry (11), complete blood count (7), blood differential (6), and liver function (5). The minimum, median, and maximum number of tests identified as relevant for a patient case are 2, 7, and 14, respectively.

5.2 OBSERVING CLINICIAN INFORMATION SEEKING BEHAVIOR WITH EYE-TRACKING

We wanted to evaluate the extent to which eye-tracking technologies can accurately determine the patient data a clinician seeks when using the LEMR interface to complete a clinical task. To sufficiently evaluate using eye-tracking for this purpose, we chose preparing for morning rounds (pre-rounding) to be the task because it is a common activity and it requires the clinician to view a diverse set of patient data.

We investigated the extent to which the patient data a clinician views (as captured by an eye-tracking device) can function as a proxy for his or her information seeking behavior. If the patient data observed by eye-tracking sufficiently matches self-reported (manually selected) relevant information, then this automatic observation method can be applied when collecting a training data set (Section 5.3).

Our hypothesis was that when preparing to present a patient case at morning rounds, the patient data a clinician gazes longest at will be the same data the clinician manually reports as seeking in the case.

5.2.1 Methods

This section describes the study participants, experimental design, and data analysis. This study was approved by the University of Pittsburgh Institutional Review Board (ID PRO16030092).

Participants

Four critical care fellows from UPMC were recruited between July 2015 and August 2015. Each participant participated for four to five hours and their time was compensated at a rate of $100 per hour.

Experimental design

Each participant participated in one study session in which they were asked to review ten patient cases. For each case, the participant was asked to follow a two-step protocol. In the first step, the participant was presented with a patient case and asked to use the available patient data to prepare for presenting the case at morning rounds. During this step, an eye-tracking device and the automatic mapping method described in Section 4.2.1 were used to record the patient data viewed by the participant (eye-tracking data set). Once the participant decided that they were prepared to present the case, they were asked to start the second step of the protocol. In this step, the participant was asked to select the patient data they used when preparing to present the current case at morning rounds (manual selection data set). Selections were indicated using features of the LEMR interface and are considered the gold standard of clinician information seeking behavior. Eye gaze was not recorded during this step.

Data analysis

In addition to the automatic mapping methods described in Section 4.2, namely, GP and DGP, we also tested augmenting the mapping method with two different fixation algorithms: Dispersion-Threshold Identification (I-DT) and Area-of-Interest Identification (I-AOI) [100]. These algorithms, rather than considering individual gaze points, combine consecutive gaze points into fixations when they meet certain criteria. Both algorithms have a time threshold (duration) parameter. For

I-AOI, this means that a certain number of consecutive gaze points must map to the same patient data variable before that variable is considered fixated on. In addition to a duration parameter, I-DT has a distance threshold (dispersion) parameter. For this algorithm, consecutive gaze points are mapped to the same fixation when they are within a certain distance of each other. We tested these two fixation algorithms across various parameter settings for their duration (2, 3, 4, and 5 consecutive gaze points) and dispersion (20, 30, 40, 50, 60, 70, and 80 pixels) thresholds.

After the study data were collected, we compared the automatically collected eye-tracking target values against a gold standard (manually selected target values) using Area Under the Curve (AUC) of the Receiver Operator Characteristic (ROC) curves and of the Precision Recall (PR) curves. To perform the analysis, time spent viewing each graphical element in each case was used as the classification measure (i.e., the curves are produced by varying the viewing time threshold).

5.2.2 Results

We recruited four University of Pittsburgh Medical Center (UPMC) ICU fellows as study participants. All four participants wore corrective lenses. The AUC-ROC and AUC-PR results for the four participants averaged across all ten patient cases are shown in Table 7. Only the best preforming I-AOI and I-DT parameter settings are shown. The GP and DGP mapping approaches, which are based on individual gaze points rather than fixations, performed the best. With nearly identical performance, it does not appear that DGP offered any benefit over GP. The two fixation algorithms resulted in reduced performance; this result may be due to the exclusion of valid gaze points that did not meet the fixation inclusion criteria.

Table 7. Averages across all ten cases of each mapping method tested.

Algorithm	Duration (data points)	Dispersion (pixels)	Participant 1		Participant 2		Participant 3		Participant 4		Average	
			AUC		AUC		AUC		AUC		AUC	
			ROC	PR	ROC	PR	ROC	PR	ROC	PR	ROC	PR
DGP	1		0.73	0.82	0.76	0.86	0.57	0.57	0.58	0.78	0.66	0.76
GP	1		0.72	0.82	0.74	0.85	0.56	0.58	0.57	0.77	0.65	0.75
I-AOI	2		0.66	0.78	0.72	0.83	0.55	0.55	0.54	0.76	0.62	0.73
I-DT	3	80	0.50	0.64	0.52	0.67	0.55	0.54	0.55	0.74	0.53	0.65

When comparing methods of observing clinician information seeking behavior, we found that eye-tracking worked well for two participants and not well for the other two participants. This result is complicated by the fact that eye-tracking performed poorly for some participants. More analyses are needed, to know when, for which users, and for which patient data eye-tracking should be used for collecting LEMR training data.

We hoped that eye-tracking could replace manual selection as the primary method for observing clinician information seeking behavior. With the results obtained, we decided to use both manual selections and eye-tracking when collecting a LEMR training data set (Section 5.3).

5.3 PRIMARY COLLECTION OF A TRAINING DATA SET

Thus far, this chapter has presented a manual method and an eye-tracking method for collecting LEMR training data. In this section, we describe the use of these methods to collect larger sets of training data, which are used in a machine learning study (Section 6.2) and in the LEMR evaluation study (Section 7.1).

5.3.1 Methods

We describe the protocol for observing clinician information seeking behavior including the participants, LEMR interface, patient cases, experimental design, and data analysis. This study was approved by the University of Pittsburgh Institutional Review Board (ID PRO16100190). Participant sessions occurred between August 2017 and October 2017.

As presented at the beginning of this chapter, the patient data a clinician seeks as relevant when preparing a case for morning rounds are assigned the target value *yes*. Data that was recorded for a patient but not sought by a clinician are assigned the target value *no*.

Participants, LEMR interface, and patient cases

The recruited participants were ICU fellows and attending clinicians from the University of Pittsburgh in the Department of Critical Care Medicine. Each participant was compensated $100 per hour of participation. They used the LEMR interface (see Section 3.6) to complete a series of tasks for about 20 cases. The cases loaded into the interface were randomly selected from a set of ICU patient cases that (1) were admitted between June 2010 and May 2012 and (2) had a diagnosis of either acute kidney failure (AKF; ICD-9 584.9 or 584.5; 93 cases) or acute respiratory failure (ARF; ICD-9 518.81; 85 cases). Case data were extracted from a research database [106] and a clinical data warehouse [107], as described in Section 2.4.1. The cases were de-identified to create a limited data set that included dates and times related to the events.

Experimental design

Data collection occurred in a meeting room where a participant sat in front of a laptop computer that was pre-loaded with 30 patient cases. The first four cases were common across all participants (there are called *burn-in cases*) and the remaining 26 cases were different for each participant so the resulting training data set would include different patient cases. Each participant reviewed and annotated as many cases as they could during one to two sessions that lasted a total of four to six hours.

The participants reviewed and annotated the cases using the interface shown in Section 3.6, Figure 12. The case review procedure included the following tasks (see Figure 17).

Task 1. For this task a random day between day two of admission to the ICU and the day before discharge from the ICU was selected as the "past patient stay". All available EMR data up until 8:00 am on the day selected for the past patient stay was displayed to the participant. Structured data were shown in graphical time series plots and free-text notes were shown in a separate area in the interface. The participant was instructed to "use the available information to become familiar with the patient case as if they are one of your own patients." After becoming familiar with the case, the participant clicked on a button to advance to Task 2.

Task 2. An additional day (from 8:00 am on the day selected for the past patient stay to 8:00 am on the next day i.e., "current time") of the patient's EMR data was added to the display. The participant was prompted with "24-hours have passed" and directed to "use the available information to prepare to present the case during morning rounds." After preparation was complete, the participant clicked on a button to advance to Task 3.

Task 3. In the interface, each available data item (e.g., glucose levels, insulin dosage regimen) was accompanied with a check box. Clicking on the area associated with an item toggled the check box. The participant was directed to "select the information you consider pertinent when preparing to present this case at morning rounds." The participant selected relevant data items by toggling the accompanying check box to the checked state.

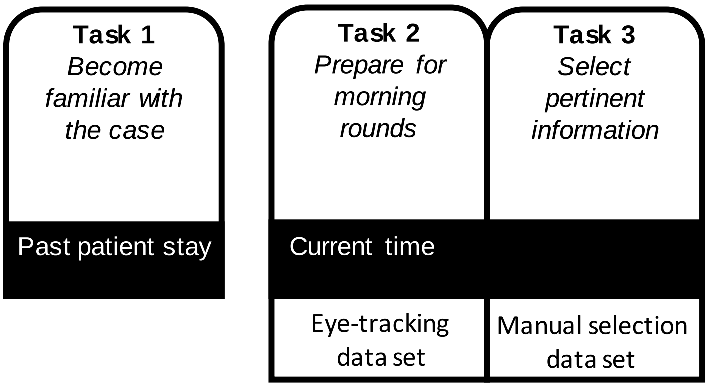

Task 1
Become familiar with the case
Past patient stay
Task 2
Prepare for morning rounds
Task 3
Select pertinent information
Current time
Eye-tracking data set
Manual selection data set

Figure 17. Case review protocol for observing clinician information seeking behavior. During this protocol, two training data sets are collected: an eye-tracking data set and a manual selection data set.

Two data sets of information seeking behavior were collected. The manual selection data set was collected during Task 3 when participants manually selected data items that were relevant to the task. The eye-tracking data set was collected during Task 2 when clinicians were preparing to present the case at morning rounds (see Section 5.2 for more details on eye-tracking). A target variable was assigned the value *yes,* if its associated data item was selected (or was gazed upon for at least 250 milliseconds), and *no,* if the associated data item was not selected (or was not gazed upon for at least 250 milliseconds).

Participant agreement

To gauge agreement among participants, we calculated an intra-class correlation coefficient (ICC) [120]. ICC ranges from 0 to 1, where values less than 0.50, between 0.50 and 0.75, and 0.75 and 0.90 are indicative of poor, moderate, and good reliability, respectively [106]. ICC was computed on the first four (burn-in) cases based on a single rater, absolute-agreement, two-way mixed-effects model (R Project; Psych package in CRAN; ICC3 method). Four cases were reviewed by all participants and, thus, were used for the ICC calculation. To increase the power of the ICC calculation, we aggregated across all data items to calculate a single ICC score. While the time to task completion for the burn-in cases is longer because users are not yet familiar with the interface, we speculated that the data items selected (target value is yes) by the participants were less affected by burn-in because relevant patient data should be the same regardless of the time it takes for a participant to use the interface.

5.3.2 Results

Participant characteristics

Table 8 summarizes the 11 critical care clinicians who participated in this study.

Table 8. Participant characteristics when collecting the primary training data sets

Gender distribution		Experience distribution		Years of experience Mean (range)	
Male	Female	Fellows	Attendings	Since medical school	In the ICU
7	4	9	2	5.3 (3.0-10.0)	1.8 (0.3-7.0)

Patient cases

A total of 178 patient cases were reviewed by the clinician participants between August and October of 2017. Of these patient cases, 52% had AKF, 48% had ARF, the average age was 60, and the median ICU day at the time of review was 7. These numbers do not include the four burn-in cases that were reviewed by all participants.

Case targets

Each case had target values assigned manually via manual selections and automatically via eye-tracking. Across the 178 patient cases, 109 different data items were manually selected, with an average of each item being selected 32 times and a maximum of 152 times. Using eye-tracking across 147 patient cases, 115 different data items were viewed, with an average of each item being viewed 35 times and a maximum of 120 times. Gaze data was not collected for the other 31 cases because the participant was not sitting within the eye-tracker's tracking range.

Participant agreement

The aggregate ICC score (agreement) of the 11 participant's manual selections on the first four burn-in cases is 0.40 (95% CI 0.36 to 0.45). This is poor agreement, suggesting that models trained on this data will start off as being very noisy.

5.4 A DISSCUSSION ON OBSERVING CLINICIAN INFORMATION SEEKING BEHAVIOR

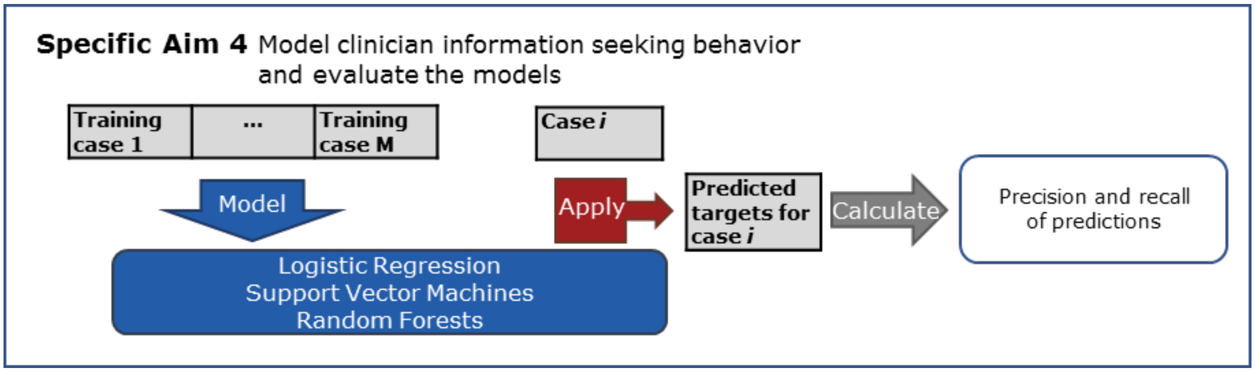
We developed and evaluated two different methods for observing clinician information seeking behavior. The manual selection method requires participants to click on the patient data they seek. This method works well enough for LEMR system development in a research setting, but it is not practical for observing clinician information usage in the real world. In real clinical settings, an implemented LEMR system would use automatic observation methods.

The automatic observation method using eye-tracking is promising for observing clinician information seeking behavior because it does not require any additional work from clinicians. Unsurprisingly, users manually selected a subset of the patient data that they viewed. In this dissertation we use manual selections as the gold standard for clinician information seeking behavior; however, clinician information viewing patterns might be more closely aligned with their information seeking behavior, than a clinician's manual judgement of their own behavior.

Tests show that clinicians have low agreement when it comes to what patient data they seek as relevant when preparing the same cases for morning rounds. The disagreement comes mainly from some clinicians being more selective and others being more liberal in the type and number of data items they choose. This suggests that the gold standard is a silver standard at best. Such disagreements will reduce the performance of models trained from it.

The manual targets collected in Section 5.3 were used to train models in the machine learning study described in Section 6.2. These trained models were applied in the LEMR evaluation study described in Section 7.1. The training data collected from eye-tracking were also applied to train models. Both sets of models (manual selection and eye-tracking) are prospectively evaluated using a gold standard collected during the evaluation study (Section 7.1).

6.0 MODELING CLINICIAN INFORMATION SEEKING BEHAVIOR


Specific Aim 4 Model clinician information seeking behavior and evaluate the models
Training case 1
...
Training case M
Model
Logistic Regression
Support Vector Machines
Random Forests
Case *i*
Apply
Predicted targets for case *i*
Calculate
Precision and recall of predictions

This chapter focuses on developing statistical models for modeling clinician information seeking behavior using training data that was collected as described in Chapter 5. Thus the data collected as described in Chapter 5 is used to generate knowledge in the form of predictive models. These models predict for a given context (e.g., user-task-case: an ICU fellow-preparing to present morning rounds-for a patient with acute kidney failure) what patient data clinicians will seek as relevant. Three different machine learning algorithms are used to train the models. In the context of a LHS, the models described in this chapter constitute the second part of the LHS loop: data to knowledge. In other words, this chapter presents methods of converting data (collected from clinicians using the EMR) to knowledge (models of clinician information seeking behavior).

Chapter 5 described methods for observing clinician information seeking behavior and the application of those methods to assign target values to patient cases. Those target variables were combined with predictor variables constructed from the same patient cases. A predictor variable denotes any patient data item and includes observations, measurements, actions, or other information that are recorded in the EMR. Examples of predictor variables in an ICU EMR include demographics, diagnosis, vital sign measurements, ventilator settings, intake and output, laboratory test results and medication administrations. A *predictor value* is the value that a predictor variable takes in a patient. Consider a patient with diabetes mellitus in whom glucose levels are recorded in the EMR. Then **diagnosis** = *diabetes mellitus* denotes that the predictor variable **diagnosis** has the value *diabetes mellitus* and **glucose** = *85, 100, 90, 105 mg/dL* denotes that the predictor variable **glucose** consists of a series of *glucose levels* over a period of time.

We use the term *variables* to denote raw patient data items that are recorded in the EMR (e.g., glucose levels) and the term *features* to denote functions of those variables (e.g., most recent measurement of the glucose levels). Predictor variables include simple atemporal variables (e.g., diagnosis), as well as more complex variables that represent multivariate time series data (e.g., glucose). We construct features from predictor variables as described below:

- For each atemporal variable such as diagnosis and demographics, we generate a single feature that is assigned a single value for a patient for the duration of ICU stay (e.g., gender = male).
- For each medication variable, we generate several features to summarize the time series of administered doses. For example, for an insulin dosing regimen we generate 4 features that include 1) an indicator of whether the patient is currently prescribed insulin, 2) the time since its first administration to the current time, 3) the time since its most recent administration to the current time, and 4) its dose at the most recent administration.

- For each laboratory test result and vital sign, we generate an extensive set of features. For example, for glucose that consists of a time series of glucose levels, we generate 36 features that include the first glucose level during the ICU stay, the most recent level, the highest and lowest levels until current time, the difference between the most recent two levels, and so on.
- Current clinician user was represented by a set of 11 binary features; one for each clinician who provided training data. For a patient case, a current user feature was assigned the value 1 if the corresponding clinician reviewed and annotated that case; otherwise the feature was assigned the value 0.

Target variables are treated like atemporal predictor variables when they are translated into features. Thus, the glucose target (i.e., whether glucose levels are sought in a given context) is translated into a single target feature (which we simply call target) in contrast to the glucose predictor which is expanded into a set of glucose features.

A patient instance (or simply *instance* or *sample*) is a vector of (predictor) feature values and corresponding target values that are derived from data from a subinterval of a patient's ICU stay that is defined from the point of admission to the ICU to the current day and time. The vector of feature values summarizes the clinical evolution of the patient's condition from the time of admission to the ICU to the current day. A data set (e.g., a training data set) is a collection of patient instances.

To train a predictive model for glucose target, for example, we train on all feature values and corresponding glucose target values (*yes* or *no*) of a data set of instances to predict if the glucose level is sought after. By changing the target, a predictive model is trained for each laboratory test, medication, ventilator setting, and vital sign. In this data representation, the temporal aspects of the predictor variables are implicitly summarized in the vector of feature values and such a representation enables standard machine learning methods to be applied.

Section 6.1 and Section 6.2 describe a preliminary machine learning study to test these methods and a larger machine learning study that trained the models applied in the LEMR evaluation study (Section 7.1), respectively.

6.1 PRELIMINARY MODELING OF A SMALL MANUALLY COLLECTED TRAINING DATA SET

To test the feasibility and accuracy of using supervised machine learning to model clinician information seeking behavior using the LEMR interface, we used the data collected as described in Section 5.1 to train and evaluate penalized logistic regression models.

6.1.1 Methods

The data consisted of 59 patient cases and 21 target variables. All target variables were laboratory tests that were selected as relevant for at least three of the cases. This was a preliminary study that occurred before the observation study described in Section 5.3. With a limited sample size (59 cases), we constructed a smaller set of features than what was described at the beginning of this chapter. The smaller set of features included, as follows, five demographic features (age, sex, weight, height, and body mass index), two features for each of the 190 distinct laboratory tests (the most recent value and a Boolean value for whether that test result had appeared within the last 24 hours of available patient data), and one feature that stored the number of days since the patient's admission to the ICU. In total there were 386 predictive features for every patient case.

A LEMR system uses a model of information seeking behavior to direct the future display of patient data. To test the feasibility and accuracy of such models, we trained 21 penalized logistic regression models that each predict if a specific laboratory test was sought as relevant by the reviewing clinician for a patient case. The models used were implemented in the Scikit-learn Python package [114] and each model is evaluated individually using leave-one-out cross-fold validation. For example, to train a predictive model for the glucose target, we trained on all feature values and corresponding glucose target values for 58 of the instances and use the predictive features from the 59th instance to predict if the glucose level is sought after for that instance. We compare the model's prediction to the glucose target value for that instance to determine if the prediction was correct. This is repeated leaving out a different patient instance until all training samples have been left out once. The results are then averaged to determine model performance for glucose target. By changing the target, a predictive model is trained for each laboratory test. Performance was measured using the Area Under the Receiver Operating Characteristic curve (AUROC).

6.1.2 Results

We trained penalized logistic regression models to decide when specific laboratory tests would be sought by clinicians for each of the 59 patient cases. There were 21 models in all, one for each of the tests that were give a target value of *yes* for at least three patient cases in the training data set. The AUROC values for those models and the number of positive training samples in their data sets are shown in Table 9. The average AUROC is 0.73. The top seven tests shown in the table have an average AUROC of greater than or equal to 0.80.

Table 9. Performance of logistic regression models when predicting clinician information seeking behavior.

		95% CI		
Target variable	**AUROC**	**Lower**	**Upper**	**Number positive**
Bilirubin Total	0.92	0.83	0.97	5
Liver Alanine Aminotransferase	0.91	0.72	0.98	4
Liver Aspartate Transaminase	0.91	0.72	0.99	4
PTT Coagulation	0.84	0.71	0.92	9
Lactate	0.83	0.58	1.00	2
Phosphorus	0.82	0.62	0.94	11
White Blood Cell	0.80	0.67	0.91	8
INR Coagulation	0.79	0.63	0.89	11
Hematocrit	0.77	0.59	0.89	37
Sodium	0.75	0.61	0.86	18
Glucose	0.73	0.55	0.87	12
Chloride	0.73	0.59	0.82	2
Blood Urea Nitrogen	0.73	0.56	0.85	22
Hemoglobin	0.71	0.54	0.83	33
Platelets	0.70	0.53	0.82	28
Lymphocytes Absolute	0.64	0.26	0.95	2
Neutrophils Absolute	0.64	0.27	0.95	2
Red Blood Cell	0.57	0.25	0.97	3
Magnesium	0.56	0.27	0.89	5
Potassium	0.52	0.37	0.68	11
Calcium	0.47	0.28	0.83	5
Average	*0.73*			

The results from this study suggest that it is possible to predict clinician information seeking behavior for a patient case (within a given clinical context). Result generalizability is limited by sample size and by having a single clinician provide all the training data. These limitations are addressed in the next section.

6.2 PRIMARY MACHINE LEARNING EVALUATION ON A TRAINING DATA SET

Following the encouraging results obtained from using data that was annotated by a single clinician, we performed a similar study with data derived from a group of clinicians to evaluate the feasibility and accuracy of using supervised machine learning to model clinician information seeking behavior. The models are trained and evaluated on the training data collected in Section 5.3 and are applied in the LEMR evaluation study, Section 7.1.

This section describes the evaluation of the performance of different models when used to predict clinician information seeking behavior when preparing for morning rounds. This study determines which types of machine learning algorithms and predictor variables work best in support of the LEMR system task.

We hypothesized that machine learning models of clinician information seeking behavior that are trained and cross-validated on the manual selection training data set collected in Section 5.3 will have a precision of at least 0.67 when recall is 0.8. We chose these thresholds, because we wanted to apply well performing models in the LEMR evaluation study. If poor performing models were applied in the evaluation study, then the utility of having highlights would be negatively affected. At this chosen level of performance, a LEMR system would highlight four out of every five data items a clinician seeks and two out of three highlighted data items would be sought.

6.2.1 Methods

We describe the training of models of clinician information seeking behavior, preprocessing of data and machine learning algorithms, and learning rate calculations. This study was approved by the University of Pittsburgh Institutional Review Board (ID PRO17030147) and occurred between November 2017 and February 2018.

Training models of clinician information seeking behavior

Before training models of clinician information seeking behavior, we preprocessed the patient cases and target values from Section 5.3 into a representation suitable for machine learning. As presented at the beginning of this chapter, each training data sample consists of a patient case that is comprised of a vector of *values* for *predictor variables* and is augmented with *values* for *target variables*. A separate model is trained for each target variable (in this case, any laboratory test, medication order, ventilator setting, or vial sign). We applied and evaluated three different machine learning algorithms, namely, lasso logistic regression, support vector classifier, and random forest classifier. Additionally, we calculated the learning rates for the best performing models and used them to estimate the sample sizes needed to train them.

Preprocessing of data and machine learning algorithms

A patient case described in terms of patient data consists of complex multivariate time series data that include laboratory test results, medication administrations, vital sign measurements, and simpler variables such as demographics and co-morbidities. From a patient case, we construct a patient instance that consists of a feature vector and a corresponding target. In a patient instance, the feature vector summarizes the clinical evolution of the patient's condition from the time of admission to the ICU to the current day. Table 10 shows the full list of features that are constructed from each predictor variable type, where an X in the table signifies a feature was included in the vector of values for a predictor variable type. For laboratory tests and ventilator settings, the features constructed will vary depended on whether the test or setting was ordinal (e.g., peripheral blood smear), nominal (e.g., urine color), or interval (e.g., temperature). Using this data representation, the temporal aspects of time series data are implicitly summarized in the vector of predictor values; this has the advantage that standard machine learning methods can be applied. Open source preprocessing software is available online at (https://github.com/ajk77/PatientPy) or see Appendix B.

Table 10. Variable expansion for machine learning. An X signifies a feature was included in the vector of values for a predictor variable type.

Feature name	**RT.**	**Laboratory tests**			**Vital signs**	**Ventilator settings**		**Med.**	**Proc.**	**Mic.**
		Or.	**No.**	**In.**	**In.**	**No.**	**In.**			
Event ever occurred	B	X	X	X	X	X	X	X	X	X
[First, last, second to last] value	F	X		X	X		X			
[First, last, second to last] value is category [1, 2, 3, 4]	B		X			X				
[Days, inverse of days] since last value	F	X	X	X	X	X	X	X	X	X
New event ordered	B	X	X	X	X	X	X			
Days since last change in value	F	X	X	X	X	X	X	X		
Number of different values	I	X	X	X	X	X	X			
Variation in event frequency	F	X	X	X	X	X	X	X		
[Apex, nadir, baseline] value	F			X	X		X			
Difference between last value and [first, second to last, apex, nadir, baseline]	F			X	X		X			
Percentage change between last value and [first, second to last, apex, nadir, baseline]	F			X	X		X			
Slope between last value and [first, second to last, apex, nadir]	F			X	X		X			
Flag is [High, Low, Abnormal, null]	B	X	X	X	X					
Absolute value of the slope between last value and second to last value	F	X	X	X	X	X	X			
[Mean, max, min] in last 30 hours	F	X	X	X	X	X	X			

Note. An X signifies a feature was included in the vector of values for a predictor variable type. RT. = result type. B = Boolean. F = float. I = integer. Or. = ordinal. No. = nominal. In. = interval. Med. = medication orders. Proc. = procedures. Mic. = microbiology. Brackets indicate sets of variables.

Table 10 (continued).

Feature name	RT.	Med.	Proc.	Mic.	IO	Dem.
Days since first value	F	X	X	X		
Ongoing event	B	X				
Number of sequential days of event	I	X				
Recency of sequential days	F	X				
Daily [urine, oral, intravenous, blood products, everything else, other/unknown, net]	F				X	
Length of stay [urine, oral, intravenous, blood products, everything else, other/unknown, net]	F				X	
Age	F					X
Height	F					X
Weight	F					X
Body mass index	F					X
Is female	B					X
Is Caucasian	B					X
Length of stay (days)	F					X

Note. An X signifies a feature was included in the vector of values for a predictor variable type. RT. = result type. B = Boolean. F = float. I = integer. Med. = medication orders. Proc. = procedures. Mic. = microbiology. IO = intake and output. Dem. = demographics. Brackets indicate sets of variables.

Target variables. A target variable, as described at the beginning of Chapter 5, is any patient data that a clinician can potentially indicate as being relevant and includes diagnosis, demographics, laboratory test results, medication administrations, and vital sign measurements. Every target variable is binary and for a patient is assigned the value *yes* (if it was measured and the participant selected it) or *no* (if it was measured but the participant did not select it). A variable is not defined if it was not measured. To train a predictive model for **glucose target**, for example, we train on the vector of predictor values and corresponding **glucose target** values of a set of patient cases to predict if the glucose level is sought after. By changing the target variable, a predictive model is trained for each laboratory test, medication, ventilator setting, and vital sign.

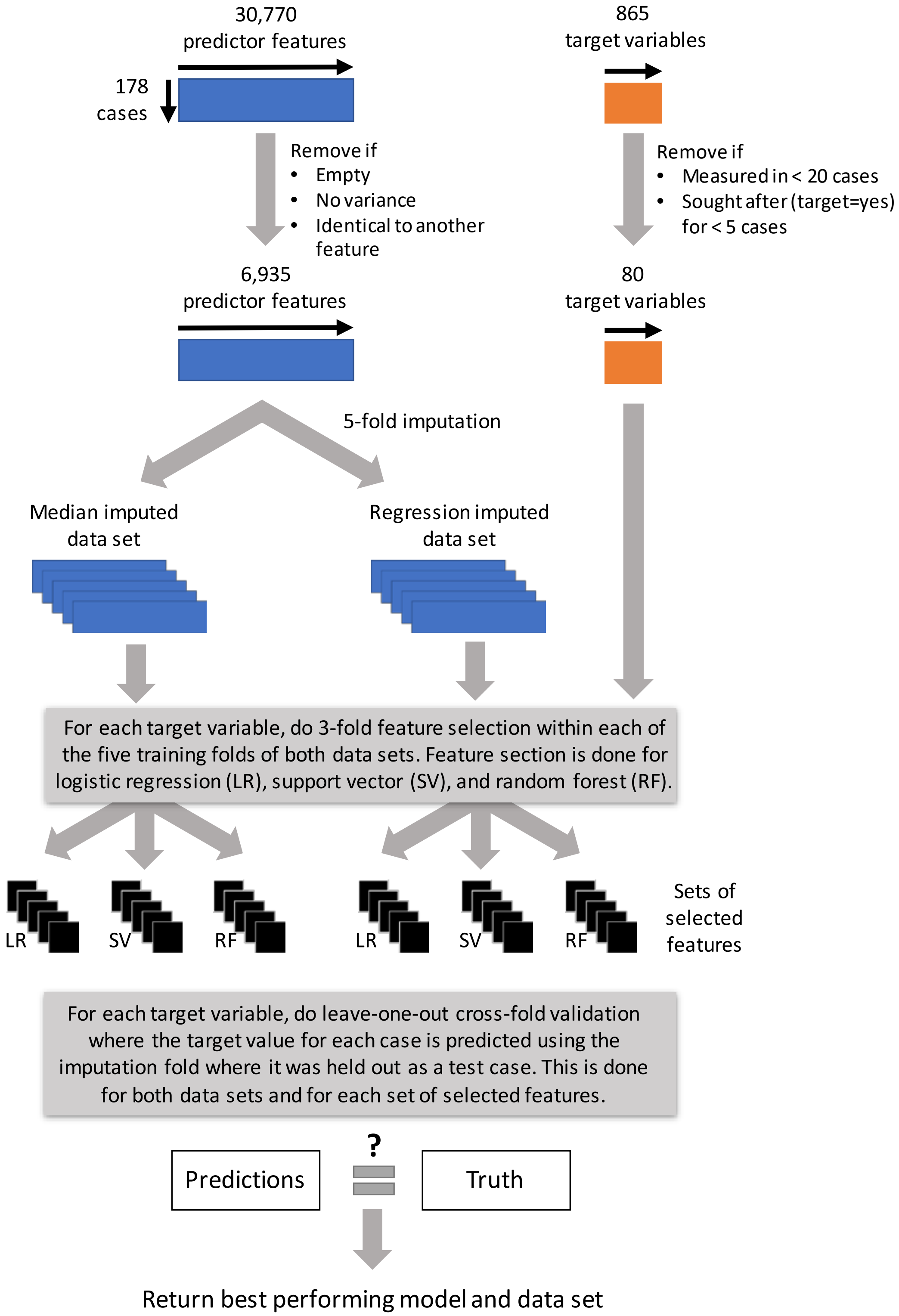
In the manual selection data set collected in Section 5.3, 80 EMR variables were measured in at least 20 patient cases and were sought (i.e., target=*yes*) in at least 5 of those cases. These are the target variables for which a model was trained in this study.

Missing values. Missing values were imputed using two different methods. In the first method, they were imputed with the median. In the second method, continuous predictor variables were imputed via linear regression and discrete predictor variables were imputed via logistic regression. To impute a feature value using regression, all cases not missing that value are used as training data. In the training data, the target feature is the feature that needs to be imputed, and all other features are used as predictor features. If a predictor feature contains missing values, those values are temporarily imputed with the median when training and applying a regressive imputer.

Both imputation methods were applied, creating two distinct data sets (a median imputed data set and a regression imputed data set). We train models on each data set separately and compare performance. Open source imputation software is available online at (https://github.com/ajk77/RegressiveImputer) or see Appendix B.

Feature selection. Feature selection was performed in two steps. First, for each set of features constructed from a single data item variable (e.g., blood glucose levels being expended into a set of features that include the most recent measurement, the slope between the two most recent measurements, etc.), we test to see if the set is predictive of the target by itself. We test this by cross-validating models. Any set of features with an area under the Receiver Operator Characteristic (AUROC) curve of less than 0.6 is removed. The features that remain after the first step are reduced further using recursive feature elimination and cross-validation (RFECV in the Python Package scikit-learn). The final set of features is used for model construction. Feature selection is target specific, so it was done separately for each of the target variable. Open source feature selection software is available online at (https://github.com/ajk77/PatientPyFeatureSelection) or see Appendix B.

Machine learning algorithms. Three different machine learning algorithms were applied: lasso logistic regression [108], support vector classifier [121], and random forest classifier [122]. Models were constructed by applying these algorithms using leave-one-out cross-fold validation. The imputation and feature selection steps were performed within the cross folds, as show in Figure 18. Results are reported as AUROC with 95% confidence intervals estimated by bootstrapping. The scikit-learn [114] implementation of each algorithm was used.



For each target variable, do 3-fold feature selection within each of the five training folds of both data sets. Feature section is done for logistic regression (LR), support vector (SV), and random forest (RF).

For each target variable, do leave-one-out cross-fold validation where the target value for each case is predicted using the imputation fold where it was held out as a test case. This is done for both data sets and for each set of selected features.

Return best performing model and data set

Figure 18. A flowchart that shows the training and evaluation of models of clinician information seeking behavior.

Learning rate calculation

Since obtaining target labels is expensive, we wanted to measure the learning rate of models to estimate the number of training cases that would be needed to reach optimal model performance. To calculate the learning rate of models, we trained each model with varying numbers of training cases. In particular, each model was trained with 25, 50, 75, and 100 percent of its respective training set. Resulting AUROCs are reported using box and whisker plots.

6.2.2 Results

Description of data set

The data set was assembled from 178 patient cases and 1,875 data items from 9 domains (Table 11). The total number of features in the final data set was 6,935. The data set consisted of 178 rows (one row for each patient case), 6,935 feature columns, and 80 target columns. Forty-one percent of data values were missing. Feature selection resulted in reducing the 6,935 features to an average of 88 features per target variable.

Performance of models

As mentioned in Section 6.2.1, models were trained to predict 80 distinct targets. These targets were chosen because they were measured for at least 20 patient cases in the training data set. Table 12 shows model performance for each target variable, including, a count of how many cases the target was selected as relevant (target = *yes*), the number of cases the target was measured, precision, recall, AUROC, a 95% confidence interval for the AUROC, and which model (lasso logistic regression or random forest classifier) and which imputation data set (median or regression) led to the highest AUROC score. Logistic regression and random forest models dominated support vector classifier for all models with AUROC performance of greater than 0.7, so we removed support vector classifier from consideration. Nineteen of the 80 models had performance meeting the criteria specified in the hypothesis (precision of at least 0.67 when recall is 0.80).

Learning rates

Learning rate calculations were performed by training all models in Table 12 at four training set sizes: 25, 50, 75, and 100 percent of each model's respective data set. The median AUROCs for the varying training set sizes are shown in Figure 19. Overall, the median AUROC increases as the number of training cases increases, but only slightly for the largest and next largest training set sizes.

Table 11. Data items and the number of features constructed from them.

Domain	Data item type	Number of data items of this type	Number of features per data item of this type	Number of features from data items of this type, before feature selection
Laboratory test results	Ordinal	94	19	1786
	Nominal	26	28	728
	Interval	519	36	18684
Vital sign measurements	Interval	14	36	504
Ventilator settings	Nominal	4	24	96
	Interval	5	32	160
Medication	Nominal	796	9	7164
Procedures	Nominal	394	4	1576
Microbiology	Nominal	10	4	40
Input and output	Interval	1	14	14
Demographics	Mixed	1	7	7
Participant	Nominal	11	1	11

Table 12. Performance of models of clinician information seeking behavior. Rows are ordered by best AUROC performance.

Target variable	Count of target=*yes*	Number of cases	Precision	Recall	AUROC	AUROC 95% CI	Model & data set
red blood cells	*18*	*165*	*0.92*	*0.61*	*0.94*	*0.86-0.99*	*Rf-r*
magnesium sulfate	9	99	0.29	0.44	0.83	0.71-0.93	LR-r
ventilator status	15	131	0.44	0.27	0.83	0.74-0.92	Rf-r
PEEP	**9**	**24**	**0.73**	**0.89**	**0.83**	**0.64-1.00**	**Rf-r**
pH	46	137	0.63	0.48	0.77	0.70-0.84	Rf-m
bicarbonate (blood gases)	11	108	0.50	0.09	0.75	0.62-0.86	Rf-m
vancomycin	37	80	0.62	0.81	0.74	0.64-0.82	Rf-m
anion gap	19	118	0.42	0.26	0.74	0.63-0.83	Rf-r
oxygen saturation	**103**	**177**	**0.70**	**0.76**	**0.74**	**0.68-0.80**	**Rf-m**
bilirubin total	36	110	0.52	0.61	0.73	0.64-0.80	Rf-m
lactate	*50*	*117*	*0.57*	*0.74*	*0.73*	*0.65-0.81*	*LR-r*
piperacillin-tazobactam	*24*	*50*	*0.64*	*0.58*	*0.73*	*0.61-0.84*	*Rf-m*
norepinephrine	17	39	0.58	0.65	0.72	0.58-0.85	Rf-r
chloride	**106**	**178**	**0.74**	**0.79**	**0.71**	**0.65-0.78**	**Rf-m**
alkaline phosphatase	14	109	0.20	0.07	0.71	0.62-0.80	Rf-m
potassium chloride	28	136	0.31	0.18	0.71	0.62-0.79	Rf-m
heparin	38	102	0.58	0.58	0.71	0.62-0.79	LR-m
glucose	**114**	**175**	**0.77**	**0.77**	**0.71**	**0.64-0.78**	**LR-m**
aspirin	15	47	0.41	0.47	0.71	0.56-0.84	LR-r
fentanyl	18	89	0.50	0.28	0.70	0.58-0.80	Rf-r
fraction of inspired O_2	**95**	**151**	**0.74**	**0.88**	**0.69**	**0.61-0.77**	**Rf-m**
central venous pressure	31	111	0.46	0.42	0.69	0.60-0.78	Rf-r
calcium	41	163	0.45	0.32	0.68	0.59-0.76	Rf-m
magnesium	74	173	0.56	0.55	0.68	0.62-0.75	Rf-m
respiratory rate	**121**	**178**	**0.73**	**0.84**	**0.68**	**0.61-0.75**	**Rf-r**
famotidine	26	84	0.43	0.35	0.68	0.58-0.78	Rf-r
blood urea nitrogen	**114**	**177**	**0.72**	**0.88**	**0.68**	**0.60-0.76**	**Rf-m**

Note. LR = lasso logistic regression. RF = random forest classifier. m = median imputed data set. r = regression imputed data set. Model and data set were selected on the basis of AUROC. Precision and recall are reported using a classification probability threshold of 0.5. Models meeting the criteria in the hypothesis, at any threshold, are bolded. Models meeting a relaxed criteria of precision >= 0.67 and recall >= 0.5, at any threshold, are italicized.

Table 12 (continued).

Target variable	Count of target=*yes*	Number of cases	Precision	Recall	AUROC	AUROC 95% CI	Model & data set
partial thromboplastin time	15	108	0.22	0.27	0.68	0.55-0.79	LR-m
ventilator mode	*71*	*148*	*0.60*	*0.70*	*0.67*	*0.59-0.73*	*Rf-m*
partial pressure of CO_2	31	138	0.42	0.26	0.67	0.58-0.75	Rf-m
neutrophils	24	156	0.35	0.25	0.67	0.56-0.78	Rf-m
temperature	**144**	**178**	**0.83**	**0.97**	**0.67**	**0.57-0.75**	**Rf-r**
intake and output	81	178	0.62	0.67	0.66	0.59-0.73	Rf-m
glomerular filtration rate	19	166	0.30	0.32	0.66	0.54-0.77	LR-r
phosphate	69	170	0.51	0.62	0.65	0.57-0.71	Rf-m
aspartate aminotransferase	25	113	0.39	0.44	0.65	0.55-0.76	LR-m
alanine aminotransferase	23	111	0.42	0.22	0.65	0.55-0.74	Rf-m
INR	62	125	0.62	0.61	0.65	0.56-0.73	LR-m
platelets	**116**	**166**	**0.78**	**0.72**	**0.65**	**0.56-0.72**	**LR-m**
creatinine	**132**	**177**	**0.77**	**0.89**	**0.65**	**0.58-0.72**	**Rf-r**
blood pressure	**151**	**178**	**0.86**	**0.97**	**0.65**	**0.56-0.75**	**Rf-m**
dextrose 5% in water	17	50	0.33	0.29	0.65	0.52-0.77	Rf-r
ampicillin-sulbactam	*9*	*22*	*0.40*	*0.44*	*0.65*	*0.46-0.85*	*Rf-m*
potassium	**121**	**178**	**0.76**	**0.78**	**0.64**	**0.57-0.72**	**LR-m**
albumin	19	114	0.33	0.26	0.64	0.53-0.76	Rf-r
venous saturation of oxygen	9	41	0.50	0.22	0.64	0.43-0.83	Rf-m
bicarbonate (chemistry)	**104**	**178**	**0.64**	**0.74**	**0.64**	**0.57-0.71**	**Rf-m**
white blood cells	**132**	**166**	**0.81**	**0.91**	**0.64**	**0.56-0.72**	**Rf-r**
sodium	**128**	**178**	**0.76**	**0.94**	**0.64**	**0.56-0.72**	**Rf-m**
venous pH	5	43	0.29	0.40	0.64	0.33-0.90	LR-m
partial pressure of O_2	30	137	0.39	0.30	0.64	0.54-0.75	Rf-r
Senna	10	46	0.29	0.40	0.64	0.50-0.77	LR-m
prothrombin time	12	125	0.13	0.17	0.64	0.52-0.76	LR-m
hemoglobin	**123**	**166**	**0.79**	**0.76**	**0.63**	**0.55-0.71**	**LR-m**

Note. LR = lasso logistic regression. RF = random forest classifier. m = median imputed data set. r = regression imputed data set. Model and data set were selected on the basis of AUROC. Precision and recall are reported using a classification probability threshold of 0.5. Models meeting the criteria in the hypothesis, at any threshold, are bolded. Models meeting a relaxed criteria of precision >= 0.67 and recall >= 0.5, at any threshold, are italicized.

Table 12 (continued).

Target variable	**Count of target=*yes***	**Number of cases**	**Precision**	**Recall**	**AUROC**	**AUROC 95% CI**	**Model & data set**
bilirubin direct	16	88	0.67	0.12	0.63	0.50-0.76	Rf-m
albuterol-ipratropium	15	61	0.33	0.4	0.63	0.49-0.77	LR-r
heart rate	**152**	**178**	**0.87**	**0.97**	**0.62**	**0.52-0.72**	**Rf-m**
ionized calcium	30	132	0.37	0.33	0.62	0.52-0.71	LR-m
midazolam	9	54	0.29	0.44	0.62	0.42-0.82	LR-m
Propofol	17	46	0.47	0.53	0.61	0.47-0.75	Rf-r
base solution	**50**	**87**	**0.61**	**0.68**	**0.61**	**0.52-0.70**	**Rf-m**
pantoprazole	16	45	0.5	0.44	0.61	0.46-0.76	Rf-r
insulin (Humulin & Novolin)	36	81	0.60	0.50	0.61	0.50-0.72	LR-r
insulin aspart (Novolog)	11	29	0.50	0.55	0.6	0.41-0.78	Rf-r
sodium chloride 0.9%	65	154	0.53	0.52	0.59	0.52-0.66	LR-r
ventilator tube status	38	130	0.39	0.32	0.59	0.51-0.68	Rf-m
metoprolol	19	62	0.31	0.21	0.58	0.46-0.70	Rf-r
vancomycin, trough	13	43	0.30	0.23	0.57	0.40-0.74	Rf-m
ammonia	12	42	0.33	0.25	0.57	0.39-0.74	Rf-m
hematocrit	7	166	0.07	0.14	0.56	0.35-0.74	LR-r
chlorhexidine topical	20	92	0.17	0.15	0.56	0.45-0.66	LR-m
metronidazole	*16*	*33*	*0.56*	*0.56*	*0.55*	*0.38-0.73*	*Rf-m*
furosemide	28	76	0.44	0.39	0.54	0.42-0.66	Rf-r
troponin	10	62	0.25	0.1	0.52	0.34-0.70	Rf-m
band neutrophils	13	85	0.12	0.23	0.52	0.38-0.66	LR-r
insulin glargine (Lantus)	**13**	**22**	**0.58**	**0.54**	**0.50**	**0.25-0.74**	**LR-r**
acetaminophen	12	72	0.25	0.08	0.47	0.34-0.61	Rf-r
Lorazepam	9	40	0.19	0.33	0.45	0.28-0.62	LR-m
fibrinogen	6	23	0.33	0.17	0.41	0.19-0.65	Rf-m
hydrocortisone	10	20	0.44	0.40	0.40	0.18-0.65	LR-m

Note. LR = lasso logistic regression. RF = random forest classifier. m = median imputed data set. r = regression imputed data set. Model and data set were selected on the basis of AUROC. Precision and recall are reported using a classification probability threshold of 0.5. Models meeting the criteria in the hypothesis, at any threshold, are bolded. Models meeting a relaxed criteria of precision >= 0.67 and recall >= 0.5, at any threshold, are italicized.

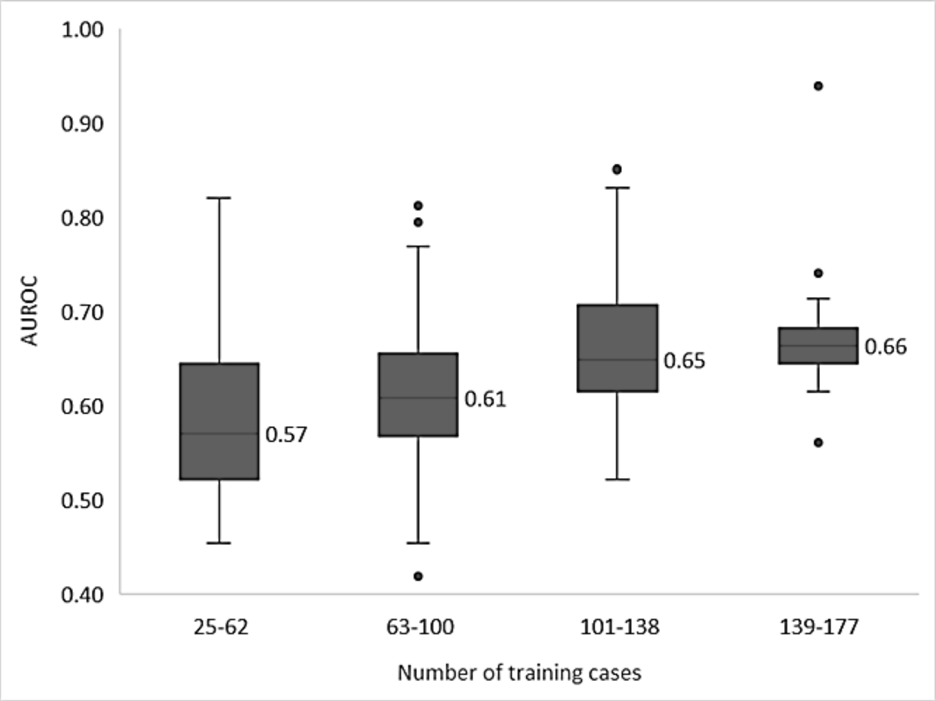


Figure 19. Learning rates of models of clinician information seeking behavior.

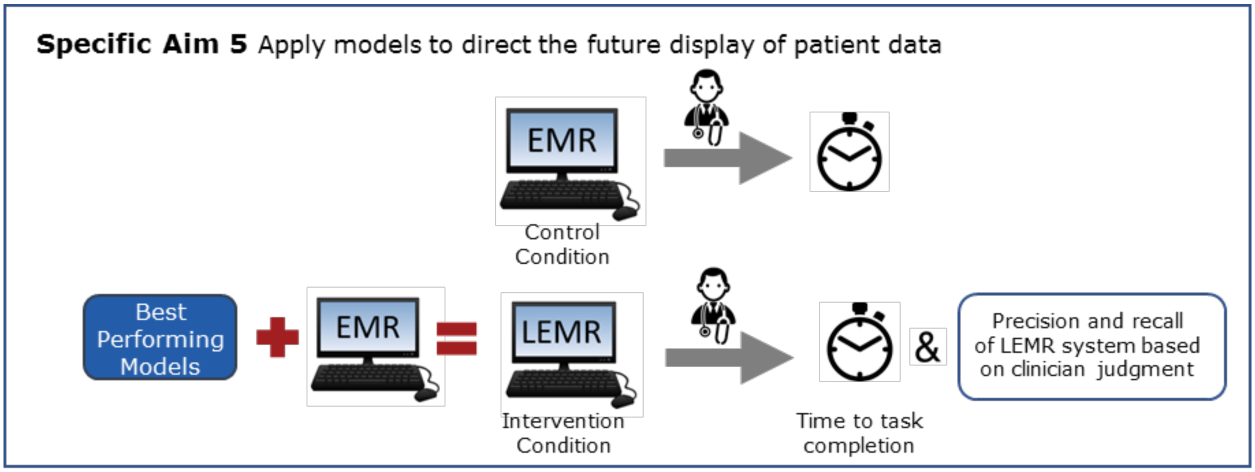
6.3 A DISSCUSSION ON MODELING RELEVANT INFORMATION

This chapter presented the training of models of clinician information seeking behavior. The target variables (e.g., blood glucose levels, insulin dosing regimen) a clinician seeks varies by context. As described in Chapter 1, context includes EMR user type, clinical task, and patient case. The models trained in Section 6.2 focus on the context of intensivists (user), preparing for morning rounds (task), for patients with AKF or ARF (cases).

Sufficient sample sizes were available for building models to predict 80 different target variables and, despite relatively small training sets, AUROC performance was at least 0.70 for a quarter of the models. These encouraging results are bolstered by the learning rate results. All but one model with at least 120 training samples had an AUROC greater than 0.60, and most models showed an upward trend in AUROC values as the number of training samples increased.

The hypothesis in Sections 6.2 was that models would have a precision of at least 0.67 when recall was 0.80. We planned on applying any model that reached this level of performance in the evaluation study (Section 7.1). When considering any classification probability threshold, 19 models did reach this level of performance. To increase the number of models applied during the evaluation study, we relaxed the performance requirements to include any model with a precision of at least 0.67 and a recall of at least 0.50. Twenty-five models met these requirements and were prospectively applied in the evaluation study, where we evaluate the impact the LEMR system has on clinicians while they prepare for morning rounds.

7.0 APPLYING MODELS TO DIRECT THE FUTURE DISPLAY OF PATIENT DATA


Specific Aim 5 Apply models to direct the future display of patient data
EMR
Control Condition
Best Performing Models
EMR
LEMR
Intervention Condition
&
Time to task completion
Precision and recall of LEMR system based on clinician judgment

This chapter focuses on the evaluation of the LEMR system. The predictive models that were developed in Chapter 6 are prospectively applied to a new set of patient cases to predict which data will be sought as relevant. These predicted data are highlighted for some LEMR users and not for others, as twelve clinician participants prepare to present eighteen patient cases for morning rounds. We measure time to task competition and compute precision and recall of the highlighted data. In the context of a LHS, the application of models described in this chapter constitute the third part of the LHS loop: knowledge to performance. In other words, this chapter presents methods of converting knowledge (models of clinician information seeking behavior) to performance (applying a model to direct the future display of patient data).

7.1 PRIMARY LEMR SYSTEM EVALUATION

Using the predictive models developed in Section 6.2, we prospectively evaluate the LEMR system. This evaluation (1) tests if the LEMR system highlights reduce the time it takes for a clinician to prepare for morning rounds and (2) assesses the adequacy of the information highlighted (precision and recall). Additional results are reported, including (3) an evaluation of each model that was applied, (4) a comparison between models trained on a data set in which the targets were manually determined and models trained on a data set in which the targets were determined using eye-tracking, (5) the clinical impact of concealing patient data that are not predicted by models, and (6) an assessment of acceptance and use of the LEMR system by the participants.

We hypothesized, that LEMR system will yield the following results on a set of test cases: (1) on average clinicians will use less time in preparing to present a patient case at morning rounds, and (2) clinicians will judge that the system highlights all of the data that they would seek in each case for the specified task.

7.1.1 Methods

This section describes the participants, patient cases, study design, the LEMR interface, study tasks, models of clinician information seeking behavior applied to determine highlighting, data collection, and the data analysis design. This study was approved by the University of Pittsburgh Institutional Review Board (ID PRO17050016). Participant sessions occurred between February 2018 and May 2018.

Participants

A total of 12 critical care specialists (intensivists) were recruited from the University of Pittsburgh in the Department of Critical Care Medicine. Participant characteristics are summarized in Table 13.

Table 13. Participant characteristics for the primary LEMR evaluation.

Gender distribution		Experience distribution		Years of experience Mean (range)	
Male	Female	Fellows	Attendings	Since medical school	In the ICU
9	3	12	0	5.4 (3.0-11.0)	1.6 (0.6-4.0)

Patient cases

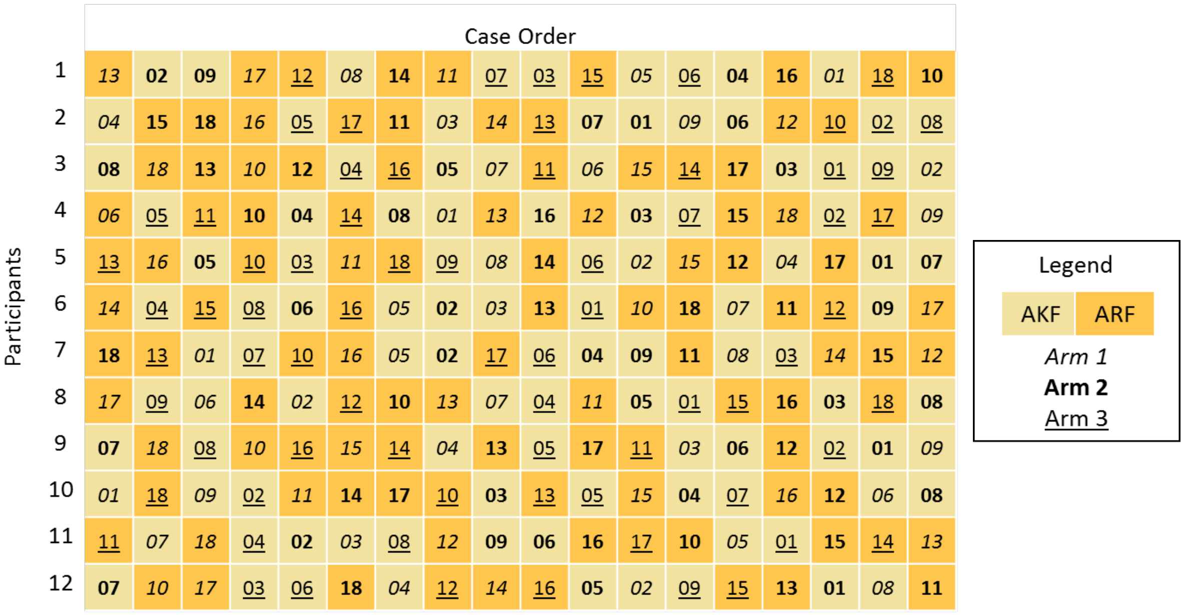
Eighteen patient cases were selected that (1) were admitted to University of Pittsburgh Medical Center ICUs between June 2012 and December 2012 and (2) had a diagnosis of either acute kidney failure (AKF; ICD-9 584.9 or 584.5; 9 cases) or acute respiratory failure (ARF; ICD-9 518.81; 9 cases). The cases were de-identified (De-ID™ Software) to create a limited data set in which all protected health information except for dates and times related to the events were removed. These cases were admitted to the ICU in the seven months after the training cases (Section 5.3).

Study design

The evaluation consisted of three arms that included a control arm (Arm 1) and two intervention arms (Arm 2 and Arm 3). In Arm 1, the procedure for reviewing the case was similar to the procedure used in the training phase. In Arm 2, the selected models were applied to the case and patient data that were predicted to be relevant were highlighted. In Arm 3, patient data was highlighted as in Arm 2 and, in addition, data that were not predicted to be relevant were removed from the interface.

Time to task completion when preparing to present a case at morning rounds was measured in all three arms, and we compared average time to task completion for Arm 1 (the control arm) to both intervention arms. To control for user effects on time to task completion, a fractional factorial design was used to assign 12 participants to one of three arms for each of the 18 patient cases. The assignment of cases to participants is shown in Figure 20.

In addition to measuring time to task completion, Arm 1 was used to collect data items that were sought by the participants for the 18 cases. These manually indicated items constitute a gold standard and was used to evaluate model performance. In Arm 3 (where patient data not predicted to be relevant was removed from the interface), we include extra tasks to evaluate the clinical impact of having some data hidden.



Participants	Case Order																	
1	*13*	**02**	**09**	*17*	12	*08*	**14**	*11*	07	03	15	*05*	06	**04**	**16**	*01*	18	**10**
2	*04*	**15**	**18**	*16*	05	17	**11**	*03*	*14*	13	**07**	**01**	*09*	**06**	*12*	10	02	08
3	**08**	*18*	**13**	*10*	**12**	04	16	**05**	*07*	11	*06*	*15*	14	**17**	**03**	01	09	*02*
4	*06*	05	11	**10**	**04**	14	**08**	*01*	*13*	**16**	*12*	**03**	07	**15**	*18*	02	17	*09*
5	13	*16*	**05**	10	03	*11*	18	09	*08*	**14**	06	*02*	*15*	**12**	*04*	**17**	**01**	**07**
6	*14*	04	15	08	**06**	16	*05*	**02**	*03*	**13**	01	*10*	**18**	*07*	**11**	12	**09**	*17*
7	**18**	13	*01*	07	10	*16*	*05*	**02**	17	06	**04**	**09**	**11**	*08*	03	*14*	**15**	*12*
8	*17*	09	*06*	**14**	*02*	12	**10**	*13*	*07*	04	*11*	**05**	01	15	**16**	**03**	18	**08**
9	**07**	*18*	08	*10*	16	*15*	14	*04*	**13**	05	**17**	11	*03*	**06**	**12**	02	**01**	*09*
10	*01*	18	*09*	02	*11*	**14**	**17**	10	**03**	13	05	*15*	**04**	07	*16*	**12**	*06*	**08**
11	11	*07*	*18*	04	**02**	*03*	08	*12*	**09**	**06**	**16**	17	**10**	*05*	01	**15**	14	*13*
12	**07**	*10*	*17*	03	06	**18**	*04*	12	*14*	16	**05**	*02*	09	15	**13**	**01**	*08*	**11**

Legend

AKF	ARF

Arm 1
Arm 2
Arm 3

Figure 20. Fractional factorial study design for the primary LEMR evaluation. Every case is viewed by all participants, but different cases have different combinations of participants assigned to the three different arms. The case order is randomized. Cases are divided evenly between the two diagnoses.

LEMR interface

Matching this study's three arms, there were three versions of the LEMR interface. The *control* version is the same as was used when collecting training data (Section 5.3); The *highlights* version of the interface is the same as the control version, except patient data predicted to be relevant were highlighted in-place, by changing the background color behind relevant data to yellow. The *highlights only* version of the interface highlights the same patient data as the highlights version and also hides patient data not predicted to be relevant; hidden data cannot be accessed and resulting blank space is compressed. Screenshots of the interface are shown in Figures 12 and 13 in Section 3.6. A slideshow presentation was used to introduce participants to the study objectives and the LEMR interface. The slides are shown in Appendix D.

Participant tasks

Evaluation patient cases were displayed in the LEMR interface and participants evaluated the cases by following instructions to complete the following tasks. In Arm 1, participants completed Tasks 1, 2, 3 and 4 in order; in Arm 2, participants completed Tasks 1, 2 and 3 in order; and in Arm 3, participants completed Tasks 1, 2, 3, 5 and 6 in order. An overview of the tasks that constitute the case review procedures for the three arms is shown in Figure 21, and details of the tasks are given below.

Task 1. For this task a random day between day two of admission to the ICU and the day before discharge from the ICU was selected as the "past patient stay". All available EMR data up until 8:00 am on the day selected for the past patient stay was displayed to the participant. Structured data were shown in graphical time series plots and free-text notes were shown in a separate area in the interface. The participant was instructed to "use the available information to become familiar with the patient case as if they are one of your own patients." All arms use the *control* version of the interface for this task. After becoming familiar with the case, the participant clicked on a button to advance to Task 2.

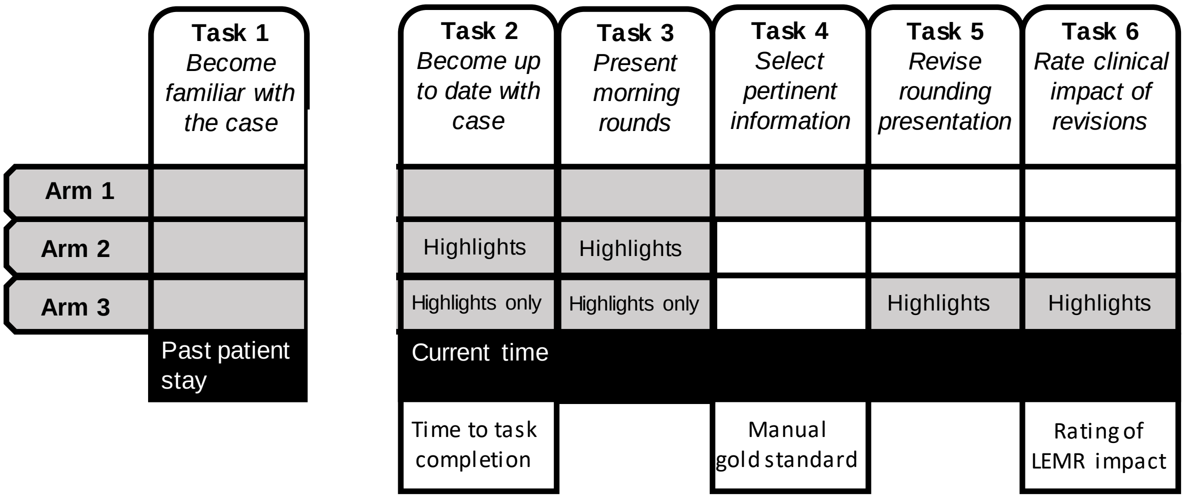
Task 2. An additional day (from 8:00 am on the day selected for the past patient stay to 8:00 am on the next day i.e., "current time") of the patient's EMR data was added to the display. The participant was prompted with "24-hours have passed" and directed to "use the available information to prepare to present the case during morning rounds." For Arm 1, Arm 2, and Arm 3, the *control* version, the *highlights* version, and the *highlights only* version of the LEMR interface are used, respectively. After preparation was complete, the participant clicked on a button to advance to Task 3.

Task 3. The participant was prompted with "now that you are up to date with this patient's problems and latest data, please present the patient as if you were presenting during morning rounds, including pertinent positives and negatives, as well as your assessment and management plan for the day. Try to make it concise." The presentation was recorded with an audio recorder. After finishing the presentation, the participant clicked a button that either advanced to Task 4 (if in Arm 1), advanced to the next patient case (if in Arm 2), or advanced to Task 5 (if in Arm 3).

Task 4 (only for Arm 1). In the interface, each available data item (e.g., glucose levels, insulin dosage regimen) was accompanied with a check box. Clicking on the area associated with data toggled the check box. The participant was directed to "select the information you consider pertinent when preparing to present this case at morning rounds." The participant selected relevant data items by toggling the accompanying check box to the checked state. The participant clicked a button to advance to the next patient case.

Task 5 (only for Arm 3). The participant was shown the case using the *highlights* interface version — i.e., the hidden data were revealed — and was prompted with "additional information is now being displayed. Considering the additional information, if you would like to revise your presentation, please do so now." Revisions to the rounding presentation were recorded using an audio recorder. After finishing the revisions (or opting not to revise), the participant clicked on a button to advance to Task 6.

Task 6 (only for Arm 3). The participant was prompted with "if you revised your presentation, rate the clinical impact those revisions would have on patient care." Clinical impact was selected on a three-point scale: "no impact", "minor impact", and "major impact", and included a fourth option labeled "no revisions". The participant clicked a button to advance to the next patient case.



	Task 1 *Become familiar with the case*	**Task 2** *Become up to date with case*	**Task 3** *Present morning rounds*	**Task 4** *Select pertinent information*	**Task 5** *Revise rounding presentation*	**Task 6** *Rate clinical impact of revisions*
Arm 1						
Arm 2		Highlights	Highlights			
Arm 3		Highlights only	Highlights only		Highlights	Highlights
	Past patient stay	Current time				
		Time to task completion		Manual gold standard		Rating of LEMR impact

Figure 21. Case review tasks for the primary LEMR system evaluation.

After all cases were reviewed, participants completed a modified Unified Theory of Acceptance and Use of Technology (UTAUT) questionnaire. The UTAUT is a theory that aims to explain the acceptance and use of information systems and information technology innovations and subsequent usage behavior [123]. According to the UTAUT theory there are four key constructs: 1) performance expectancy, 2) effort expectancy, 3) social influence, and 4) facilitating conditions. The modified UTAUT questionnaire that was used in the evaluation study is provided in Appendix C.

Models of clinician information seeking behavior applied to determine highlighting

In Section 6.2, we described the training of models of clinician information seeking behavior. This study provides a prospective evaluation of those models. To compare and contrast the two studies, five of the participants participated in both studies, the training study cases were admitted to the ICU in the 17 months before the cases in this study, the training study interface was the same as the control version used in this study, and the training study tasks were the same as the tasks for Arm 1 in this study.

In the training study, sufficient sample sizes were available for building models to predict clinician information seeking behavior of 80 different target variables (i.e., data items). Of these models, 25 of them met the selection criteria of having a precision of at least 0.67 and a recall of at least 0.50. The best performing models were retrained on the entire training data set (178 cases) and applied to the 18 cases of this study. Any target with a predicted probability greater than 0.5 was highlighted in the *highlights* and *highlights only* versions of the LEMR interface.

Data collection

Data collected during the tasks included time to task completion during Task 2, a list of pertinent (i.e., context relevant) data items in Task 4, and, based on data revealed in Task 5, a rating of the LEMR system clinical impact in Task 6.

Data analysis

We performed six analyses to evaluate the LEMR system. They are described below.

(1) *To evaluate the impact of LEMR system highlights on time to task completion when preparing for morning rounds*. This evaluation was performed using a one-way ANOVA with post hoc analysis. A Bartlett test of homogeneity of variance was performed before performing ANOVA to verify that the variance did not differ between groups. The post hoc analysis was performed using Tukey's Honest Significant Difference test which also assumes homogeneity of variance. All three tests were performed in the R statistical computing language using the following functions: Bartlett.test() from the stats package, aov() from the stats package, and HSD.test() from the agricolae package.

(2) *To evaluate the adequacy of highlighted patient data*. First, we compared and summarized the number of patient data items displayed, highlighted, and manually selected in each case during Task 2 through Task 4. Note that patient data were manually selected in Task 4 if a clinician considered it pertinent when preparing a case for morning rounds. Next, we assumed the selected data were the same data clinicians sought as relevant, which allowed us to use this set of manual selections as a gold standard for calculating precision and recall of the highlights. Finally, we compared the performance of model-based highlighting to the performance of random highlighting. To generate a 95% confidence interval for the precision and recall of random highlights, we randomly selected (in each case) *h* data items (where *h* is the number of items

highlighted by the models) from a set of *n* data items (where *n* is the total number of displayed items). Of the *n* data items, the number of positive items was the average number of items manually selected in each case. Then, precision and recall were calculated by evaluating the number of positive items randomly selected. This process was repeated 1,000 times for each case in a to estimate confidence intervals.

(3) *To evaluate the performance of each model that was applied in this study*. Performance is reported using precision and recall.

(4) *To compare the performance of models trained on different training sets (manual selection vs. eye-tracking) and with different levels of personalization (general vs. semi-personalized)*. Model training and selection were described in Section 6.2. Model performance was compared using Wilcoxon signed rank test. To determine if models trained on the manual selection data set perform better than models trained on the eye-tracking data set, the AUROCs of the models for each data item were tested against an alternative hypothesis: the mean AUROC for manually trained models is greater than the mean AUROC for eye-tracking trained models. This test was conducted twice, once for generalized models and a second time for semi-personalized models. Semi-personalized models were trained and evaluated using only the data from the five clinicians who participated in both the training (Section 5.3) and evaluation (Section 7.1) studies. Personalization is achieved through the inclusion of five Boolean variables, where the variable corresponding to a participant is set to true if that participant provided the manually selected target labels. To determine if personalization changes model performance, Wilcoxon signed rank test was performed a third time to compare the AUROC performance of manually trained generalized models to manually trained semi-personalized models against a two-sided, alternative hypothesis.

Wilcoxon signed ranked tests were performed using the R function wilcox.test() from the stats package.

(5) *To estimate the clinical impact of clinicians not seeing any of the patient data that were not highlighted by the model*. Impact is reported as summary statistics from the ratings that participants provided in Task 6.

(6) *To evaluate user acceptance and use of the LEMR system*. Participants answered relevant questions from the unified theory of acceptance and use of technology (UTAUT) model. The results are reported using box-and-whisker plots.

7.1.2 Results

Results are reported for six different analyses. The first two analyses reported are to the study hypothesis and include impact on time to task completion and adequacy of highlighted patient data.

Impact on time to task completion

Time to task completion was measured when participants were preparing to present each of the 18 patient cases at morning rounds (Task 2). The Bartlett test of homogeneity of variances showed no statistically significant difference in the variance of time to task completion for the three arms (Bartlett's K-squared = 2.1683, df = 2, p-value = 0.3382); therefore, both ANOVA and Tukey’s tests are appropriate. The ANOVA test showed a statistically significant difference in time to task completion among the three arms of this experiment (Table 14). Summary statistics of the time to task completion in each arm are shown in Table 15, and the pairwise results of Tukey’s Honest Significant Difference test are shown in Table 16. The times in Arm 1 and Arm 2 were not statistically significantly different; thus, the data do not support that in-place highlighting saves time when preparing for morning rounds. In contrast, the times in Arm 3 were statistically significantly smaller than the times in Arm 1 ($\alpha = 0.1$; p-value = 0.0912); thus, the data suggest that clinicians used less time when preparing to present morning rounds if viewing a case in Arm 3.

We found that clinicians take less time to review data when less data are available to review (i.e., when data are hidden in Arm 3). While this result is unsurprising, it does not provide a sense of whether the highlights are beneficial or not, because we do not know if the highlighted patient data was adequate in satisfying the needs of clinicians. In the next section we present an evaluation on the adequacy of highlighted patient data.

Table 14. ANOVA comparing time to task completion in the three arms of the evaluation study.

Source	*SS*	*df*	*MS*	*F*	*Sig.*
Between	40,356	2	20,178	3.821	0.0234
Within	1,124,869	213	5,281		
Total	1,165,225	215			

Table 15. Mean time to task completion in the three arms of the evaluation study.

Arm	Mean time to task completion (sec)	std	n	Min	Max	Q25	Q50	Q75
1	140.4	76.9	72	9.1	513.3	95.1	128.3	181.6
2	146.4	75.3	72	22.9	362.2	94.8	125.2	190.3
3	114.9	65.3	72	21.5	334.6	71.1	100.6	154.5

Table 16. Results of Tukey's Honestly Significant Difference post hoc test on time to task completion. Arm 1 is a control arm, using the plain LEMR interface. Arm 2 is an intervention arm, where patient data predicted to be relevant are highlighted in the LEMR interface. Arm 3 is also an intervention arm, where, in addition to the highlights of Arm 2, patient data are removed from the LEMR interface if not predicted to be relevant.

Comparison (arms)	Difference (95% CI)	p-value	Is significant
1 – 2	-6.0 (-34.6, 22.5)	0.8718	No
1 – 3	25.5 (-3.1, 54.1)	0.0912	At α = 0.1
2 – 3	31.5 (3.0, 60.1)	0.0265	At α = 0.05

Adequacy of highlighted patient data

First, we analyzed the data to determine if the right proportion of available data were highlighted. Figure 22 shows the number of data items available for each of the 18 patient cases, the number of data items highlighted for each case, and the minimum, maximum, and average number of data items manually selected as relevant for each case. We anticipated that the number of items selected was substantially smaller than the number of items available, because the LEMR system is based on the premise that there is a subset of all available patient data that is context-relevant and will be sought by clinicians when reviewing a case. Supporting this premise, we found that the cases had, on average, 108.9 data items available, and participants selected (sought as relevant) 22.6 of those items.

Next, we examined if the number of highlighted data items matched the number of items sought as relevant for each case. We found the number of highlighted items to be within the range of the number of items selected (sought as relevant) by participants for 14 of the cases (averaged highlighted = 15.1 and average selected = 15.7). The remaining four cases were within two and three items of the maximum, and two and five items of the minimum number of items selected for each patient case (average highlighted = 14.5 and average selected = 17.4).

Finally, we examined the precision and recall of the highlighted data. Table 17 shows these performance metrics computed in two ways. First, it shows precision and recall when only considering the patient data for which a predictive model was available. In other words, if the model performance for a data item did not make the inclusion thresholds in Section 6.2, then that item did not have a model in the evaluation study and, therefore, is not considered in the first analysis. Second, we report precision and recall when considering all available patient data. We call these two analyses, ‘model active patient data’ and ‘all patient data’, respectively.

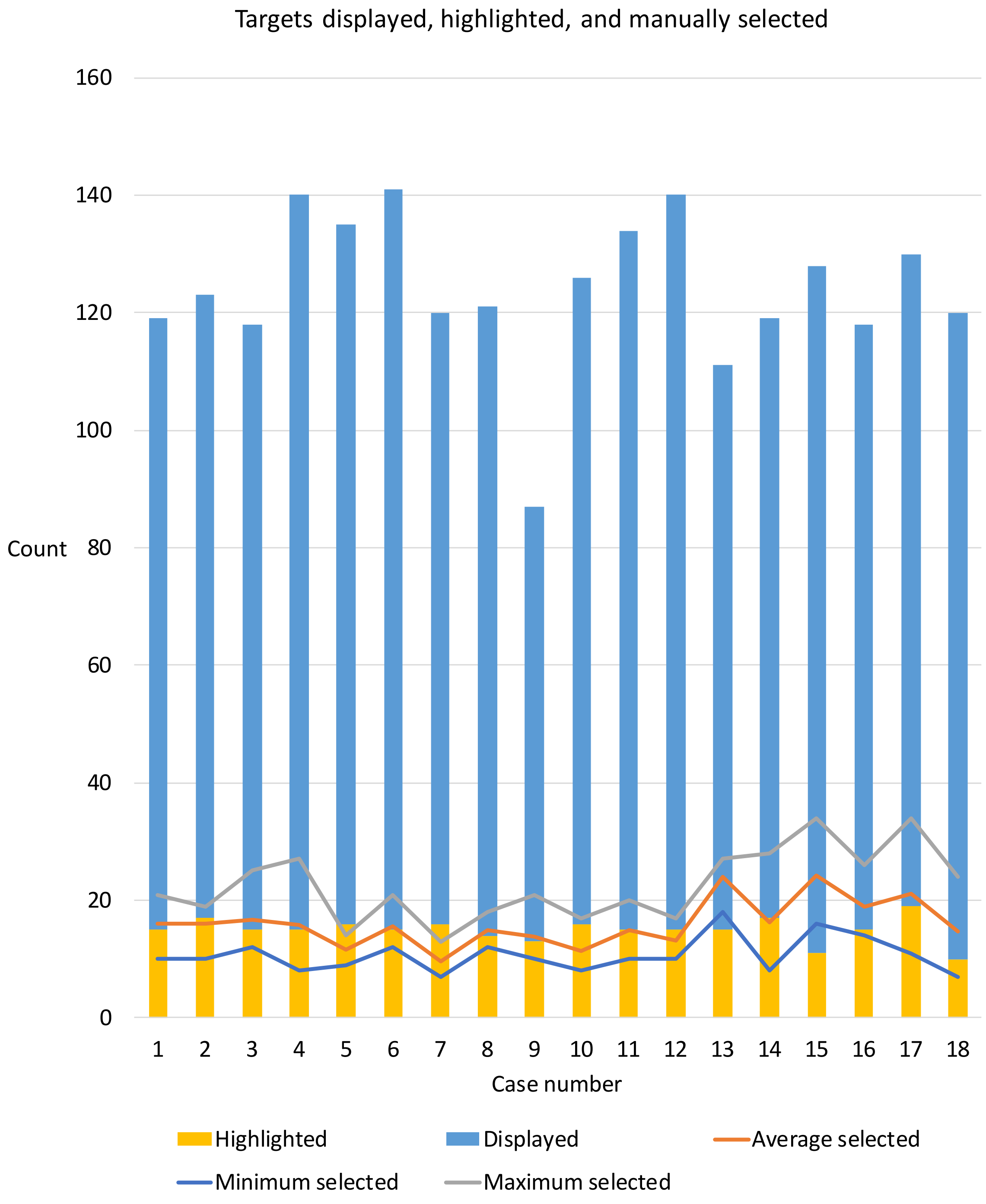


Figure 22. Summary of targets displayed, highlighted, and manually selected during the evaluation study. Participants selected the data they considered pertinent when preparing to present each case at morning rounds. We assume that selected (pertinent) data are the data clinicians sought as relevant.

The results show that the models do not identify all the patient data that the participants seek when preparing for morning rounds. The standard of "all the patient data" is a high bar that this experiment was very unlikely to achieve. On the opposite end of the performance spectrum is randomly selecting data to highlight. To ensure that the models were informative and performing better than chance, we randomly selected data to estimate precision and recall confidence intervals for random performance. The resulting intervals are shown in Table 18. The confidence intervals do not overlap with the model performance reported in Table 17, providing support that the models are performing better than random.

Table 17. Overall performance of models applied during the evaluation study.

	Precision (95% CI)	Recall (95% CI)
Model active patient data	0.52 (0.49, 0.54)	0.77 (0.75, 0.80)
All patient data	0.52 (0.49, 0.54)	0.43 (0.41, 0.45)

Table 18. Mean and confidence interval estimates of random selection.

	Precision (95% CI)	Recall (95% CI)
Random highlights of all patient data	0.15 (0.00, 0.33)	0.14 (0.00, 0.29)

Evaluation of each model

As was shown in Section 6.2, model performance varies greatly between target variables. Precision and recall performance of each model applied to highlight data in this study is shown in Figure 23.

Comparison of models trained on different data sets

The results of the Wilcoxon signed rank tests show that the AUROCs of models trained on manually selections and the AUROCs of models trained on eye-tracking selections are not statistically significantly different for both general models (W = 2198, p-value = 0.690) and semi-personalized models (W = 852, p-value = 0.310). These results show promise for using eye-tracking as an automatic means of training LEMR system models.

The results of the Wilcoxon signed rank tests also show that there is no statistically significant difference between the AUROCs of manually trained general models and manually trained semi-personalized models (W = 688, p-value = 0.283). These results indicate that personalized models do not perform better; however, the sample size for this test was small and included data for only five participants. The AUROC performance of all 80 models trained in Section 6.2 and their performance when applied to the manual selection gold standard collected in this study are shown in Appendix E.

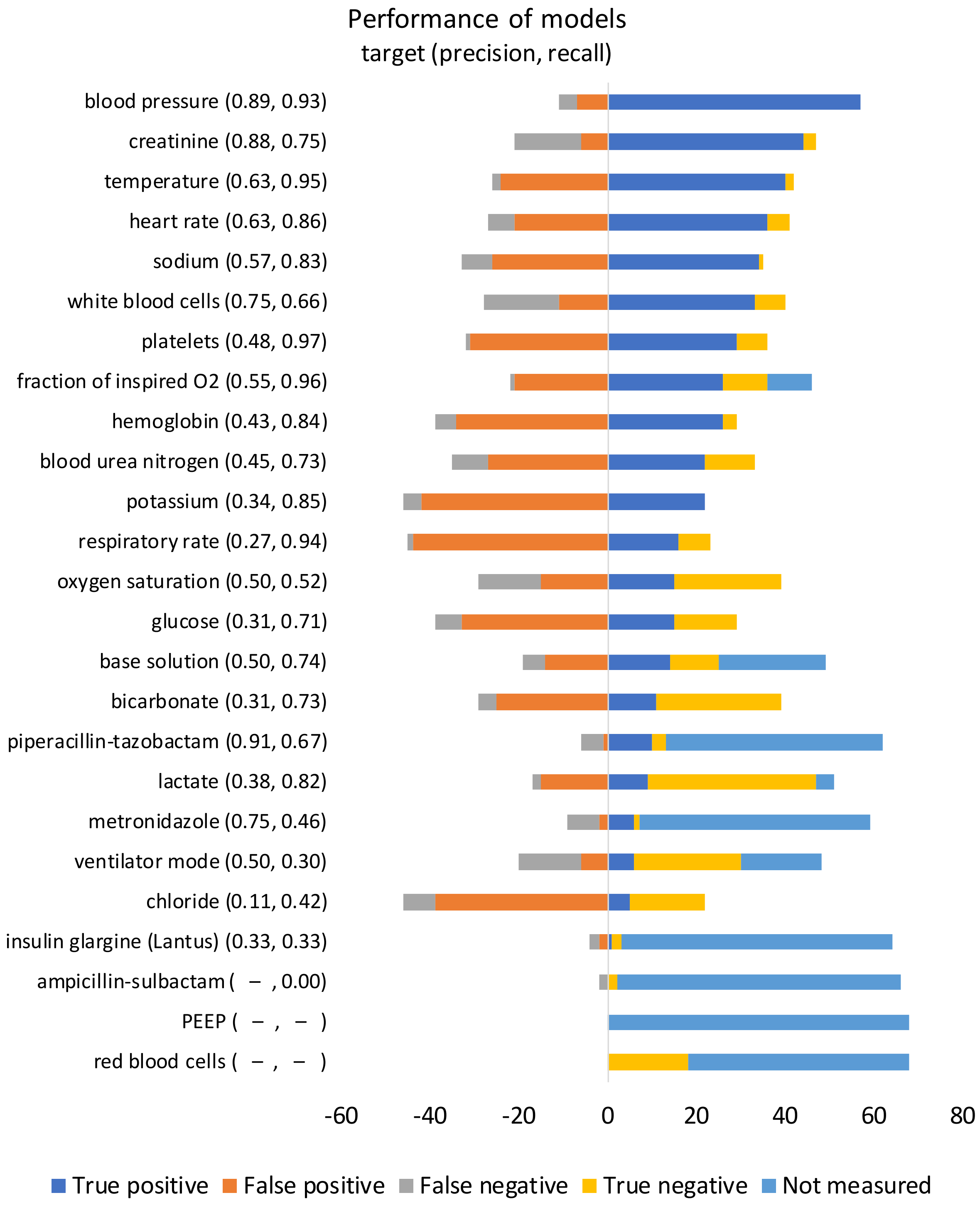


Figure 23. Performance of models applied during the evaluation study. All counts are positive values; false positives and false negatives are plotted negatively to show contrast with true positives and true negatives.

Clinical impact

A summary of self-rated clinical impact of not seeing patient data that were not highlighted is shown in Table 19. In summary, in over half the cases (54.1%), the participants when shown the hidden data did not revise their rounding presentation, or the revision had no clinical impact. However, in 18.1% of the cases, the participants made a revision to their rounding presentation that would have had a major impact on clinical care of the patient. It may be the case that the revisions causing a major impact in patient care where due to just a few data items that did not have a model; further assessment is needed.

Table 19. Clinical impact of not seeing the patient data that were not highlighted in the evaluation study.

	I did not revise	1. No impact	2. Minor impact	3. Major impact
Counts	33 (45.8%)	6 (8.3%)	20 (27.8%)	13 (18.1%)

Acceptance and use of LEMR system

After completing all case tasks, participants completed a modified version of the unified theory of acceptance and use of technology (UTAUT) questionnaire. The results of the questionnaire are shown in Figure 24. For performance expectancy, effort expectancy, attitude, and self-efficacy, the higher the score the better; for anxiety, the lower the score the better. Overall, participants had a slightly positive feeling on their expected performance when using the LEMR system, a more positive feeling on the effort required to use the LEMR interface, and they are at ease (i.e., not anxious) when using it.

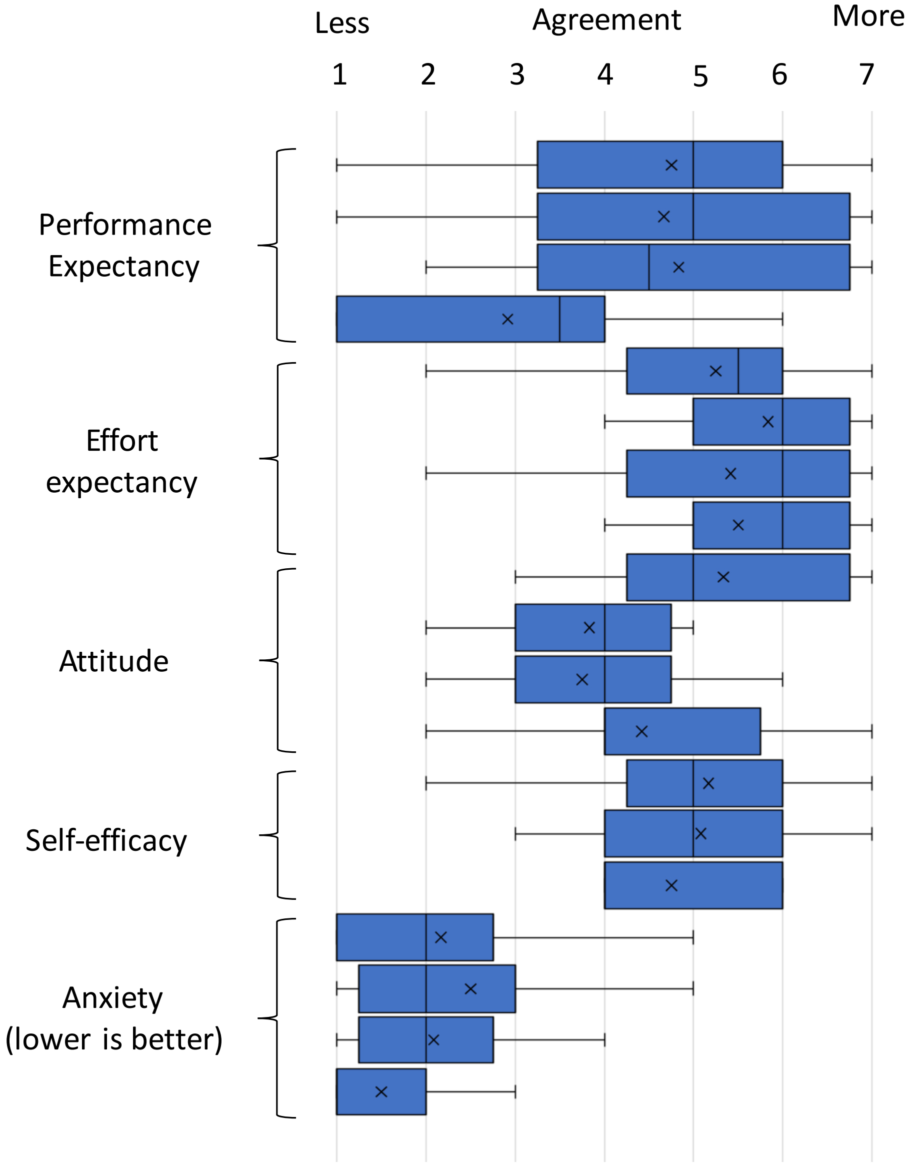


Figure 24. Results from a modified Unified Theory of Acceptance and Use of Technology (UTAUT) questionnaire.

7.2 A DISSCUSSION OF APPLYING MODELS TO DIRECT THE FUTURE DISPLAY OF PATIENT DATA

The results reported in this chapter show that a LEMR system that applies models to direct the display of patient data may reduce the time it takes for clinicians to prepare for morning rounds. Average time to task completion was less when model output was used to highlight data that are predicted to be sought as relevant and to hide data that are not predicted to be sought as relevant. Hiding data comes at the risk of hiding data that are relevant. To investigate this issue, we calculated precision and recall of the highlights and found overall recall to be 43%. At this recall level, the system is missing five or six out of every ten data items a clinician seeks. Encouragingly, however, even with this modest level of recall, not seeing the hidden data only had a major clinical impact for 18% of the cases and had no impact for over half of the cases. The results also show prospects for improvement. When considering only the 25 data items that had an active model during the study, recall increases to nearly 80%. This result suggests that with more training data and more models meeting inclusion threshold, the overall recall is likely to increase.

The second part of evaluating model performance is precision. Results showed a precision of 52%, meaning about half of the highlighted data were sought as relevant. Model performance was substantially better than randomly selected highlights.

To increase model coverage of data items and contexts, a fully developed LEMR system would be trained on tens of thousands of cases. To collect thousands of training cases, automatic methods for observing clinician information seeking behavior are needed. The current study supports that eye-tracking may be a viable, automatic alternative to manually labeled training data, as shown by performance that was not significantly worse than manually trained models.

Personalization of the models did not seem to improve performance. This may be due to lack of training data for each individual user or due to higher variance in performance because of a smaller evaluation set (i.e., data from five participants instead of twelve).

In conclusion, our hypothesis was that the LEMR system will yield the following results on a set of test cases: (1) on average clinicians will use less time in preparing to present a patient case at morning rounds, and (2) clinicians will judge that the system highlights all the information that they would use in each case for the specified task. This hypothesis was partially supported by our results. Addressing part 1, time to task completion was less when models were applied to determine which data to highlight and which data to hide, but was not reduced when only highlighting. It is important to emphasize that we do not expect a working LEMR system to ever completely 'hide' information as in Arm 3, which was introduced for the purpose of experimentation. In a clinical setting, clinicians will always retain access to all the information they would otherwise have access. Highlighting in some situations might be done in-place, as it was in Arm 2, and, at other times, be done as an optional HID (highlighted information display). Addressing part 2, the models did not highlight all the information clinicians would use, but compared to random, performed substantially better.

8.0 DISSCUSSION

Reducing cognitive load is a top priority in improving EMR usability [31]. The LEMR (learning electronic medical record) system presented in this dissertation observes clinician information seeking behavior and applies it to direct the future display of patient data. Patient data are highlighted if statistical models predict that a clinician will seek them as relevant; thus, highlighting provides concise and context sensitive data that are uncluttered by extraneous information. The LEMR system has a major advantage over other context sensitive EMR systems in that it is data-driven rather than expert-driven. This distinction potentially enables the LEMR system to be more readably adaptable to different contexts and to changes in care practices.

The LEMR system exemplifies a LHS (learning health system) approach to EMR system design (as described in Section 1.5). When clinicians use a LEMR system in clinical practice, the system observes clinician information seeking behavior, which is an example of *practice to data*. The data are then used to train a model of clinician information seeking behavior — a form of *data to knowledge*. The knowledge of the model is applied to direct the future display of patient data — an instance of *knowledge to practice*. How clinicians seek information using the directed display is then observed, so the learning cycle can continue indefinitely to drive improvement.

The development of this system was divided into five phases. First (in Chapter 3), we developed a prototype LEMR interface that served as a test bed for LEMR experimentation. The LEMR interface was evaluated in two studies: a think aloud study and a usability study. The results from these studies were used to iteratively improve the interface.

Second (in Chapter 4), we evaluated the accuracy of an inexpensive eye-tracking device and developed an automatic method for mapping eye gaze to patient data displayed in the LEMR interface. In two studies we showed that an inexpensive eye-tracking device can perform as well as a costlier device intended for research and that an automatic mapping method accurately captures the patient information a user is viewing.

Third (in Chapter 5), we collected observations of clinician information seeking behavior in the LEMR system. In three studies we evaluated both manual and automatic methods for collecting the observations to train the system. In the last of those studies, we used both manual selection and (automatic) eye-tracking methods to assign target values to the patient data that clinicians sought as relevant in a set of 178 patient cases.

Fourth (in Chapter 6), we applied machine learning to the training data to model clinician information seeking behavior from a manual set of training cases. In total, there were enough training data to train models for 80 different data items (i.e., target variables). Twenty-five of these models met the performance criteria for inclusion in the evaluation study.

Fifth (in Chapter 7), we applied the models to direct the display of patient data in a prospective evaluation of the LEMR system. The evaluation found that, when the models were applied to highlight data predicted to be sought and hide data not predicted to be sought, the system reduced time to task completion for clinicians who are preparing to present cases at morning rounds. More work is needed before the models adequately highlight all the patient data clinicians seek, but at current performance the study clinicians assessed that differences (between their rounding presentation when only seeing the highlighted data and their revised presentation when seeing all data) do not have a major clinical impact on patient care in over 80% of the cases.

LEMR systems are an important area of computerized clinical decision support. As such, they should be evaluated by the same criteria as other clinical decision support systems. Ideally, a LEMR system would anticipate user needs, deliver support in a timely manner, fit into the user's workflow, and maintain an effective knowledge base. These are four important features of effective clinical decision support in general [17]. The LEMR system also addresses three grand challenges of clinical decision support as described by Sittig et al. [13]: it summarizes patient level data, prioritizes and filters recommendations (highlights) to the user, and combines recommendations (highlights) for patients with comorbidities.

The remainder of this chapter is divided into three parts: (1) insights into LEMR systems that were gained while completing this dissertation, (2) future work inspired by this dissertation research, including its limitations, and (3) concluding remarks.

8.1 INSIGHTS INTO LEMR SYSTEMS

This section lists insights into LEMR systems that were gained while completing this dissertation. The list is not exhaustive, but presents key concepts that may help further develop LEMR systems.

What data are highlighted

Insight 1. There are three conditions in which context-relevant patient data are highlighted for an EMR user.

Condition 1: The user knows a data item is relevant. In this condition, the LEMR system should focus on ease of access: highlighting to save time and cognitive effort. It should reduce the number of clicks required to retrieve data and display together the different data a clinician combines in making a decision. Doing so will reduce screen switching time, will reduce the memory load on the user, and will reduce the need for writing details down on paper.

Condition 2: The user does not know a data item is relevant. In this scenario, the LEMR system should focus on increasing situational awareness: highlighting to focus attention on relevant, overlooked data. The overlooked data may be an ignored laboratory test result, an unnoticed negative trend across temporal results, an imaging study that the clinician did not know was available, or any of many other aspects of patient data. Data may be overlooked due to reasons such as anchoring to a certain diagnosis or experiencing information overload. Highlighting to focus attention on relevant, overlooked data may result in the clinician considering new or alternative diagnoses and associated treatment plans.

Condition 3: The user forgets a data item is relevant. In this situation, as in the second one, the LEMR system should use highlighting to focus attention on relevant, overlooked data.

Insight 2. With three conditions of highlighting patient data, three different methods are applied to determine which data to highlight.

In condition 1, a model is applied to predict when the user knows a data item is relevant. This model is a clinician-specific model that is personalized to a clinician's own information seeking behavior in different contexts. Personalization is desirable because it will make the EMR interaction more seamless.

In condition 2, two models are applied: a clinician-specific model and a general model. The clinician-specific model predicts the data the current clinician user will seek. The general model predicts the data a population of similar clinicians would seek in the same context (crowd wisdom [47]). Data that are not predicted to be sought as relevant by the clinician-specific model but are predicted to be sought in the general model could be assumed by the system to be data items that the current clinician user may not know are relevant.

In condition 3, the data items that are highlighted are those predicted by the clinician-specific model to be sought by the current user in the current case, but which the user has not yet viewed upon reaching the end of viewing the case. This form of highlighting would occur toward the end of viewing a case, compared to the highlighting described for conditions 1 and 2, which occurs at the beginning of case viewing. Such late highlighting has the advantage of being relatively non-directive; data items are only highlighted when they are believed relevant (by the system) and not viewed by the user (according to eye tracking or other methods).

Combinations of the above conditions may be useful to investigate as well. One possibility involves combining conditions 2 and 3. Here both the clinician-specific and the general model are applied to predict data seeking behavior and the union of the data items predicted by these two models is combined using set union to produce set *U*. The data in *U* that is not viewed by the clinician would be highlighted toward the end of viewing the case in order to emphasize data items that the user either did not know or remember are relevant.

Insight 3**.** A clinician cannot be over reliant on highlighting when seeking data they know to be relevant.

If the clinician knows a data item is relevant, over-reliance on LEMR-system highlighting is not a concern because the clinician knows what data they seek; therefore, if data they seek are not highlighted, then they will use traditional EMR navigation to find it. Traditional EMR navigation is how most clinician information seeking currently occurs.

If the clinician does not know a data item is relevant or forgets a data item is relevant, then there is a risk of over-reliance because the clinician may assume the system highlights all relevant patient data. However, using current EMR systems, clinicians are missing relevant data, which may lead to problems, including treatment delays [2,26,30]. The highlights provided by a LEMR system could help reduce such delays. Thus, there appears to be a tradeoff between costs (e.g., over-reliance) and benefits (e.g., reduction of treatment delays) in using a LEMR system. However, the use of late highlighting, as described above for condition 3, may reduce the costs of highlighting, by making it less directive and intrusive of the usual clinical workflow. Additional issues about the timing of data highlighting are discussed in the next insight as well.

When are data highlighted

Insight 4. Relevant data are those data necessary for making a clinical decision well, when the user is making that decision; therefore, the current task of the EMR user must be taken into account.

EMR systems store many data because clinicians use many different types of data when making different types of clinical decisions. Which data are relevant depends on the decisions being made. Data relevant to one treatment decision (ordering insulin) may not be relevant to a second treatment decision (perform a spontaneous breathing trial), even if both treatment decisions are for the same patient at the same time. The LEMR system should highlight the data relevant to a decision while that decision is being made, then highlight the data relevant to the next decision while that decision is being made.

Doing otherwise, highlighting different relevant data for many decisions at the same time, might confuse EMR users and add cognitive burden, due to interruption and multi-tasking, as the user starts thinking of other decisions that need to be made. This insight comes with the acknowledgement that some tasks are very broad and may have a large set of relevant data, e.g., differential diagnosis.

Insight 5. Since current task is needed to determine when to highlight data, the current EMR task should be determined as part of the training data.

Some tasks are easily captured by an EMR system, as for example if a user starts to place an order for insulin, then then system can infer that the current task is "ordering insulin." Other tasks, particularly when reviewing patient data, may be less obvious to capture. The LEMR system may do so through a combination of active user specification (such as clicking on the current task from a list of potential tasks), interaction with LEMR-system highlighting, or estimation using machine-learning-driven prediction of the task; the model would estimate the current task, based on the data being sought and the EMR actions being taken.

Where are data highlighted

Insight 6. Where to highlight relevant data depends on what and when they are highlighted.

Relevant data may be highlighted in different places, for example, in-place by changing a data item's background color, in a HID (highlighted information display) where relevant data are shown together, near other data that are relevant to a decision, or in an alert. In-place highlighting is the most subtle and may work well for ensuring a data item (e.g., downward trending blood urea nitrogen results) are not overlooked when the clinician is assessing other laboratory test results. A HID is useful for consolidating data relevant to a decision in one place. Decision relevant data may be grouped in electronic “cards” where each card is for a single decision or task. Which cards to display could be determined using either a model of the tasks a user will perform next or through user specification of the tasks they wish to complete currently. Neighborhood highlights will show relevant data next to another piece of relevant data. For example, if a clinician is on the medication ordering screen and evaluating warfarin dosing, the results of bleeding tests could be displayed next to the warfarin dosing regimen. Finally, highlighting can be achieved through alerting. Alerting is interruptive and should only be used when the model has high certainty that a relevant data item was overlooked. This point illustrates, however, that highlighting can (and we believe should) be viewed quite broadly.

Insight 7. Combinations of highlighting methods may be used.

The appropriate method for highlighting a data item may depend on the condition in which it is highlighted. If the system predicts that a clinician knows a data item is relevant for a task, then perhaps it should be highlighted in a HID with all the other data relevant for that task. On the other hand, if the system believes that a user forgot to view data that are relevant for a task, then it may be most appropriate to notify the clinician once it becomes probable that the item will not be viewed in the near term.

How are training data collected

Insight 8. A LEMR system works interactively with users and continuously observes clinician information seeking behavior.

Currently, clinicians interact with an EMR system to retrieve the data they seek. Interaction details, such as page visits, are usually captured as meta-data [124,125]. These meta-data can function as high-level observations of clinician information seeking behavior, and, therefore, can be used when training LEMR system models of clinician information seeking behavior.

In addition to interactions like those with current EMR systems, clinicians will interact in new ways with LEMR systems. The new interactions will be captured as more comprehensive meta-data. For the first of three examples, if a LEMR system displays in its HID a series of cards — where each card contains the patient data relevant to a single clinical decision or task — the use of those cards will be observed by the system. Some cards may be dismissed with no action. Other cards may be used as is. A few cards may be missing a data item needed for the decision and a user may manually add the item to the card using system functionalities, such as a search bar.

For the second example, the LEMR system may sometimes highlight context-relevant data in the same 'neighborhood' together. The example provided in Insight 6 was displaying the results of blood clotting tests next to a patient's warfarin dosing regimen. To determine if model predictions are correct and data are appropriately displayed together, the system could solicit feedback from the user. In this example, the names of the blood clotting tests could be present, but the results could be blurred out. The user can clear the blur by clicking on it. So if the user clicks to reveal a test result, then the test was appropriately highlighted. If the user does not click to reveal, then the test was not appropriately highlighted. This is active learning and it, or similar methods, will occasionally be applied selectively to model predictions with low certainty.

For the third and final example, this dissertation presented how eye-tracking may be used to infer clinician information seeking behavior. Eye-tracking may also prove useful when determining when a clinician forgets data are relevant (i.e., an item was predicted to be sought, but was never viewed). In the long run, we think eye tracking is likely to be the most practical way to obtain extensive and detailed training data for the type LEMR system described here. It seems plausible that eventually eye tracking will become highly accurate and sufficiently inexpensive to incorporate routinely into computer display monitors. At that point, the potential for using eye tracking for LEMR training would be high.

8.2 FUTURE WORK TO ADDRESS LIMITATIONS

This dissertation research has explored the initial design, implementation, and evaluation of a LEMR system. Before such a system is ready for clinical use, current limitations must be addressed with future work.

LEMR interface

The LEMR system was developed in conjunction with a LEMR display interface. This interface had limited functionality that prevents it from being classified as a full EMR system. Namely, the interface does not have data input functionality — it only displays patient data. This was sufficient for LEMR system experiments reported here, but future work needs to involve real EMR systems that are being actively used.

The LEMR system could be tightly integrated with an existing commercial EMR or it could be a standalone interface providing clinicians with a second way of accessing the patient data they seek. Imagine a LEMR tablet device that can be carried around during rounds or to a patient's bedside. The lightweight LEMR interface would adapt to show the clinician just the data they are predicted to seek. A LEMR interface could also be useful to clinicians who want remote EMR access on their mobile phones. When a clinician is at home and gets a phone call about a treatment decision, a LEMR interface on their phone could show the patient data they are likely to need to efficiently make the decision; nevertheless, at all times the user could access any of the EMR data. A final possibility is to develop a LEMR Fast Healthcare Interoperability Resources (FHIR) application [126]. FHIR has the potential to revolutionize the healthcare information technology space, as EMR vendors open app stores. A LEMR system could be made available as an app.

To address issues raised in Section 8.1, more experiments are needed to determine when and where to highlight patient data to maximize effectiveness and user acceptance.

Eye-tracking

Inexpensive and automatic eye-tracking is a promising method for observing clinician information seeking behaviors. A limitation is that this dissertation applied eye-tracking in a laboratory setting on an in-house interface. It is important to determine how to apply the technology in a hospital and to determine if the results will hold up in a dynamic environment.

If eye-tracking reaches widespread adoption, like the cameras that are now found on almost every laptop and mobile phone, then we will enter a new era of eye-tracking for clinical decision support [119]. For example, if data seen (as captured by an eye-tracking device) is a reasonable approximation for what information a clinician knows about a patient case, then perhaps we can estimate which diagnoses a clinician has considered for a case and which diagnoses the clinician has not considered. If based on characteristics of a patient case, an unconsidered diagnosis is more likely than a considered diagnosis, then the LEMR system can highlight for the clinician data that are suggestive of the unconsidered diagnosis.

Observing

In this dissertation, the primary means of observing clinician information seeking behavior was a manual process. This is a limitation that needs to be addressed through the continued development of automatic labeling methods. We have discussed the promise of eye-tracking and, in Section 8.1, discussed using EMR and LEMR meta-data. Additional methods should be developed and studied.

To address issues raised in Section 8.1, we believe that at least three lines of observational research should be pursued. The first is descriptive research into clinician information seeking behavior: (a) how often are patient data sought; (b) how does information seeking vary between contexts (clinicians, tasks and patient cases); and (c) how much data are known to be relevant, not known to be relevant, and forgotten that they are relevant? Second, research into determining the discrepancies between the data a model predicts a clinician will seek and the data they actually seek. Third, research into methods of determining what is the current clinical task.

Modeling

The models trained for this dissertation address a problem that is a bit unusual: what patient data will be sought as relevant. The dissertation applied traditional model learning methods, including logistic regression, support vector machines, and random forests. While these methods produced positive results, more sophisticated approaches, such as hierarchical learning [127], are applicable to model a wider range of clinical contexts (different types of clinicians, performing different clinical tasks, for any patient case). Hierarchical modeling will allow gathering and using training data across a hierarchy of contexts, including across hospitals, clinical departments, specific wards, and even specific clinicians. LEMR models to predict information seeking for a given clinician will be trained using *all* the available data, with the data more specific to the clinician given greater weight.

To get to a wider range of clinical contexts, larger training sets are needed. Modestly sized training sets are a limitation of this dissertation that will be addressed in future work. Eventually, a method for observing clinician information seeking behavior will be implemented in a clinical environment, so obtaining a plethora of training data will be an ongoing, natural by-product of the care experience.

LEMR system evaluation

We evaluated the LEMR system on time to task completion. While it is important to address the time clinicians spend using an EMR system, this is not a measure of whether LEMR system will make patient care safer. The gains in patient safety are in insuring that a clinician does not overlook important information (e.g., miss an important test result [2] or overlook international travel of the patient to a region with Ebola [128]), and it still needs further investigation. The current LEMR system highlights data predicted to be sought as relevant by intensivists who are preparing to present patient cases for morning rounds. The patient cases each have either AKF or ARF as a diagnosis upon ICU admission. The limitations of this initial work suggest additional types of evaluations.

The evaluation study (Section 7.1) has several limitations. (1) We only tested one means of highlighting: in-place (both with and without hiding surrounding data). Other approaches such as HID's, neighborhood highlighting, and alerting, as well as a combination of highlighting methods, should be explored. (2) Our experimental context was limited. Future work should seek to add different types of clinicians, additional clinical tasks, and more patient cases. (3) We did not measure the LEMR system's impact on clinician cognitive load. Additional experiments should be conducted to test (a) if the LEMR system highlights affect clinician cognitive load, (b) if clinicians using the LEMR system succumb to automation bias and become over reliant on highlighting [129], and (c) if the LEMR system highlights improve medical decision making.

8.3 CONCLUSIONS

Learning electronic medical systems, like the LEMR system, are a LHS approach to improving EMRs. This dissertation shows that LEMRs can reduce the time it takes for clinicians to use EMRs while highlighting about half of the patient data they seek. Highlighting patient data was explored in the data-rich ICU environment. The LEMR system may prove to be equally or more useful in other clinical environments, like ambulatory care, where many of the patients have chronic conditions and decades of history recorded in an ambulatory EMR.

Regular automated observation of clinician information seeking behavior opens many possibilities for supporting clinical decision support, including intelligent alerts, automated documentation, and LEMR system highlights.

The current dissertation describes an initial investigation of LEMR systems. The potential impact of LEMR systems on the future of EMR systems in particular and clinical care more broadly seems substantial. We hope that this dissertation research proves useful in helping realize that potential.

APPENDIX A. LIST OF CONDUCTED STUDIES

To test the hypothesis and complete the specific aims in Section 1.3, we conducted a series of studies. Table 20 provides a high-level overview of each of them.

Table 20. List of studies conducted.

Section	Date Conducted	Purpose	Summary of Results
Developing a LEMR Interface			
3.4	11/7/2014	Elicit feedback on a prototype LEMR interface and the LEMR concept.	Potential issues were identified, and interface changes were made to address the issues. An individual's inefficiencies could be lessened by the LEMR and the addition of outlier detection may be beneficial.
3.5	2/5/2015-2/10/2015	Elicit feedback on the LEMR system concept and test the usability of a prototype interface.	Participants were enthusiastic about an EMR that learns from user behavior and provided design recommendations.
Developing Automatic Eye-Tracking for the LEMR Interface			
4.1	3/24/2016-3/28/2016	Evaluate two different eye-tracking devices.	The accuracy of an inexpensive eye-tracking device performs at least as well as a more expensive one.
4.2	5/2/2016	Evaluate a method that maps eye-tracking data to graphical elements in the LEMR interface.	This mapping method has high accuracy after participants become familiar with the LEMR interface.

Table 20 (continued).

Observing Clinician Information Seeking Behaviors			
5.1	4/8/2014	Collect a preliminary training data set of labelled EMR cases.	A clinician manually labeled the laboratory tests that he used when assessing 59 patient cases.
5.2	8/15/2016-9/7/2016	Test using eye-tracking as an automatic approach for observing clinician information seeking behavior.	Eye-tracking performance was moderate; thus, the primary training data set will be collected using both manual and eye-tracking methods.
5.3	8/15/2017-10/17/2017	Collect a **primary** training data set of labeled EMR cases.	176 cases were manually labeled by 11 critical care fellows. 147 of the cases were also automatically labeled via eye-tracking.
Modeling Clinician Information Seeking Behavior			
6.1	1/16/2015-1/30/2015	Test the feasibility and accuracy of LEMR models using a preliminary data set.	Model performance suggests that we can predict the laboratory tests that a clinician will seek as relevant.
6.2	11/1/2017-2/2/2018	Train models of clinician information seeking behavior for the **primary** LEMR evaluation study.	Models were developed for data items with enough training data. The 25 best performing models were applied in the LEMR evaluation study.
Applying a Model to Direct the Future Display of Patient Data			
7.1	2/8/2018-5/6/2018	Evaluate LEMR system impact on time to task completion and the adequacy of highlights during a **primary** LEMR evaluation study.	The LEMR system required less clinician time to use when applying models to highlight and hide patient data. The models predicted nearly half of the sought-after patient data, which is significantly better than random.

APPENDIX B. SOFTWARE DEVELOPED

LEMRinterface: A web interface written using the Bitnami Django Stack that displays EMR data in a temporal fashion.

(https://github.com/ajk77/LEMRinterface)

EyeBrowserPy: Eye (gaze) tracking in your browser, plus area of interest analysis software.

(https://github.com/ajk77/EyeBrowserPy)

PatientPy: Patient state construction from clinical databases for machine learning.

(https://github.com/ajk77/PatientPy)

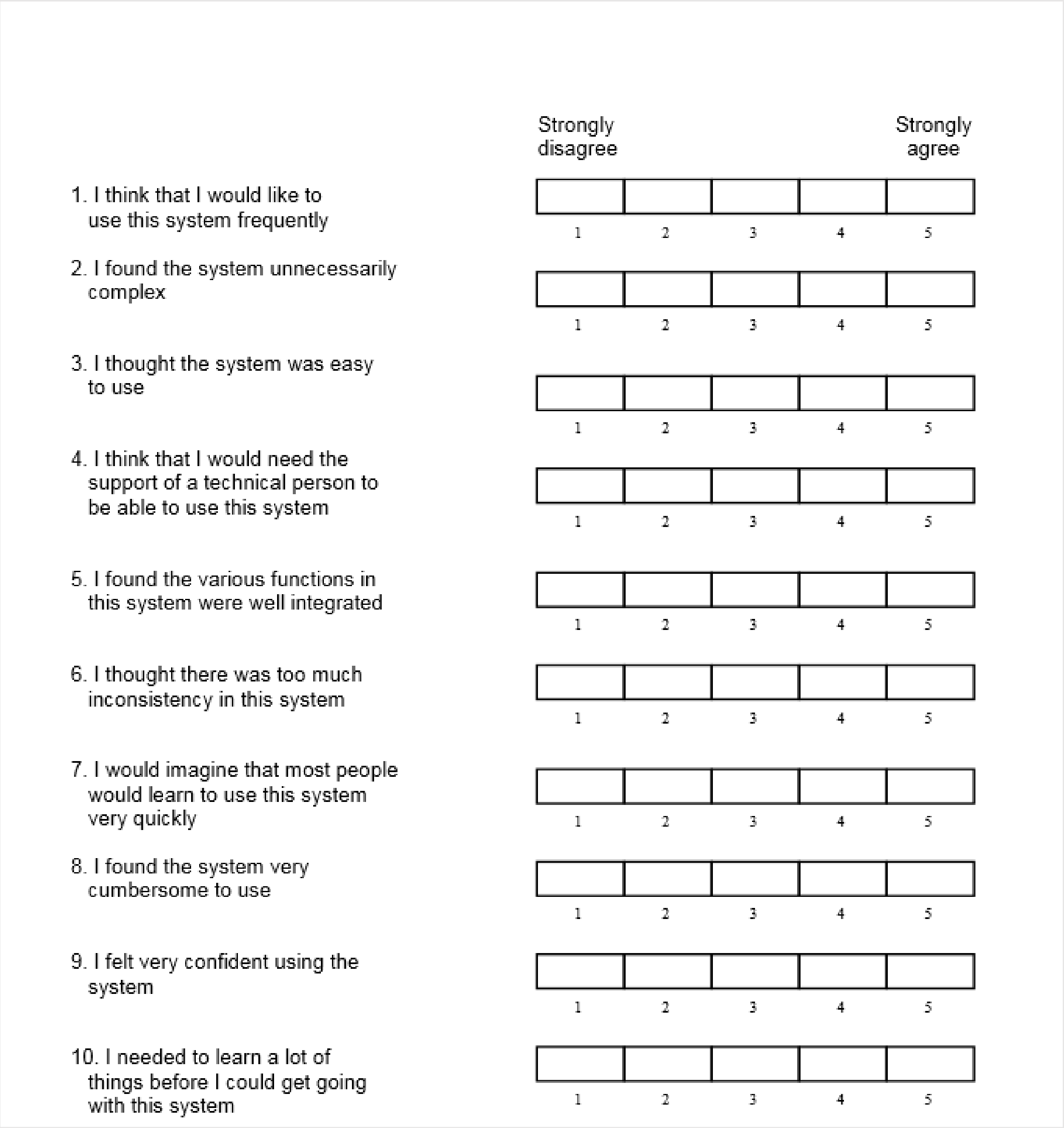
RegressiveImputer: Impute missing values via a regression model.

(https://github.com/ajk77/RegressiveImputer)

PateintPyFeatureSelection: Feature selection for constructed sets of features, such as the temporal expansions used in PatientPy.

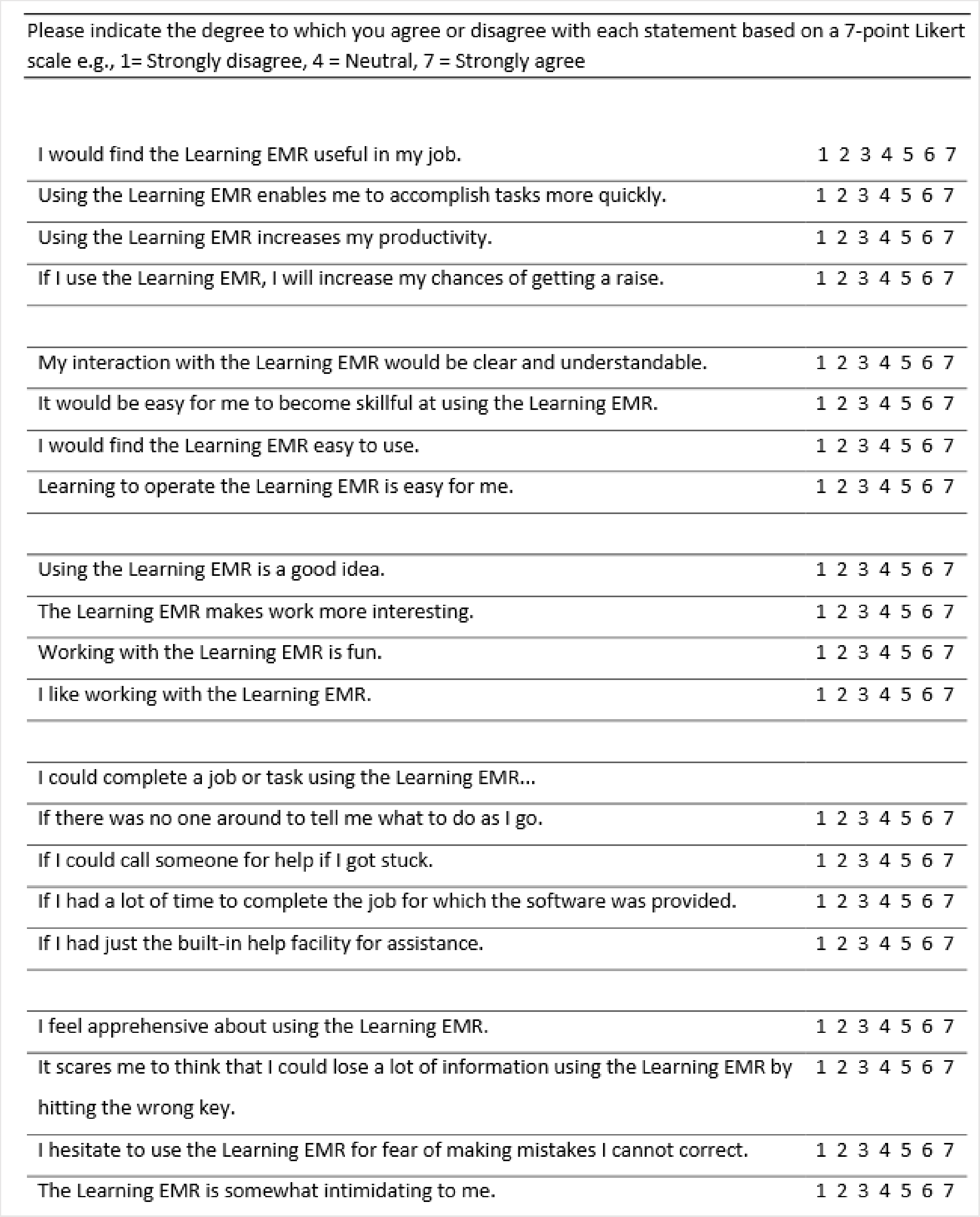
(https://github.com/ajk77/PatientPyFeatureSelection)

APPENDIX C. SURVEYS AND QUESTIONNAIRES



	Strongly disagree				Strongly agree
1. I think that I would like to use this system frequently	□ 1	□ 2	□ 3	□ 4	□ 5
2. I found the system unnecessarily complex	□ 1	□ 2	□ 3	□ 4	□ 5
3. I thought the system was easy to use	□ 1	□ 2	□ 3	□ 4	□ 5
4. I think that I would need the support of a technical person to be able to use this system	□ 1	□ 2	□ 3	□ 4	□ 5
5. I found the various functions in this system were well integrated	□ 1	□ 2	□ 3	□ 4	□ 5
6. I thought there was too much inconsistency in this system	□ 1	□ 2	□ 3	□ 4	□ 5
7. I would imagine that most people would learn to use this system very quickly	□ 1	□ 2	□ 3	□ 4	□ 5
8. I found the system very cumbersome to use	□ 1	□ 2	□ 3	□ 4	□ 5
9. I felt very confident using the system	□ 1	□ 2	□ 3	□ 4	□ 5
10. I needed to learn a lot of things before I could get going with this system	□ 1	□ 2	□ 3	□ 4	□ 5

Figure 25. System usability scale (SUS) used in the usability study (Section 3.5).

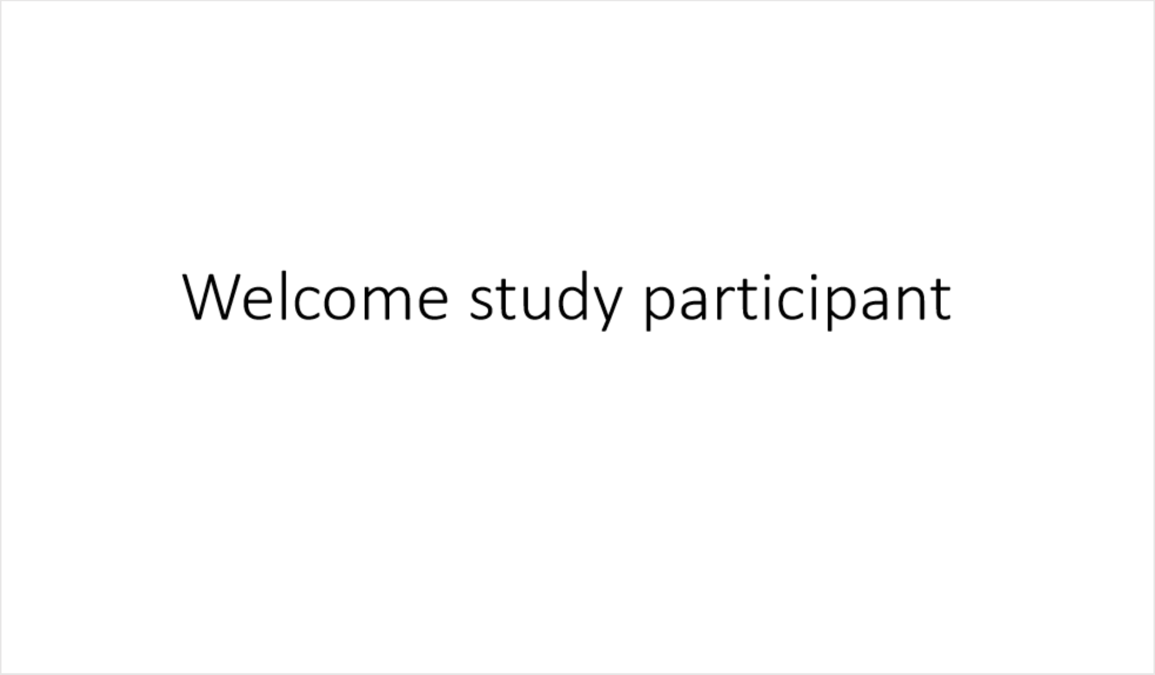


Please indicate the degree to which you agree or disagree with each statement based on a 7-point Likert scale e.g., 1= Strongly disagree, 4 = Neutral, 7 = Strongly agree

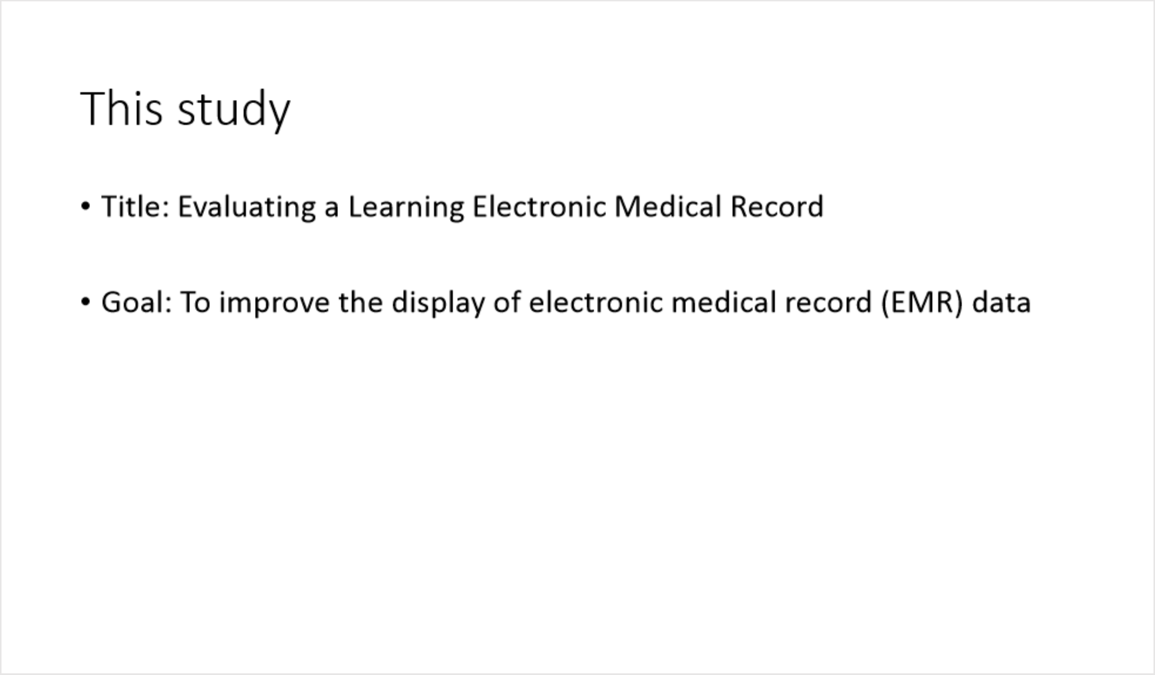
I would find the Learning EMR useful in my job.	1 2 3 4 5 6 7
Using the Learning EMR enables me to accomplish tasks more quickly.	1 2 3 4 5 6 7
Using the Learning EMR increases my productivity.	1 2 3 4 5 6 7
If I use the Learning EMR, I will increase my chances of getting a raise.	1 2 3 4 5 6 7
My interaction with the Learning EMR would be clear and understandable.	1 2 3 4 5 6 7
It would be easy for me to become skillful at using the Learning EMR.	1 2 3 4 5 6 7
I would find the Learning EMR easy to use.	1 2 3 4 5 6 7
Learning to operate the Learning EMR is easy for me.	1 2 3 4 5 6 7
Using the Learning EMR is a good idea.	1 2 3 4 5 6 7
The Learning EMR makes work more interesting.	1 2 3 4 5 6 7
Working with the Learning EMR is fun.	1 2 3 4 5 6 7
I like working with the Learning EMR.	1 2 3 4 5 6 7
I could complete a job or task using the Learning EMR…	
If there was no one around to tell me what to do as I go.	1 2 3 4 5 6 7
If I could call someone for help if I got stuck.	1 2 3 4 5 6 7
If I had a lot of time to complete the job for which the software was provided.	1 2 3 4 5 6 7
If I had just the built-in help facility for assistance.	1 2 3 4 5 6 7
I feel apprehensive about using the Learning EMR.	1 2 3 4 5 6 7
It scares me to think that I could lose a lot of information using the Learning EMR by hitting the wrong key.	1 2 3 4 5 6 7
I hesitate to use the Learning EMR for fear of making mistakes I cannot correct.	1 2 3 4 5 6 7
The Learning EMR is somewhat intimidating to me.	1 2 3 4 5 6 7

Figure 26. Modified unified theory of acceptance and use of technology (UTAUT) model used in the evaluation study (Section 7.1).

APPENDIX D. INTRODUCTORY PRESENTATION FOR THE PRIMARY LEMR EVALUATION

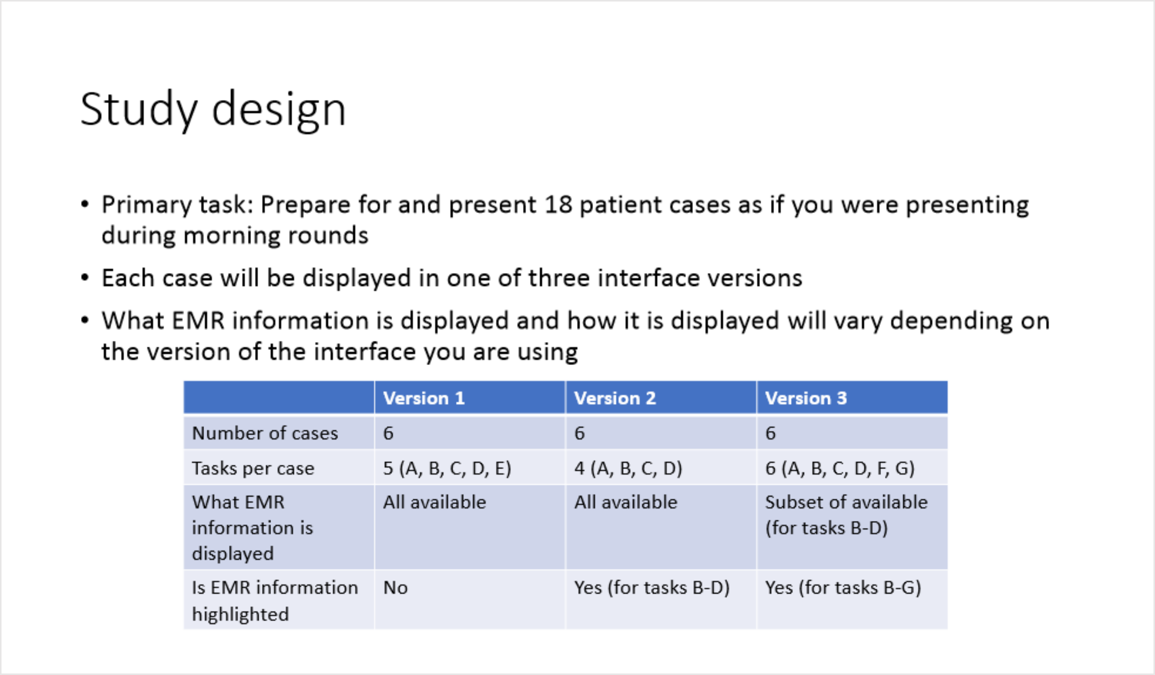


Welcome study participant



This study

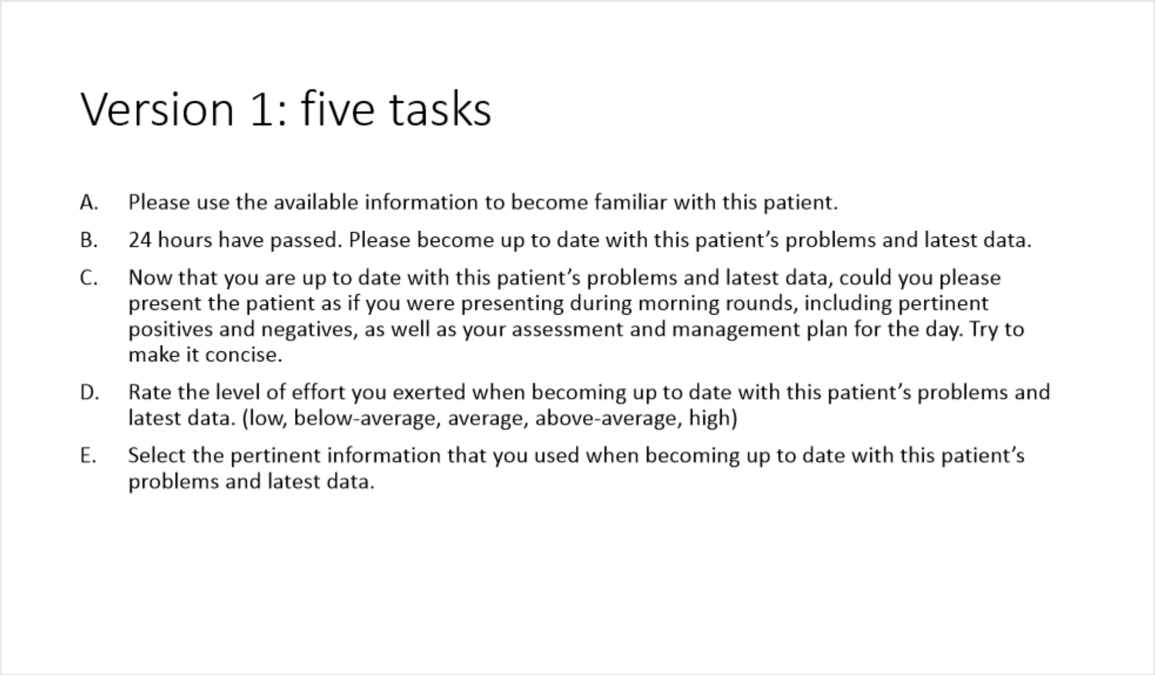
- Title: Evaluating a Learning Electronic Medical Record
- Goal: To improve the display of electronic medical record (EMR) data



Study design

- Primary task: Prepare for and present 18 patient cases as if you were presenting during morning rounds
- Each case will be displayed in one of three interface versions
- What EMR information is displayed and how it is displayed will vary depending on the version of the interface you are using

	Version 1	Version 2	Version 3
Number of cases	6	6	6
Tasks per case	5 (A, B, C, D, E)	4 (A, B, C, D)	6 (A, B, C, D, F, G)
What EMR information is displayed	All available	All available	Subset of available (for tasks B-D)
Is EMR information highlighted	No	Yes (for tasks B-D)	Yes (for tasks B-G)



Version 1: five tasks

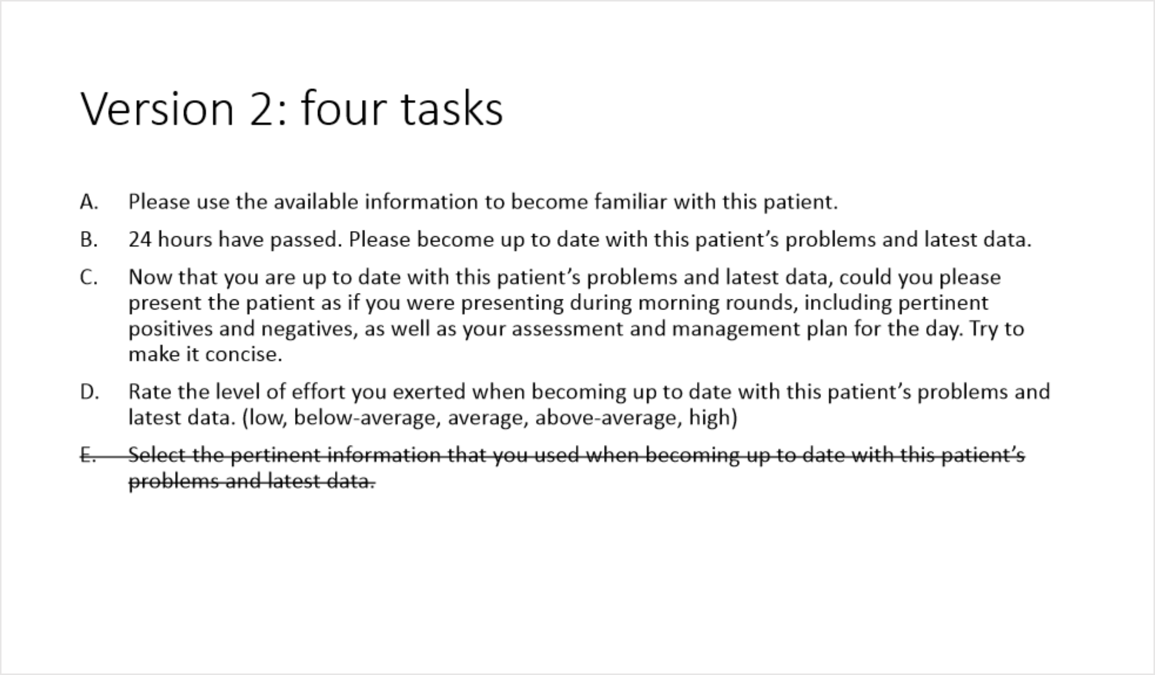
A. Please use the available information to become familiar with this patient.

B. 24 hours have passed. Please become up to date with this patient's problems and latest data.

C. Now that you are up to date with this patient's problems and latest data, could you please present the patient as if you were presenting during morning rounds, including pertinent positives and negatives, as well as your assessment and management plan for the day. Try to make it concise.

D. Rate the level of effort you exerted when becoming up to date with this patient's problems and latest data. (low, below-average, average, above-average, high)

E. Select the pertinent information that you used when becoming up to date with this patient's problems and latest data.



Version 2: four tasks

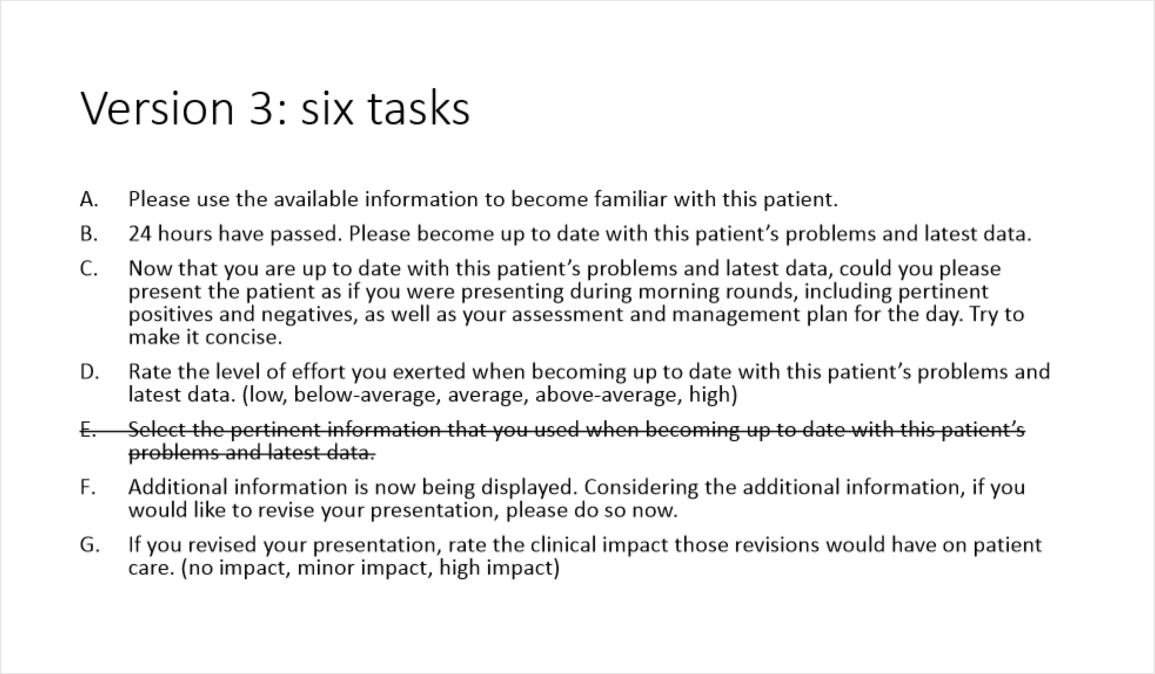
A. Please use the available information to become familiar with this patient.

B. 24 hours have passed. Please become up to date with this patient’s problems and latest data.

C. Now that you are up to date with this patient’s problems and latest data, could you please present the patient as if you were presenting during morning rounds, including pertinent positives and negatives, as well as your assessment and management plan for the day. Try to make it concise.

D. Rate the level of effort you exerted when becoming up to date with this patient’s problems and latest data. (low, below-average, average, above-average, high)

~~E. Select the pertinent information that you used when becoming up to date with this patient’s problems and latest data.~~



Version 3: six tasks

A. Please use the available information to become familiar with this patient.

B. 24 hours have passed. Please become up to date with this patient’s problems and latest data.

C. Now that you are up to date with this patient’s problems and latest data, could you please present the patient as if you were presenting during morning rounds, including pertinent positives and negatives, as well as your assessment and management plan for the day. Try to make it concise.

D. Rate the level of effort you exerted when becoming up to date with this patient’s problems and latest data. (low, below-average, average, above-average, high)

~~E. Select the pertinent information that you used when becoming up to date with this patient’s problems and latest data.~~

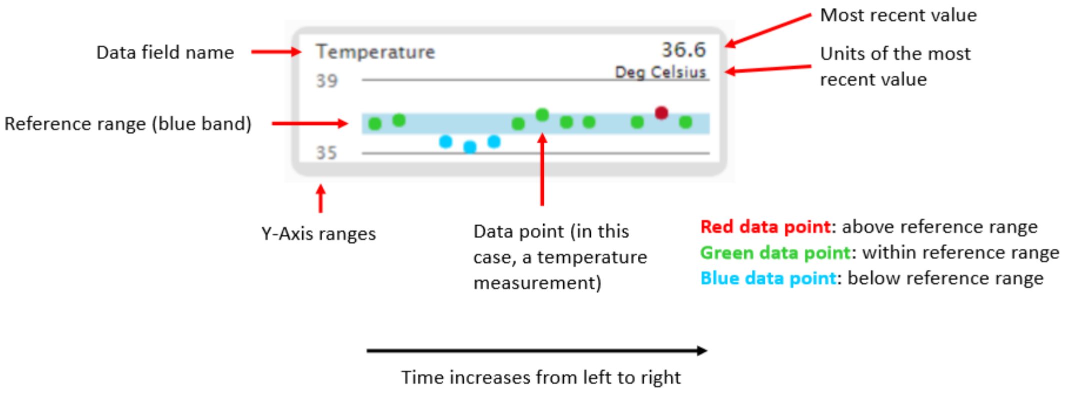
F. Additional information is now being displayed. Considering the additional information, if you would like to revise your presentation, please do so now.

G. If you revised your presentation, rate the clinical impact those revisions would have on patient care. (no impact, minor impact, high impact)

Interface design

Some information is displayed in a temporal manner:

- vital sign measurements
- medication administration
- laboratory test results
- ventilator settings
- intake and output measurements



Data field name

Reference range (blue band)

Y-Axis ranges

Data point (in this case, a temperature measurement)

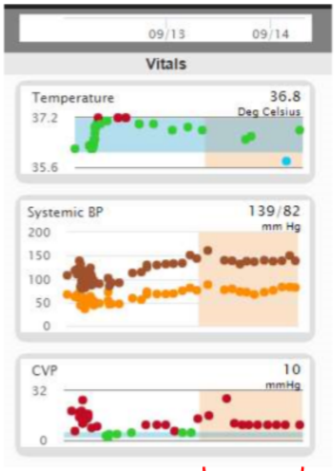
Most recent value

Units of the most recent value

Red data point: above reference range
Green data point: within reference range
Blue data point: below reference range

Time increases from left to right

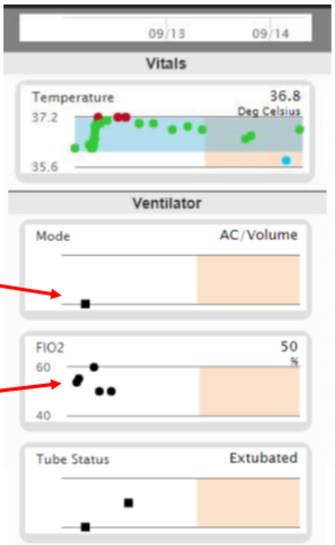
Date axis



For blood pressure charts:

- systolic measurements are **brown**
- diastolic measurements are **orange**

The tan band covers the last 24 hours of patient data



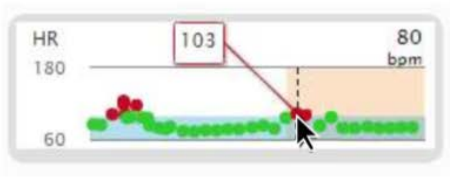
Discrete data points are represented with black squares

Data points on a chart without a blue reference range are black circles

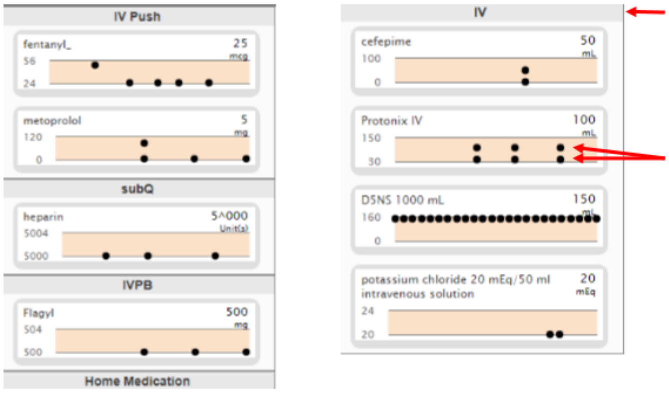
Discrete data fields do not have a defined y-axis. Hover over a data point with curser to see value (shown on next slide)

Discrete data points with different y-axis height will have different values. Points with the same y-axis height will have the same value

Hover over data point with curser to see value



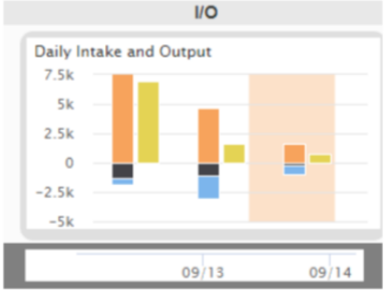
Medication charts will follow the same rules as other charts. Data points provided correspond to medication administration. The data field names and data points displayed are based on the information we have access to in our database of de-identified patient cases. Make sense of it the best you can.

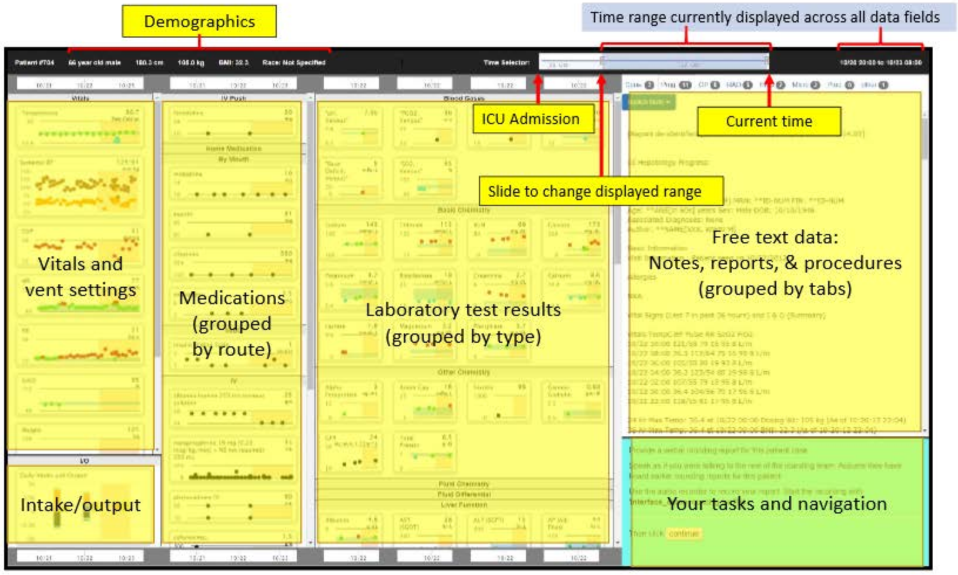


Meds are grouped by route

Some charts will display both volume and mass. Hover over the points to see the difference.

Intake and output uses a bar chart where there are two bars for each day. A multi-color bar that shows the differed types of measured intake and output and a yellow bar that shows the daily net. Hover over a bar to see the name of that type of i/o and the value of the measurement.





Demographics

Time range currently displayed across all data fields

ICU Admission

Current time

Slide to change displayed range

Vitals and vent settings

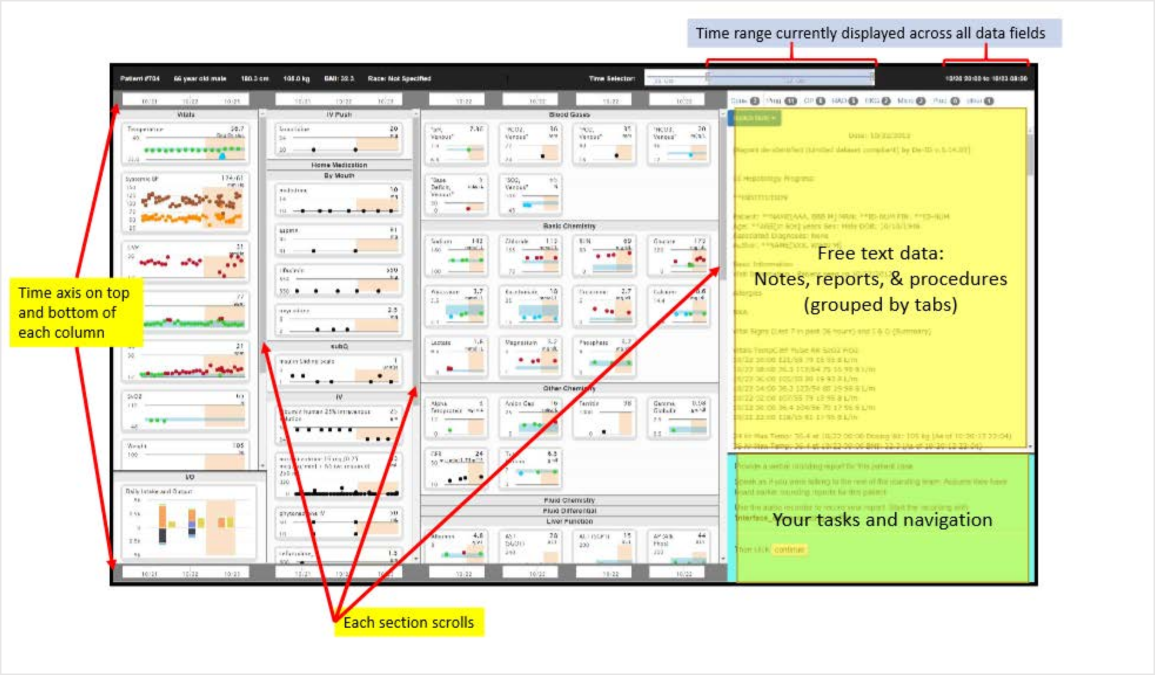
Medications (grouped by route)

Laboratory test results (grouped by type)

Free text data: Notes, reports, & procedures (grouped by tabs)

Intake/output

Your tasks and navigation



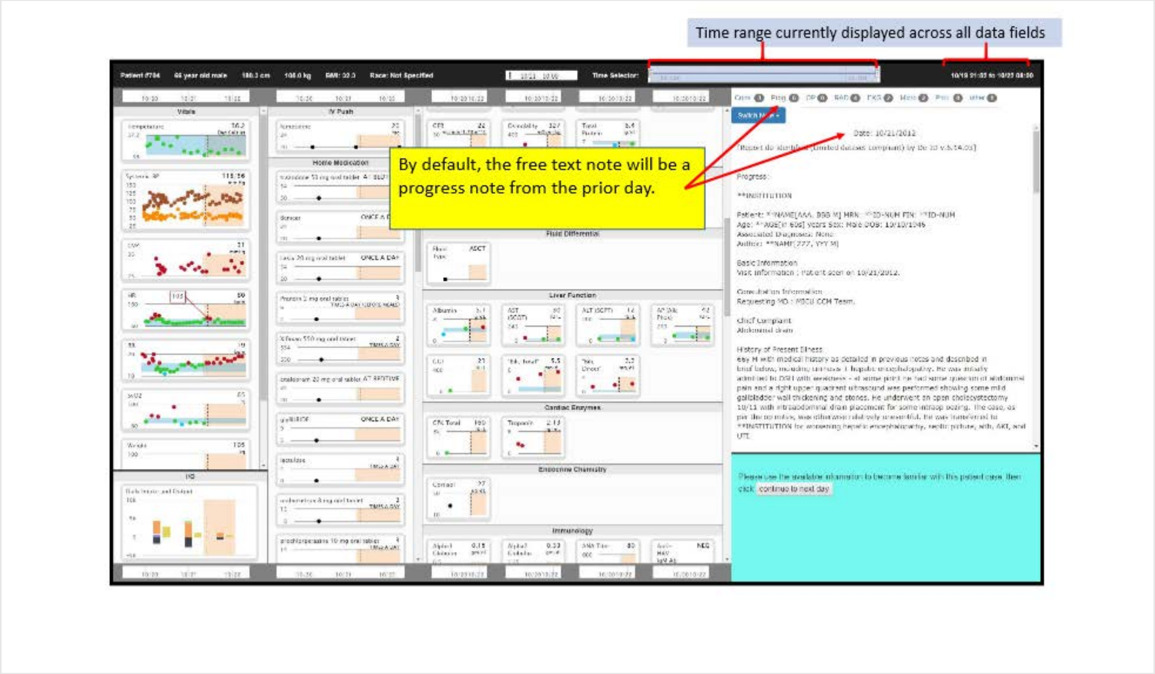
Time range currently displayed across all data fields

Time axis on top and bottom of each column

Free text data: Notes, reports, & procedures (grouped by tabs)

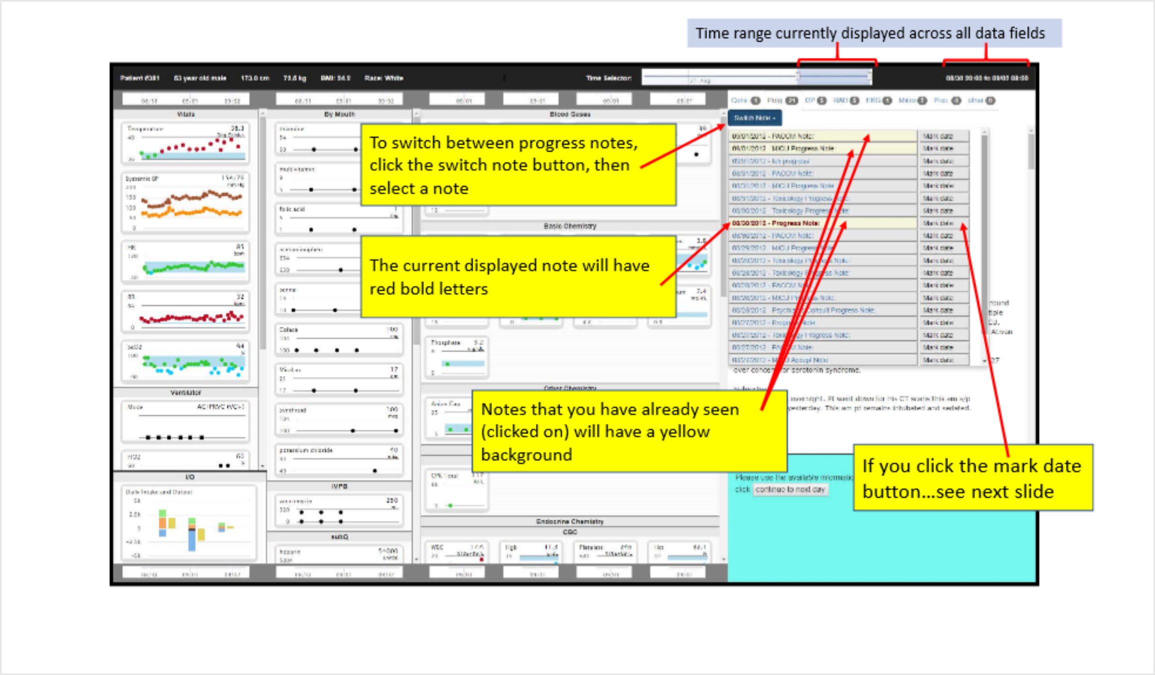
Your tasks and navigation

Each section scrolls



Time range currently displayed across all data fields

By default, the free text note will be a progress note from the prior day.



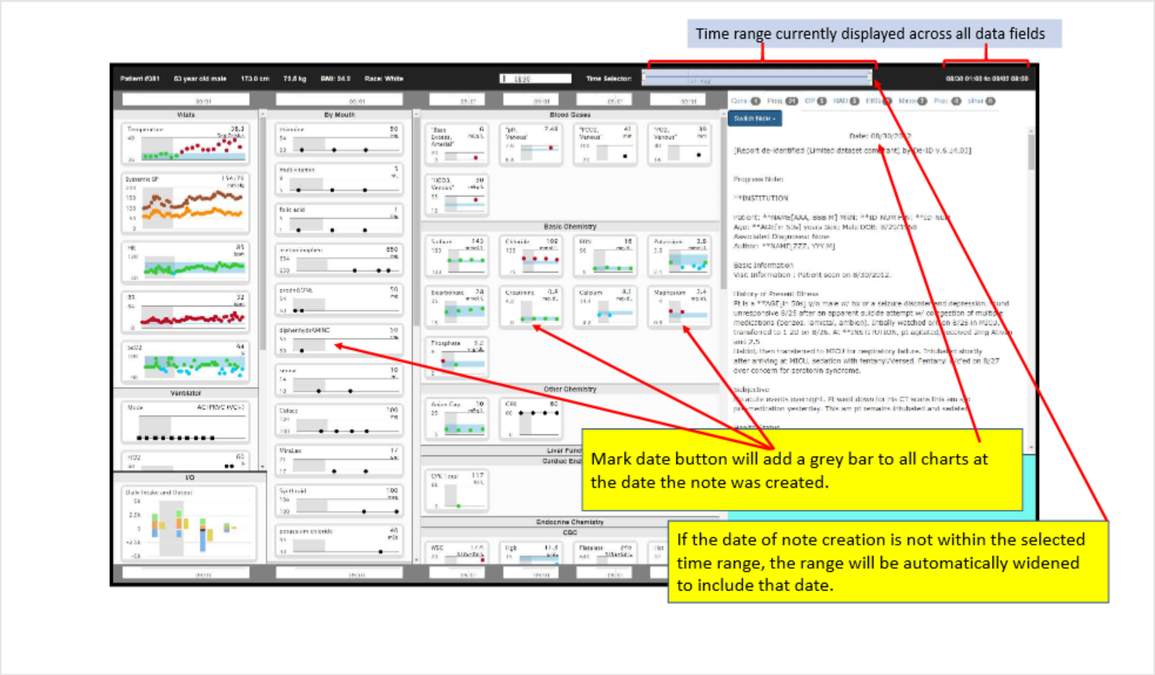
Time range currently displayed across all data fields

To switch between progress notes, click the switch note button, then select a note

The current displayed note will have red bold letters

Notes that you have already seen (clicked on) will have a yellow background

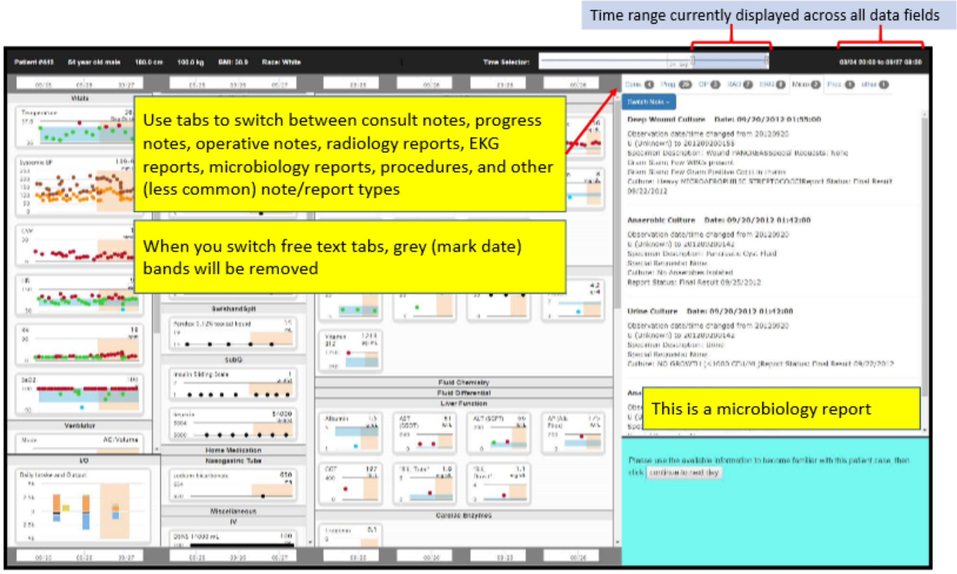
If you click the mark date button…see next slide



Time range currently displayed across all data fields

Mark date button will add a grey bar to all charts at the date the note was created.

If the date of note creation is not within the selected time range, the range will be automatically widened to include that date.



Time range currently displayed across all data fields

Use tabs to switch between consult notes, progress notes, operative notes, radiology reports, EKG reports, microbiology reports, procedures, and other (less common) note/report types

When you switch free text tabs, grey (mark date) bands will be removed

Deep Wound Culture Date: 09/20/2012 01:55:00

Observation date/time changed from 20120920
U (Unknown) to 201209200155
Specimen Description: Wound
Gram Stain: Few WBCs present
Gram Stain: Few Gram Positive Cocci in chains
Culture: Heavy MICROAEROPHILIC STREPTOCOCCIReport Status: Final Result 09/22/2012

Anaerobic Culture Date: 09/20/2012 01:42:00

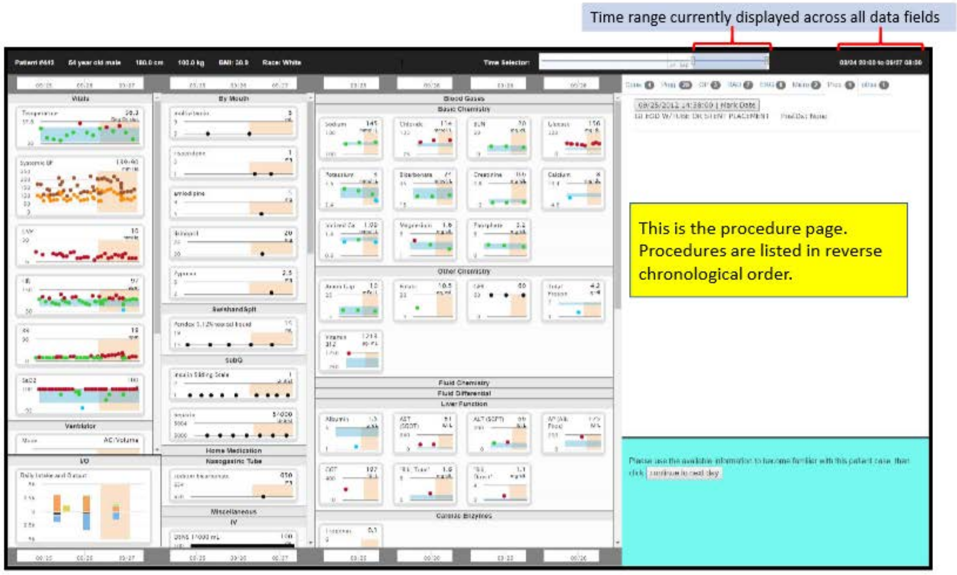
Observation date/time changed from 20120920
U (Unknown) to 201209200142
Specimen Description: Pancreatic Cyst Fluid
Special Requests: None
Culture: No Anaerobes Isolated
Report Status: Final Result 09/25/2012

Urine Culture Date: 09/20/2012 01:42:00

Observation date/time changed from 20120920
U (Unknown) to 201209200142
Specimen Description: Urine
Special Requests: None
Culture: NO GROWTH (<1000 CFU/ML)Report Status: Final Result 09/22/2012

This is a microbiology report

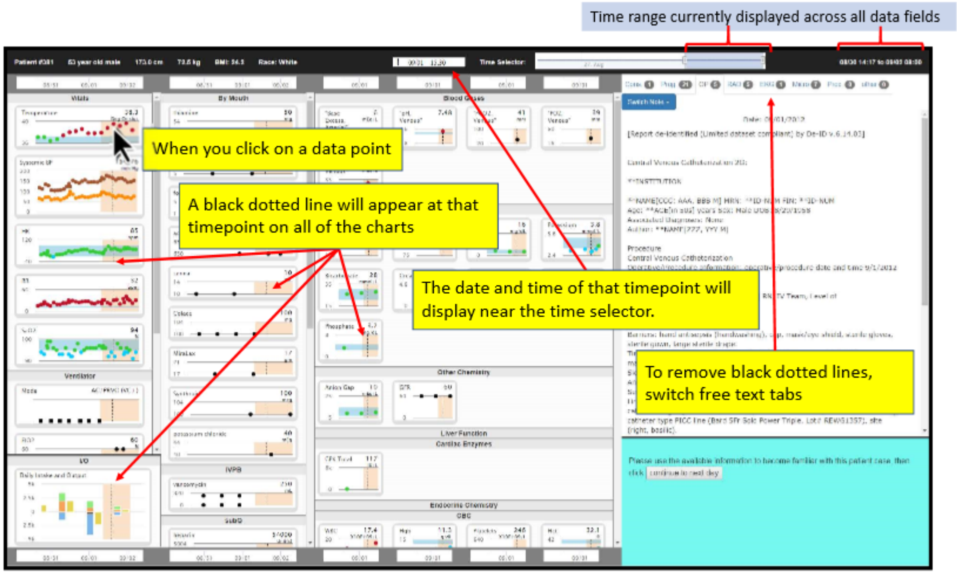
Please use the available information to become familiar with this patient case, then click continue to next day



Time range currently displayed across all data fields

This is the procedure page. Procedures are listed in reverse chronological order.

Please use the available information to become familiar with this patient case, then click continue to next day



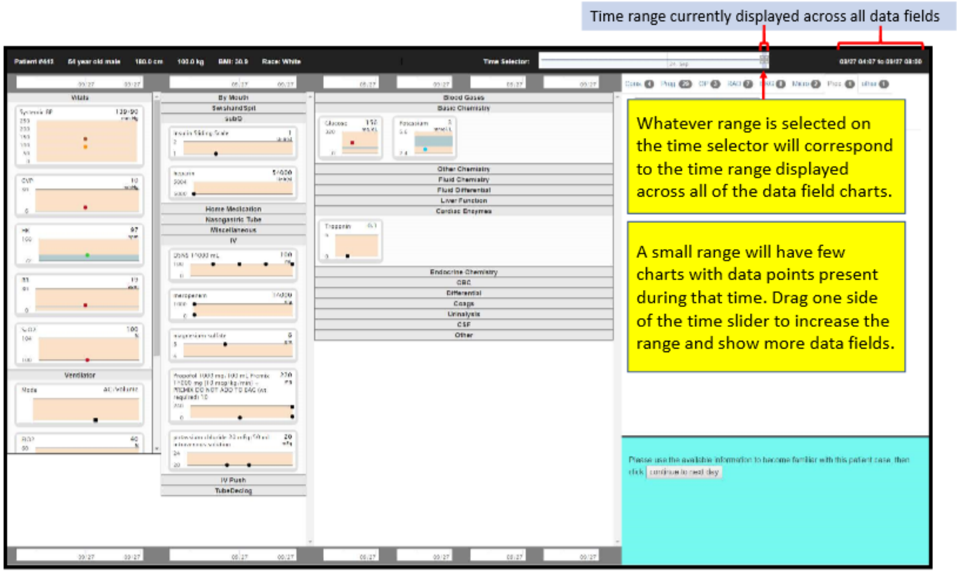
Time range currently displayed across all data fields

When you click on a data point

A black dotted line will appear at that timepoint on all of the charts

The date and time of that timepoint will display near the time selector.

To remove black dotted lines, switch free text tabs



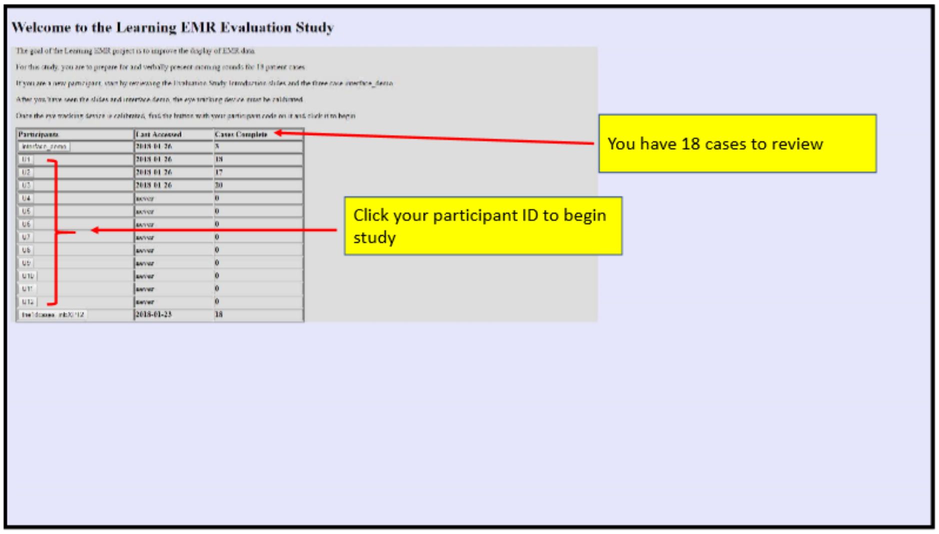
Time range currently displayed across all data fields

Whatever range is selected on the time selector will correspond to the time range displayed across all of the data field charts.

A small range will have few charts with data points present during that time. Drag one side of the time slider to increase the range and show more data fields.

Interface demo

Study home screen



Welcome to the Learning EMR Evaluation Study

The goal of the Learning EMR project is to improve the display of EMR data.

For this study, you are to prepare for and verbally present morning rounds for 18 patient cases.

If you are a new participant, start by reviewing the Evaluation Study Introduction slides and the three case interface_demo.

After you have seen the slides and interface demo, the eye tracking device must be calibrated.

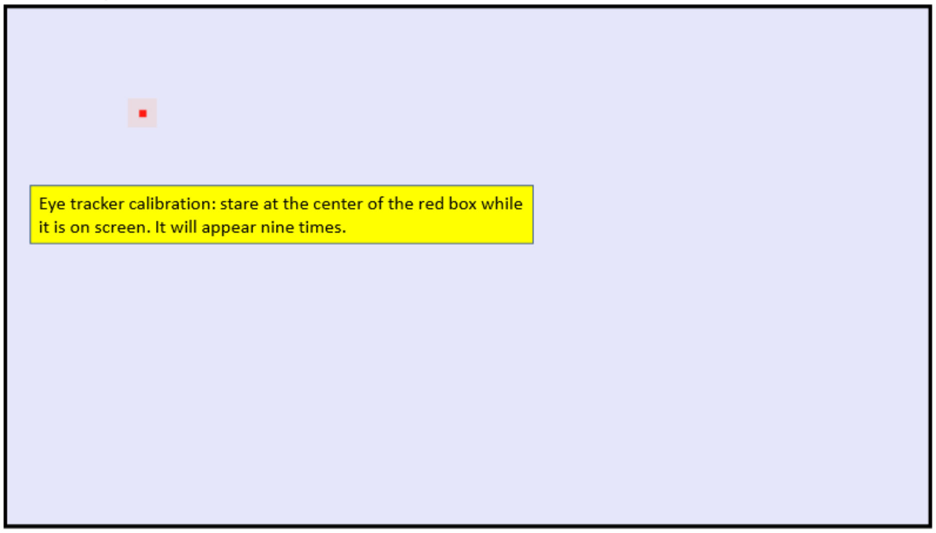
Once the eye tracking device is calibrated, find the button with your participant code on it and click it to begin.

Participants	Last Accessed	Cases Complete
interface_demo	2018 01 26	3
U1	2018 01 26	18
U2	2018 01 26	17
U3	2018 01 26	20
U4	never	0
U5	never	0
U6	never	0
U7	never	0
U8	never	0
U9	never	0
U10	never	0
U11	never	0
U12	never	0
[illegible]	2018-01-23	18

You have 18 cases to review

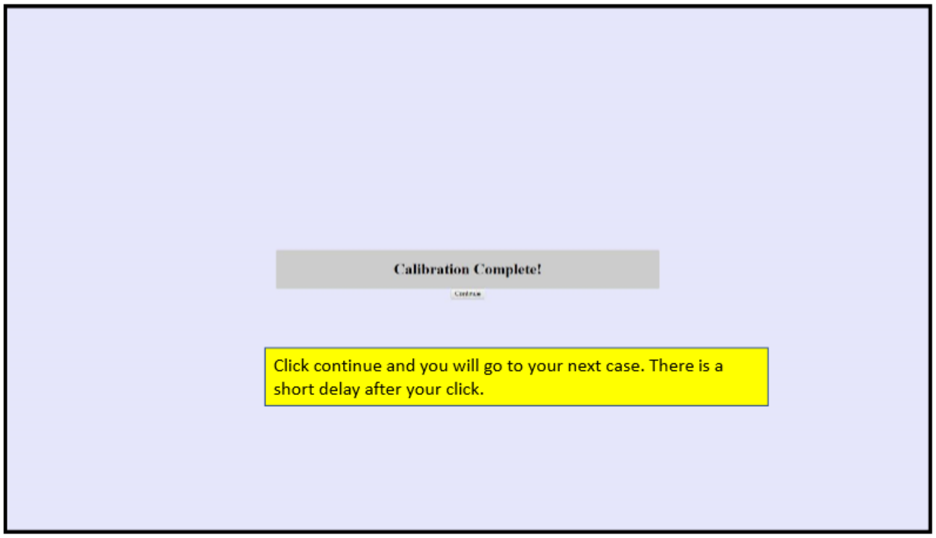
Click your participant ID to begin study

First you will see the eye tracker calibration screen



Eye tracker calibration: stare at the center of the red box while it is on screen. It will appear nine times.

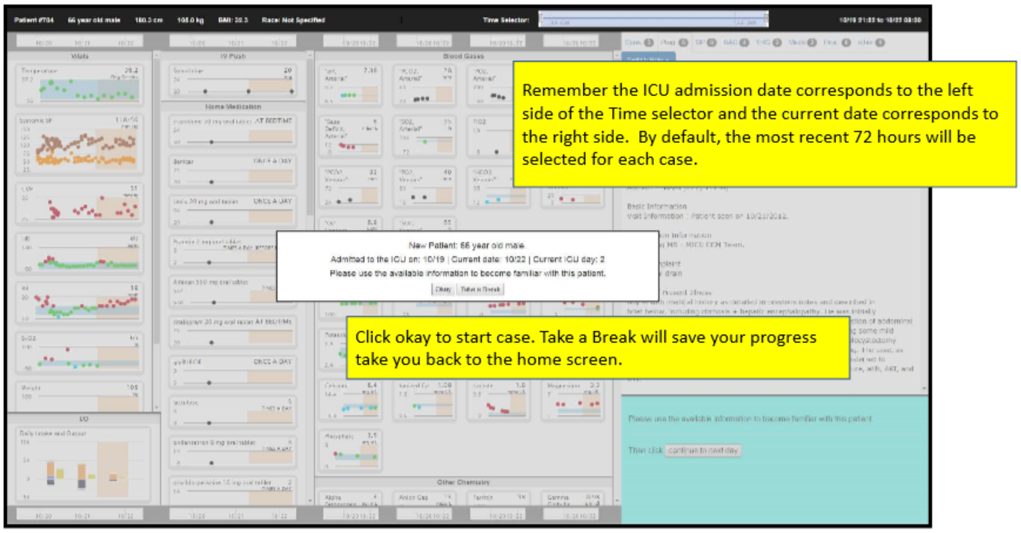
End of the eye tracker calibration screen



Calibration Complete!

Click continue and you will go to your next case. There is a short delay after your click.

Demo Case 1 shown in Version 1



Patient #784 66 year old male 180.3 cm 108.0 kg BMI: 33.3 Race: Not Specified Time Selector: 10/19 21:53 to 10/22 08:00

New Patient: 66 year old male

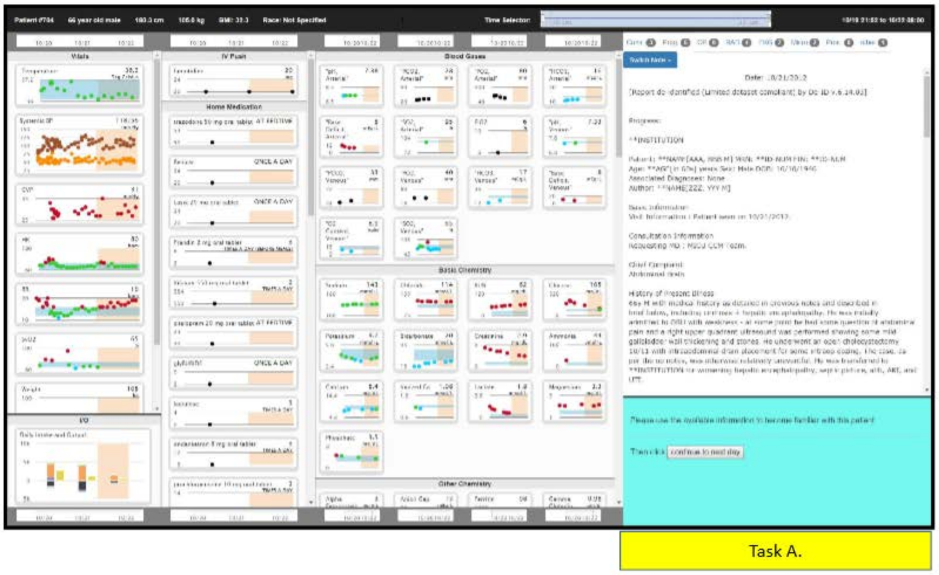
Admitted to the ICU on: 10/19 | Current date: 10/22 | Current ICU day: 2

Please use the available information to become familiar with this patient.

Remember the ICU admission date corresponds to the left side of the Time selector and the current date corresponds to the right side. By default, the most recent 72 hours will be selected for each case.

Click okay to start case. Take a Break will save your progress take you back to the home screen.

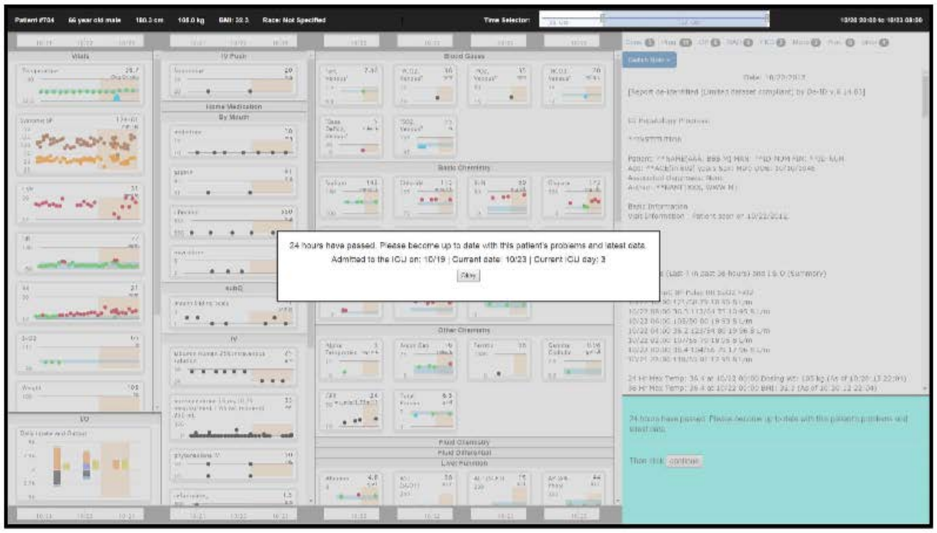
Demo Case 1 shown in Version 1



Patient #784 66 year old male 180.3 cm 108.0 kg BMI: 33.3 Race: Not Specified Time Selector: 10/19 21:53 to 10/22 08:00

Task A.

Demo Case 1 shown in Version 1



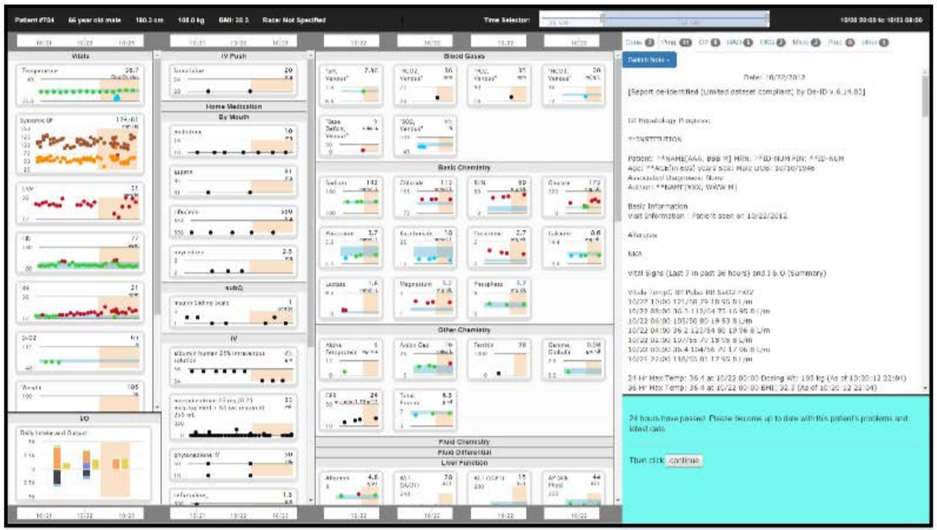
24 hours have passed. Please become up to date with this patient's problems and latest data.

Admitted to the ICU on: 10/19 | Current date: 10/23 | Current ICU day: 3

24 hours have passed. Please become up to date with this patient's problems and latest data.

Then click continue

Demo Case 1 shown in Version 1

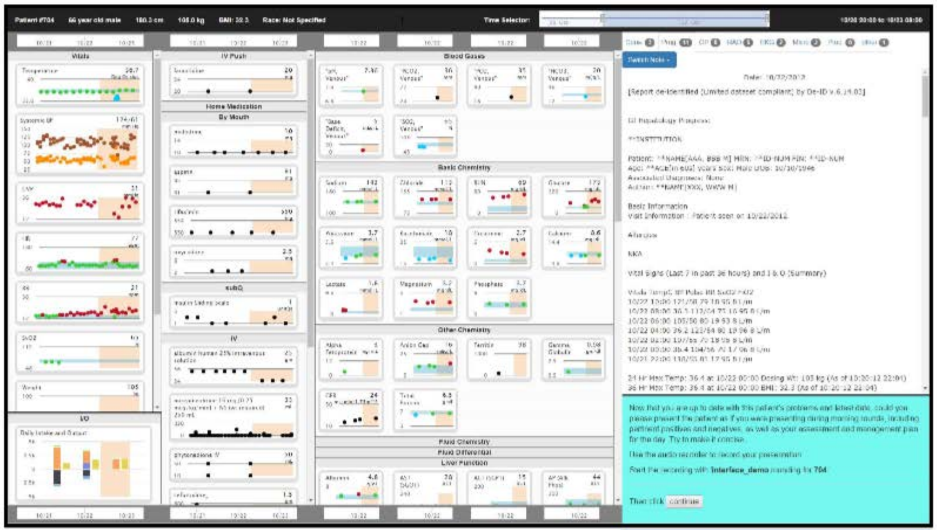


24 hours have passed. Please become up to date with this patient's problems and latest data.

Then click continue

Task B.

Demo Case 1 shown in Version 1



Date: 10/22/2012

[Report de-identified (Limited dataset compliant) by De-ID v.6.14.02]

GI Hepatology Progress

**INSTITUTION

Patient: **NAME[AAA, BBB M] MRN: **ID-NUM FIN: **ID-NUM
Age: **AGE[in 60s] years Sex: Male DOB: 10/10/1946
Associated Diagnoses: None
Author: **NAME[XXX, WWW M]

Basic Information
Visit Information : Patient seen on 10/22/2012.

Allergies

NKA

Vital Signs (Last 7 in past 36 hours) and I & O (Summary)

Vitals TempC BP Pulse RR SaO2 FiO2
10/22 10:00 121/58 79 18 95 8 L/m
10/22 08:00 36.1 112/54 73 16 95 8 L/m
10/22 06:00 105/50 80 19 93 8 L/m
10/22 04:00 36.2 123/54 80 19 96 8 L/m
10/22 02:00 107/55 70 18 95 8 L/m
10/22 00:00 36.4 104/56 70 17 96 8 L/m
10/21 22:00 110/55 81 17 95 8 L/m

24 Hr Max Temp: 36.4 at 10/22 00:00 Dosing Wt: 103 kg (As of 10:20:12 22:04)
36 Hr Max Temp: 36.4 at 10/22 00:00 BMI: 32.3 (As of 10:20:12 22:04)

Now that you are up to date with this patient's problems and latest data, could you please present the patient as if you were presenting during morning rounds, including pertinent positives and negatives, as well as your assessment and management plan for the day. Try to make it concise.

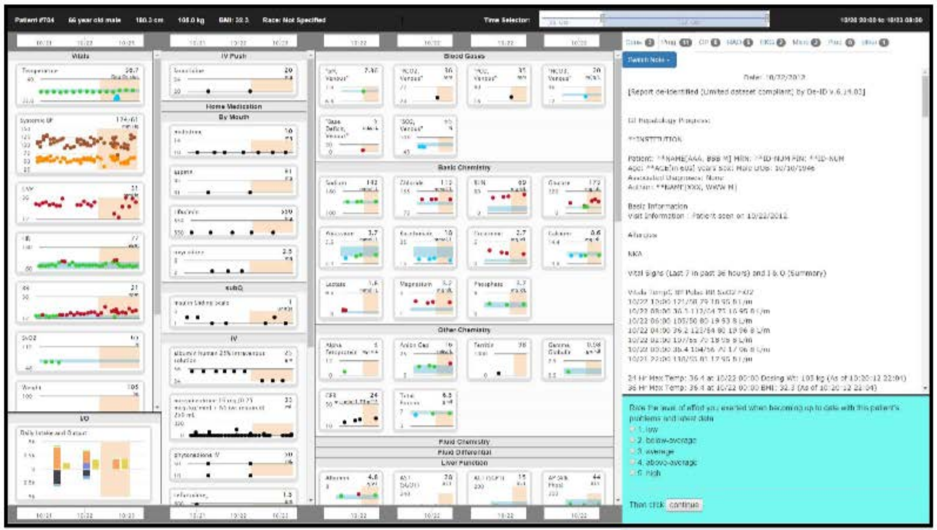
Use the audio recorder to record your presentation.

Start the recording with **Interface_demo** standing for **704**.

Then click continue

Task C.

Demo Case 1 shown in Version 1



Date: 10/22/2012

[Report de-identified (Limited dataset compliant) by De-ID v.6.14.02]

GI Hepatology Progress

**INSTITUTION

Patient: **NAME[AAA, BBB M] MRN: **ID-NUM FIN: **ID-NUM
Age: **AGE[in 60s] years Sex: Male DOB: 10/10/1946
Associated Diagnoses: None
Author: **NAME[XXX, WWW M]

Basic Information
Visit Information : Patient seen on 10/22/2012.

Allergies

NKA

Vital Signs (Last 7 in past 36 hours) and I & O (Summary)

Vitals TempC BP Pulse RR SaO2 FiO2
10/22 10:00 121/58 79 18 95 8 L/m
10/22 08:00 36.1 112/54 73 16 95 8 L/m
10/22 06:00 105/50 80 19 93 8 L/m
10/22 04:00 36.2 123/54 80 19 96 8 L/m
10/22 02:00 107/55 70 18 95 8 L/m
10/22 00:00 36.4 104/56 70 17 96 8 L/m
10/21 22:00 110/55 81 17 95 8 L/m

24 Hr Max Temp: 36.4 at 10/22 00:00 Dosing Wt: 103 kg (As of 10:20:12 22:04)
36 Hr Max Temp: 36.4 at 10/22 00:00 BMI: 32.3 (As of 10:20:12 22:04)

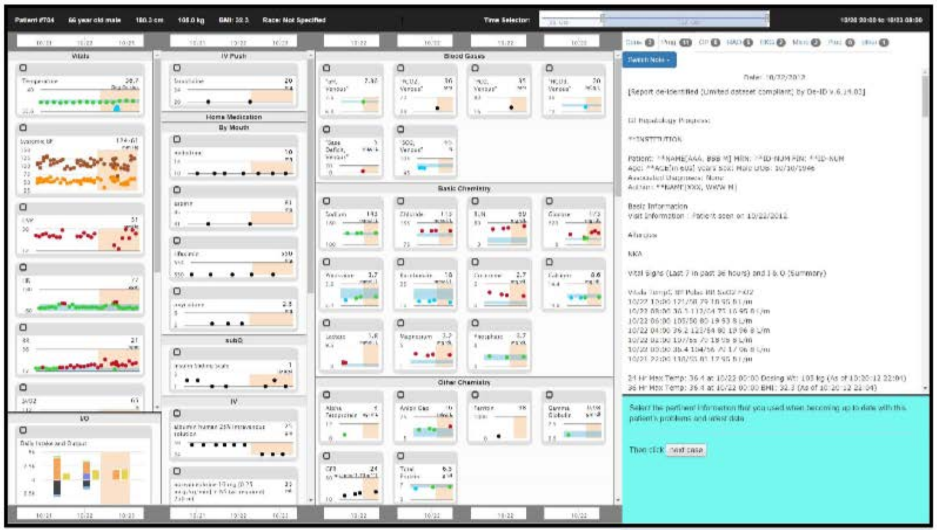
Rate the level of effort you exerted when becoming up to date with this patient's problems and latest data

○ 1. low
○ 2. below-average
○ 3. average
○ 4. above-average
○ 5. high

Then click continue

Task D.

Demo Case 1 shown in Version 1



Patient #734 66 year old male 180.3 cm 105.0 kg BMI: 32.3 Race: Not Specified

Time Selector:

10/20 20:00 to 10/23 08:00

Date: 10/22/2012

[Report de-identified (Limited dataset compliant) by De-ID v.6.19.03]

GI Hepatology Progress

**INSTITUTION

Patient: **NAME[AAA, BBB M] MRN: **ID-NUM FIN: **ID-NUM
Age: **AGE[in 60s] years Sex: Male DOB: 10/10/1946
Associated Diagnoses: None
Author: **NAME[XXX, WWW M]

Basic Information
Visit Information : Patient seen on 10/22/2012.

Allergies

NKA

Vital Signs (Last 7 in past 36 hours) and I & O (Summary)

Vitals Temp C BP Pulse RR SaO2 FiO2
10/22 10:00 121/58 79 18 95 8 L/m
10/22 08:00 36.1 112/64 73 16 95 8 L/m
10/22 06:00 105/50 80 19 93 8 L/m
10/22 04:00 36.2 122/54 80 19 96 8 L/m
10/22 02:00 107/55 70 18 95 8 L/m
10/22 00:00 36.4 104/56 72 17 96 8 L/m
10/21 22:00 110/55 81 17 95 8 L/m

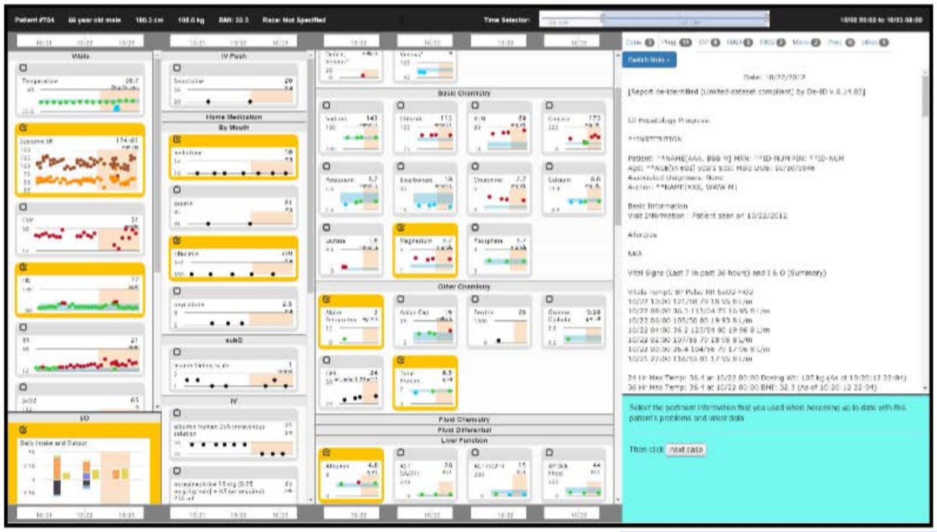
24 Hr Max Temp: 36.4 at 10/22 00:00 Dosing Wt: 105 kg (As of 10:20:12 22:04)
36 Hr Max Temp: 36.4 at 10/22 00:00 BMI: 32.3 (As of 10:20:12 22:04)

Select the pertinent information that you used when becoming up to date with this patient's problems and latest data

Then click next case

Task E.

Demo Case 1 shown in Version 1



Patient #734 66 year old male 180.3 cm 105.0 kg BMI: 32.3 Race: Not Specified

Time Selector:

10/20 20:00 to 10/23 08:00

Date: 10/22/2012

[Report de-identified (Limited dataset compliant) by De-ID v.6.19.03]

GI Hepatology Progress

**INSTITUTION

Patient: **NAME[AAA, BBB M] MRN: **ID-NUM FIN: **ID-NUM
Age: **AGE[in 60s] years Sex: Male DOB: 10/10/1946
Associated Diagnoses: None
Author: **NAME[XXX, WWW M]

Basic Information
Visit Information : Patient seen on 10/22/2012.

Allergies

NKA

Vital Signs (Last 7 in past 36 hours) and I & O (Summary)

Vitals Temp C BP Pulse RR SaO2 FiO2
10/22 10:00 121/58 79 18 95 8 L/m
10/22 08:00 36.1 112/64 73 16 95 8 L/m
10/22 06:00 105/50 80 19 93 8 L/m
10/22 04:00 36.2 122/54 80 19 96 8 L/m
10/22 02:00 107/55 70 18 95 8 L/m
10/22 00:00 36.4 104/56 72 17 96 8 L/m
10/21 22:00 110/55 81 17 95 8 L/m

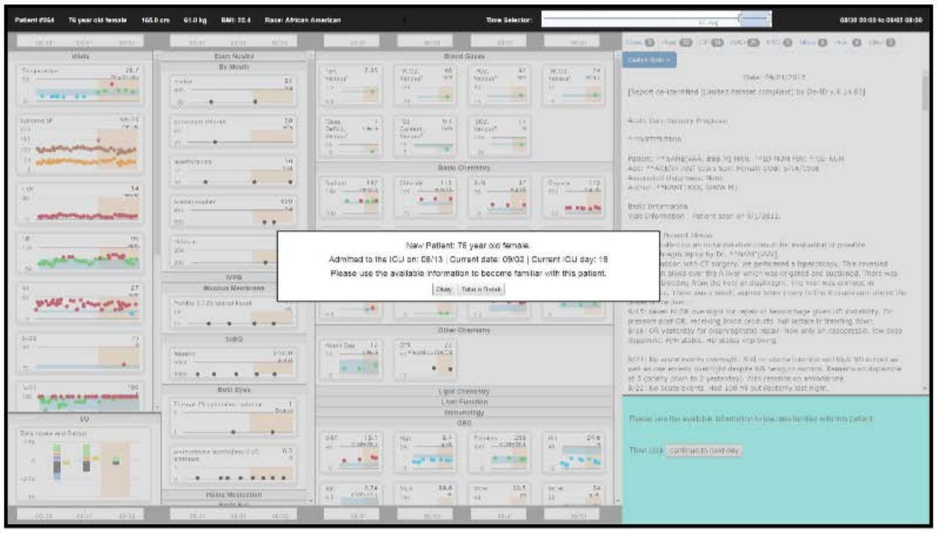
24 Hr Max Temp: 36.4 at 10/22 00:00 Dosing Wt: 105 kg (As of 10:20:12 22:04)
36 Hr Max Temp: 36.4 at 10/22 00:00 BMI: 32.3 (As of 10:20:12 22:04)

Select the pertinent information that you used when becoming up to date with this patient's problems and latest data

Then click next case

Task E. (with some selections being shown)

Demo Case 2 shown in Version 2



New Patient: 76 year old female

Admitted to the ICU on: 08/13 | Current date: 09/02 | Current ICU day: 18

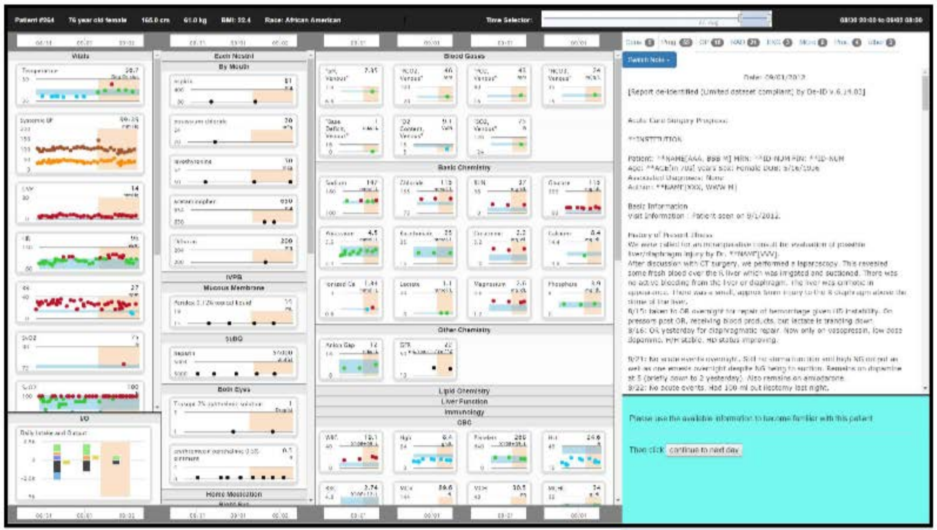
Please use the available information to become familiar with this patient.

Okay | Take a Break

Please use the available information to become familiar with this patient

Then click continue to next day

Demo Case 2 shown in Version 2



Date: 09/01/2012

[Report de-identified (Limited dataset compliant) by De-ID v.6.14.03]

Acute Care Surgery Progress

**INSTITUTION

Patient: **NAME[AAA, BBB M] MRN: **ID-NUM FIN: **ID-NUM
Age: **AGE[in 70s] years Sex: Female DOB: 5/16/1936
Associated Diagnoses: None
Author: **NAME[XXX, WWW M]

Basic Information
Visit Information : Patient seen on 9/1/2012.

History of Present Illness
We were called for an intraoperative consult for evaluation of possible liver/diaphragm injury by Dr. **NAME[VVV].
After discussion with CT surgery, we performed a laparoscopy. This revealed some fresh blood over the R liver which was irrigated and suctioned. There was no active bleeding from the liver or diaphragm. The liver was cirrhotic in appearance. There was a small, approx 5mm injury to the R diaphragm above the dome of the liver.
8/15: taken to OR overnight for repair of hemorrhage given HD instability. On pressors post OR, receiving blood products, but lactate is trending down.
8/16: OR yesterday for diaphragmatic repair. Now only on vasopressin, low dose dopamine. H/H stable, HD status improving.

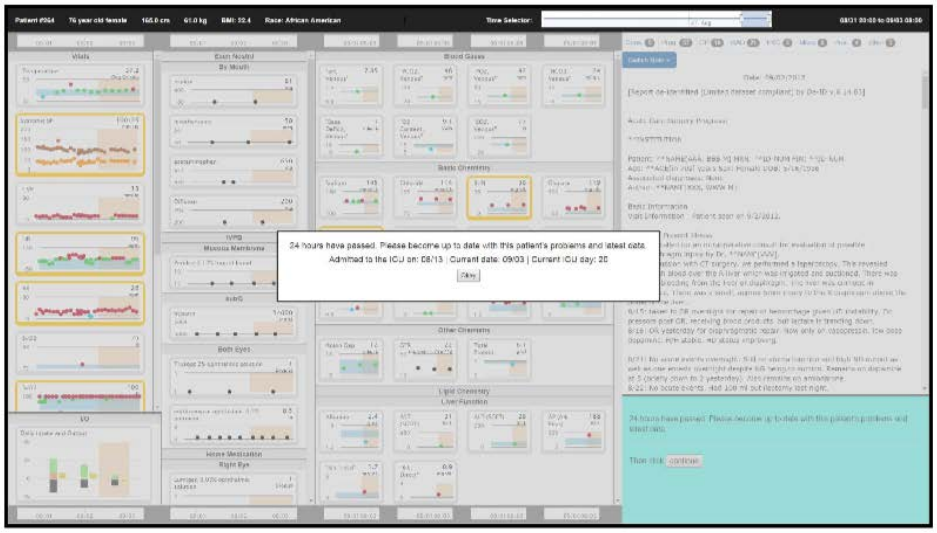
8/21: No acute events overnight. Still no stoma function and high NG output as well as one emesis overnight despite NG being to suction. Remains on dopamine at 5 (briefly down to 2 yesterday). Also remains on amiodarone.
8/22: No acute events. Had 100 ml out ileostomy last night.

Please use the available information to become familiar with this patient

Then click continue to next day

Task A.

Demo Case 2 shown in Version 2



Patient #264 76 year old female 165.0 cm 61.0 kg BMI: 22.4 Race: African American Time Selector: 08/31 00:00 to 09/03 08:00

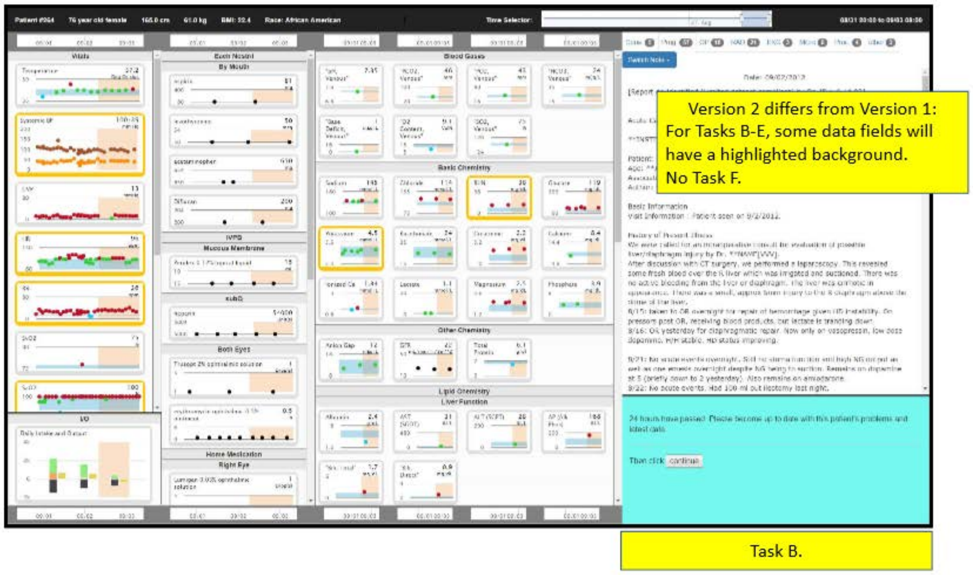
24 hours have passed. Please become up to date with this patient's problems and latest data.

Admitted to the ICU on: 08/13 | Current date: 09/03 | Current ICU day: 20

24 hours have passed. Please become up to date with this patient's problems and latest data.

Then click: continue

Demo Case 2 shown in Version 2



Patient #264 76 year old female 165.0 cm 61.0 kg BMI: 22.4 Race: African American Time Selector: 08/31 00:00 to 09/03 08:00

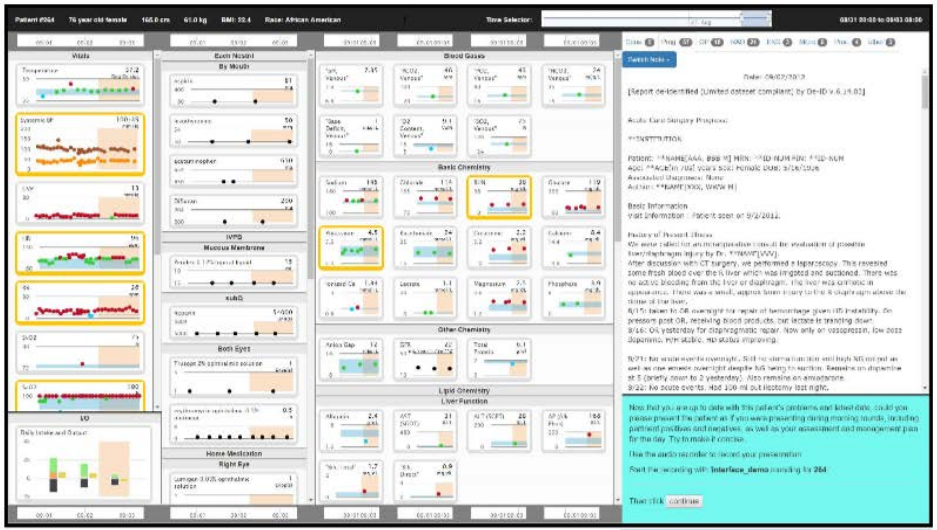
24 hours have passed. Please become up to date with this patient's problems and latest data.

Then click: continue

Version 2 differs from Version 1: For Tasks B-E, some data fields will have a highlighted background. No Task F.

Task B.

Demo Case 2 shown in Version 2



Patient #264 76 year old female 165.0 cm 61.0 kg BMI: 22.4 Race: African American Time Selector: 08/31 00:00 to 09/03 08:00

Date: 09/02/2012

[Report de-identified (Limited dataset compliant) by De-ID v.6.19.03]

Acute Care Surgery Progress

**INSTITUTION

Patient: **NAME[AAA, BBB M] MRN: **ID-NUM FIN: **ID-NUM
Age: **AGE[in 70s] years Sex: Female DOB: 5/16/19xx
Associated Diagnoses: None
Author: **NAME[XXX, WWW M]

Basic Information
Visit Information : Patient seen on 9/2/2012.

History of Present Illness
We were called for an intraoperative consult for evaluation of possible liver/diaphragm injury by Dr. **NAME[VVV].
After discussion with CT surgery, we performed a laparoscopy. This revealed some fresh blood over the R liver which was irrigated and suctioned. There was no active bleeding from the liver or diaphragm. The liver was cirrhotic in appearance. There was a small, approx 5mm injury to the R diaphragm above the dome of the liver.
8/15: taken to OR overnight for repair of hemorrhage given HD instability. On pressors post OR, receiving blood products, but lactate is trending down.
8/16: OR yesterday for diaphragmatic repair. Now only on vasopressin, low dose dopamine. H/H stable. HD status improving.

8/21: No acute events overnight. Still no stoma function and high NG output as well as one emesis overnight despite NG being to suction. Remains on dopamine at 5 (briefly down to 2 yesterday). Also remains on amiodarone.
8/22: No acute events. Had 100 ml out ileostomy last night.

Now that you are up to date with this patient's problems and latest data, could you please present the patient as if you were presenting during morning rounds, including pertinent positives and negatives, as well as your assessment and management plan for the day. Try to make it concise.

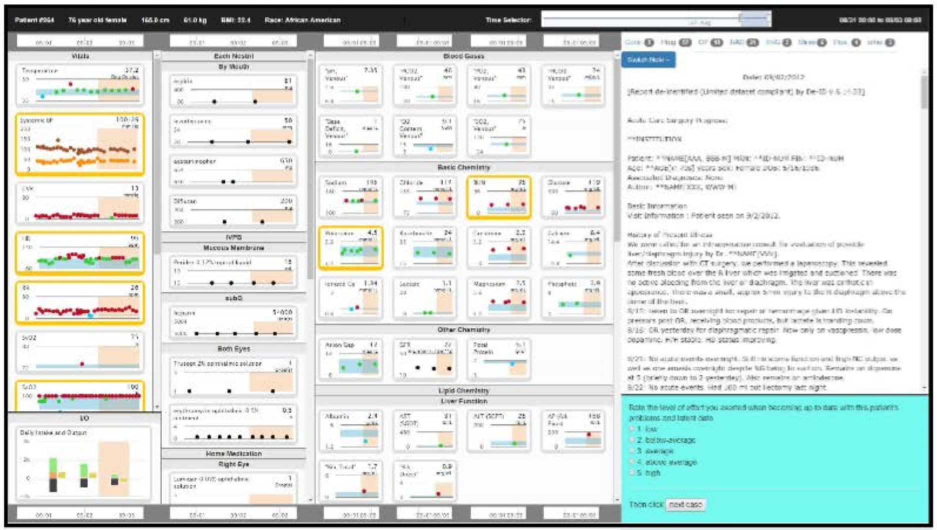
Use the audio recorder to record your presentation.

Start the recording with **Interface_demo** standing for **264**

Then click continue

Task C.

Demo Case 2 shown in Version 2



Patient #264 76 year old female 165.0 cm 61.0 kg BMI: 22.4 Race: African American Time Selector: 08/31 00:00 to 09/03 08:00

Date: 09/02/2012

[Report de-identified (Limited dataset compliant) by De-ID v.6.19.03]

Acute Care Surgery Progress

**INSTITUTION

Patient: **NAME[AAA, BBB M] MRN: **ID-NUM FIN: **ID-NUM
Age: **AGE[in 70s] years Sex: Female DOB: 5/16/19xx
Associated Diagnoses: None
Author: **NAME[XXX, WWW M]

Basic Information
Visit Information : Patient seen on 9/2/2012.

History of Present Illness
We were called for an intraoperative consult for evaluation of possible liver/diaphragm injury by Dr. **NAME[VVV].
After discussion with CT surgery, we performed a laparoscopy. This revealed some fresh blood over the R liver which was irrigated and suctioned. There was no active bleeding from the liver or diaphragm. The liver was cirrhotic in appearance. There was a small, approx 5mm injury to the R diaphragm above the dome of the liver.
8/15: taken to OR overnight for repair of hemorrhage given HD instability. On pressors post OR, receiving blood products, but lactate is trending down.
8/16: OR yesterday for diaphragmatic repair. Now only on vasopressin, low dose dopamine. H/H stable. HD status improving.

8/21: No acute events overnight. Still no stoma function and high NG output as well as one emesis overnight despite NG being to suction. Remains on dopamine at 5 (briefly down to 2 yesterday). Also remains on amiodarone.
8/22: No acute events. Had 100 ml out ileostomy last night.

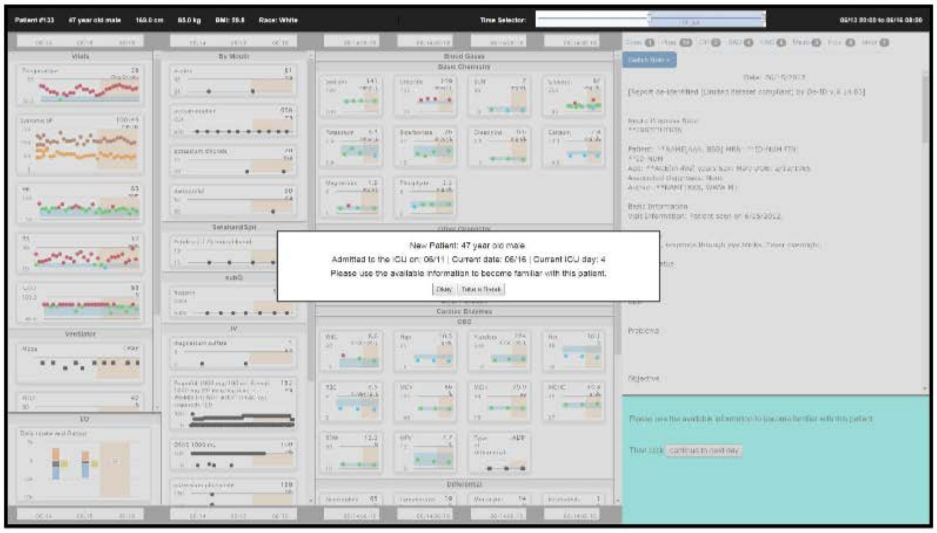
Rate the level of effort you exerted when becoming up to date with this patient's problems and latest data

○ 1: low
○ 2: below-average
○ 3: average
○ 4: above average
○ 5: high

Then click next case

Task D.

Demo Case 3 shown in Version 3



New Patient: 47 year old male

Admitted to the ICU on: 06/11 | Current date: 06/16 | Current ICU day: 4

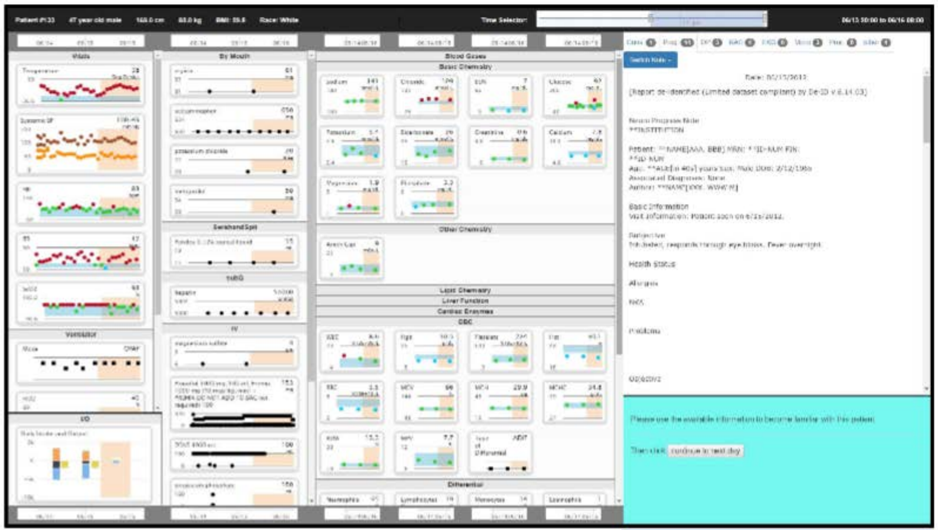
Please use the available information to become familiar with this patient.

Okay | Take a Break

Please use the available information to become familiar with this patient.

Then click continue to next day

Demo Case 3 shown in Version 3



Date: 06/15/2012

[Report de-identified (Limited dataset compliant) by De-ID v.6.14.03]

Neuro Progress Note
**INSTITUTION

Patient: **NAME[AAA, BBB] MRN: **ID-NUM FIN:
**ID-NUM
Age: **AGE[in 40s] years Sex: Male DOB: 2/12/1965
Associated Diagnoses: None
Author: **NAME[XXX, WWW M]

Basic Information
Visit information: Patient seen on 6/15/2012.

Subjective
Intubated, responds through eye blinks. Fever overnight.

Health Status

Allergies

NKA

Problems

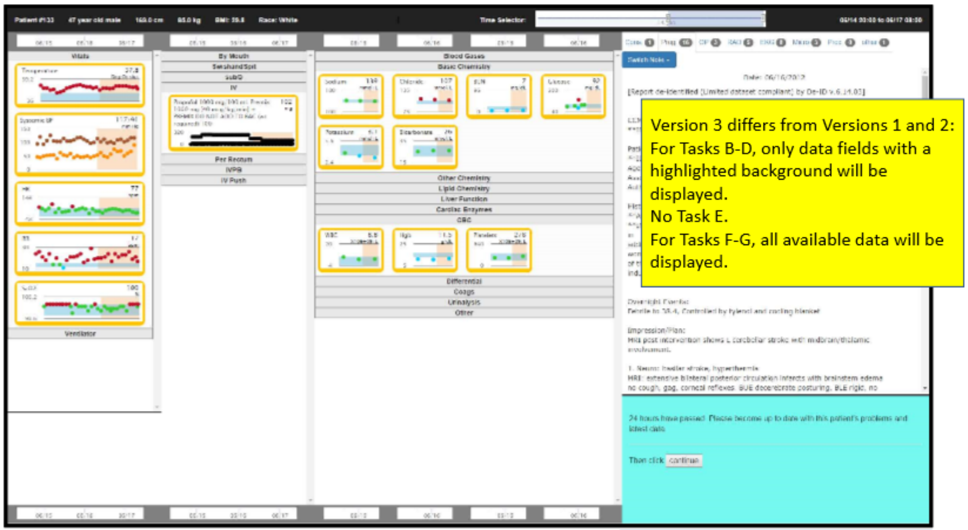
Objective

Please use the available information to become familiar with this patient.

Then click continue to next day

Task A.

Demo Case 3 shown in Version 3



Date: 06/16/2012

[Report de-identified [Limited dataset compliant] by De-ID v.6.14.03]

Overnight Events:
Febrile to 38.4, Controlled by tylenol and cooling blanket

Impression/Plan:
MRI post intervention shows L cerebellar stroke with midbrain/thalamic involvement.

1. Neuro: basilar stroke, hyperthermia
MRI: extensive bilateral posterior circulation infarcts with brainstem edema
no cough, gag, corneal reflexes. BUE decerebrate posturing. BLE rigid, no

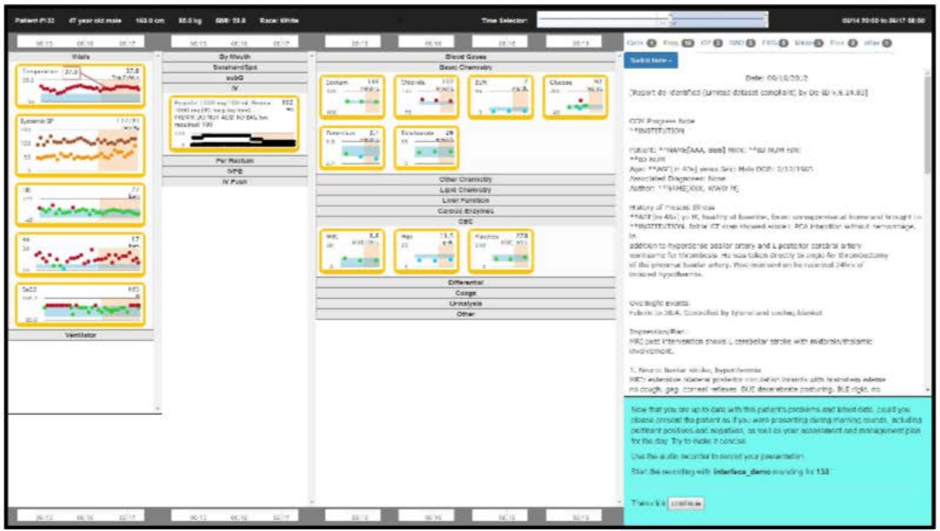
24 hours have passed. Please become up to date with this patient's problems and latest data.

Then click continue

Version 3 differs from Versions 1 and 2:
For Tasks B-D, only data fields with a highlighted background will be displayed.
No Task E.
For Tasks F-G, all available data will be displayed.

Task B.

Demo Case 3 shown in Version 3



Date: 06/16/2012

[Report de-identified [Limited dataset compliant] by De-ID v.6.14.03]

CCM Progress Note
**INSTITUTION

Overnight events:
Febrile to 38.4, Controlled by tylenol and cooling blanket

Impression/Plan:
MRI post intervention shows L cerebellar stroke with midbrain/thalamic involvement.

1. Neuro: basilar stroke, hyperthermia

Now that you are up to date with this patient's problems and latest data, could you please present the patient as if you were presenting during morning rounds, including pertinent positives and negatives, as well as your assessment and management plan for the day. Try to make it concise.

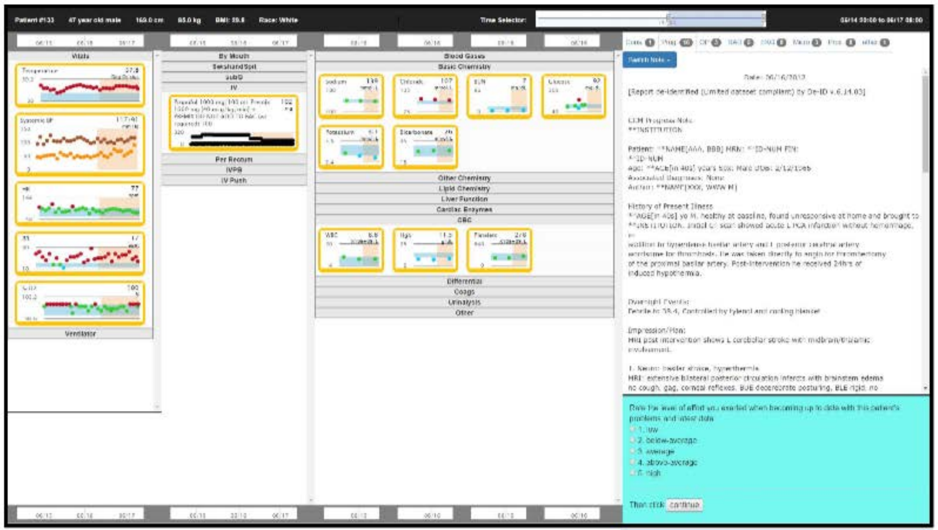
Use the audio recorder to record your presentation.

Start the recording with **interface_demo** standing for **133**

Then click continue

Task C.

Demo Case 3 shown in Version 3



Date: 06/16/2012

[Report de-identified (Limited dataset compliant) by De-ID v.6.14.03]

CCM Progress Note
**INSTITUTION

Patient: **NAME[AAA, BBB] MRN: **ID-NUM FIN:
**ID-NUM
Age: **AGE[in 40s] years Sex: Male DOB: 2/12/1965
Associated Diagnoses: None
Author: **NAME[XXX, WWW M]

History of Present Illness
**AGE[in 40s] yo M, healthy at baseline, found unresponsive at home and brought to
**INSTITUTION. Initial CT scan showed acute L PCA infarction without hemorrhage,
in
addition to hyperdense basilar artery and L posterior cerebral artery
worrisome for thrombosis. He was taken directly to angio for thrombectomy
of the proximal basilar artery. Post-intervention he received 24hrs of
induced hypothermia.

Overnight Events:
Febrile to 38.4, Controlled by tylenol and cooling blanket

Impression/Plan:
MRI post intervention shows L cerebellar stroke with midbrain/thalamic
involvement.

1. Neuro: basilar stroke, hyperthermia
MRI: extensive bilateral posterior circulation infarcts with brainstem edema
no cough, gag, corneal reflexes. BUE decerebrate posturing, BLE rigid, no

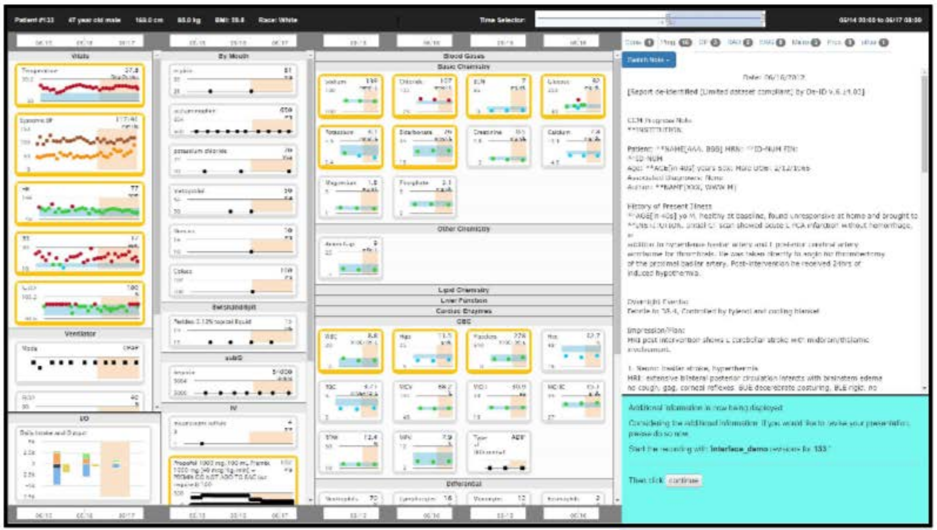
Rate the level of effort you exerted when becoming up to date with this patient's problems and latest data

○ 1. low
○ 2. below-average
○ 3. average
○ 4. above-average
○ 5. high

Then click continue

Task D.

Demo Case 3 shown in Version 3



Date: 06/16/2012

[Report de-identified (Limited dataset compliant) by De-ID v.6.14.03]

CCM Progress Note
**INSTITUTION

Patient: **NAME[AAA, BBB] MRN: **ID-NUM FIN:
**ID-NUM
Age: **AGE[in 40s] years Sex: Male DOB: 2/12/1965
Associated Diagnoses: None
Author: **NAME[XXX, WWW M]

History of Present Illness
**AGE[in 40s] yo M, healthy at baseline, found unresponsive at home and brought to
**INSTITUTION. Initial CT scan showed acute L PCA infarction without hemorrhage,
in
addition to hyperdense basilar artery and L posterior cerebral artery
worrisome for thrombosis. He was taken directly to angio for thrombectomy
of the proximal basilar artery. Post-intervention he received 24hrs of
induced hypothermia.

Overnight Events:
Febrile to 38.4, Controlled by tylenol and cooling blanket

Impression/Plan:
MRI post intervention shows L cerebellar stroke with midbrain/thalamic
involvement.

1. Neuro: basilar stroke, hyperthermia
MRI: extensive bilateral posterior circulation infarcts with brainstem edema
no cough, gag, corneal reflexes. BUE decerebrate posturing, BLE rigid, no

Additional information is now being displayed

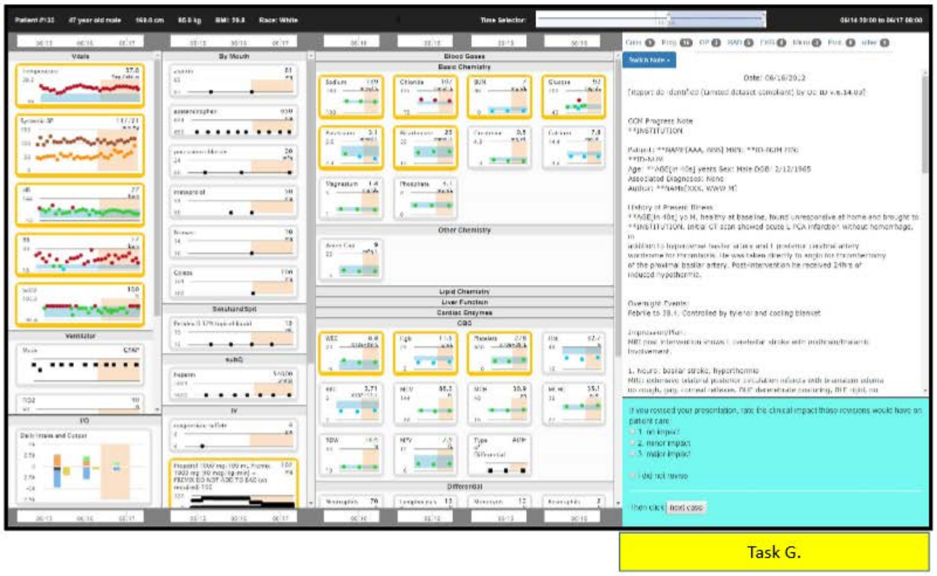
Considering the additional information, if you would like to revise your presentation, please do so now.

Start the recording with "**interface_demo** revision for 133"

Then click continue

Task F. (no Task E in Version 3)

Demo Case 3 shown in Version 3



Date: 06/16/2012

[Report de-identified (Limited dataset compliant) by DE-ID v.6.14.02]

CCN Progress Note
**INSTITUTION

Patient: **NAME[AAA, BBB] MRN: **ID-NUM FIN:
**ID-NUM
Age: **AGE[in 40s] years Sex: Male DOB: 2/12/1965
Associated Diagnoses: None
Author: **NAME[XXX, WWW M]

History of Present Illness
**AGE[in 40s] yo M, healthy at baseline, found unresponsive at home and brought to **INSTITUTION. Initial CT scan showed acute L PCA infarction without hemorrhage, in
addition to hyperdense basilar artery and L posterior cerebral artery worrisome for thrombosis. He was taken directly to angio for thrombectomy of the proximal basilar artery. Post-intervention he received 24hrs of induced hypothermia.

Overnight Events:
Febrile to 38.1. Controlled by tylenol and cooling blanket

Impression/Plan:
MRI post intervention shows [illegible] cerebellar stroke with [illegible] involvement.

1. Neuro: basilar stroke, hyperthermia
MRI: extensive bilateral posterior circulation infarcts with brainstem edema
no cough, gag, corneal reflexes, [illegible]

If you revised your presentation, rate the clinical impact those revisions would have on patient care:

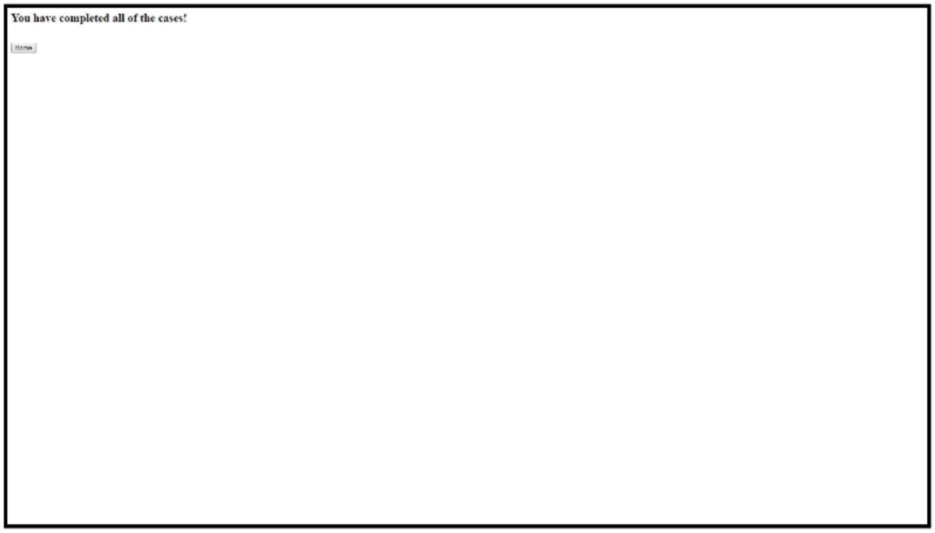
○ 1. no impact
○ 2. minor impact
○ 3. major impact

○ I did not revise

Then click next case

Task G.

Study completed screen


You have completed all of the cases!

APPENDIX E. PERFORMANCE OF MODELS FOR EIGHTY DATA ITEMS

Table 21. AUROC performance of models for eighty data items.

Model type	Semi-personalized	General		Semi-personalized	
Training data set	178 case manual selection	178 case manual selection	147 case eye-tracking	178 case manual selection	147 case eye-tracking
Evaluation data set (gold standard)	leave-one-out on training data set	68 case manual selection		26 case manual selection	
Target	**AUROC**				
red blood cells	0.94	**–**	–	–	–
magnesium sulfate	0.83	–	–	–	0.44
ventilator status	0.83	0.42	0.47	–	–
PEEP	0.83	**–**	–	–	–
pH	0.77	0.66	0.76	0.46	0.58
bicarbonate (blood gases)	0.75	0.33	0.59	0.9	0.46
vancomycin	0.74	0.47	0.57	–	0.60
anion gap	0.74	–	–	–	–
oxygen saturation	0.74	**0.66**	0.48	0.69	0.44
bilirubin total	0.73	0.89	0.71	0.67	0.72
lactate	0.73	*0.79*	0.53	0.78	0.57
piperacillin-tazobactam	0.73	*0.67*	0.69	0.76	0.67
norepinephrine	0.72	0.79	0.49	–	0.67
chloride	0.71	**0.44**	0.59	0.51	0.50
alkaline phosphatase	0.71	0.64	0.50	0.7	0.42
potassium chloride	0.71	–	–	–	–
heparin	0.71	0.34	0.56	0.35	0.65
glucose	0.71	**0.45**	0.46	0.52	0.54
aspirin	0.71	0.48	0.85	–	0.85
fentanyl	0.70	0.51	0.74	–	0.30
fraction of inspired O2	0.69	**0.75**	0.78	0.75	0.79
central venous pressure	0.69	0.48	0.50	–	–
calcium	0.68	0.31	0.56	–	–
magnesium	0.68	0.48	0.60	0.29	0.55
respiratory rate	0.68	**0.62**	0.5	0.78	0.71

See next page for table notes

Table 21 (continued).

Model type	Semi-personalized	General		Semi-personalized	
Training data set	178 case manual selection	178 case manual selection	147 case eye-tracking	178 case manual selection	147 case eye-tracking
Evaluation data set (gold standard)	leave-one-out on training data set	68 case manual selection		26 case manual selection	
Target	**AUROC**				
famotidine	0.68	0.5	0.29	–	0.40
blood urea nitrogen	0.68	**0.38**	0.39	0.85	0.35
partial thromboplastin time	0.68	0.63	0.50	–	–
ventilator mode	0.67	*0.66*	0.72	0.79	0.76
partial pressure of CO2	0.67	0.98	0.70	0.94	0.76
neutrophils	0.67	0.29	0.44	0.33	0.38
temperature	0.67	**0.51**	0.36	0.53	0.70
intake and output	0.66	0.45	0.68	0.53	0.70
glomerular filtration rate	0.66	0.47	0.35	–	–
phosphate	0.65	0.49	0.58	0.61	0.52
aspartate aminotransferase	0.65	0.78	0.56	0.69	0.59
alanine aminotransferase	0.65	0.8	0.64	0.74	0.58
INR	0.65	0.61	0.71	0.76	0.76
platelets	0.65	**0.66**	0.54	0.49	0.64
creatinine	0.65	**0.58**	0.50	0.14	0.60
blood pressure	0.65	**0.57**	0.50	–	–
dextrose 5% in water	0.65	0.64	0.52	–	0.61
ampicillin-sulbactam	0.65	*0.83*	0.59	0.82	0.88
potassium	0.64	**0.42**	0.57	0.62	0.63
albumin	0.64	0.72	0.30	0.86	0.64
venous saturation of oxygen	0.64	–	–	–	–
bicarbonate (chemistry)	0.64	**0.68**	0.44	0.56	0.64
white blood cells	0.64	**0.58**	0.50	–	–
sodium	0.64	**0.38**	0.59	0.49	0.53
venous pH	0.64	–	–	–	–
partial pressure of O2	0.64	0.79	0.48	0.95	0.69
Senna	0.64	–	–	–	–
prothrombin time	0.64	0.3	0.5	0.4	1.00

Targets that are bolded or italicized were applied in the evaluation study (Chapter 7.1). The version applied was the general model trained on the manual selection data set. The 68 case manual selection data set consists of multiple participants evaluating the same 18 patient cases. The 26 case manual selection data set is a five-participant subset of the 68 cases.

Table 21 (continued).

Model type	Semi-personalized	General		Semi-personalized	
Training data set	178 case manual selection	178 case manual selection	147 case eye-tracking	178 case manual selection	147 case eye-tracking
Evaluation data set (gold standard)	leave-one-out on training data set	68 case manual selection		26 case manual selection	
Target	**AUROC**				
hemoglobin	0.63	**0.47**	0.42	0.42	0.57
bilirubin direct	0.63	0.50	0.49	–	0.54
albuterol-ipratropium	0.63	0.27	0.48	–	0.5
heart rate	0.62	**0.55**	0.48	0.67	0.56
ionized calcium	0.62	0.31	0.59	0.42	0.56
midazolam	0.62	0.41	0.50	–	–
Propofol	0.61	0.68	0.58	–	0.72
base solution	0.61	**0.67**	0.38	0.83	0.44
pantoprazole	0.61	0.56	0.72	–	0.81
insulin (Humulin & Novolin)	0.61	0.59	0.84	–	0.85
insulin aspart (Novolog)	0.60	–	–	–	–
sodium chloride 0.9%	0.59	0.59	0.68	0.64	0.74
ventilator tube status	0.59	0.42	0.27	0.30	0.72
metoprolol	0.58	0.12	0.49	–	0.26
vancomycin, trough	0.57	0.61	0.40	–	0.41
ammonia	0.57	–	–	–	–
hematocrit	0.56	–	–	–	–
chlorhexidine topical	0.56	0.54	0.66	–	0.28
metronidazole	0.55	*0.74*	0.45	0.85	0.31
furosemide	0.54	0.05	0.51	–	0.56
Troponin	0.52	0.52	0.95	–	0.98
band neutrophils	0.52	0.50	0.39	–	0.63
insulin glargine (Lantus)	0.50	0.46	0.58	0.30	0.38
acetaminophen	0.47	0.19	0.80	–	–
Lorazepam	0.45	0.47	0.28	–	0.33
fibrinogen	0.41	–	–	–	–
hydrocortisone	0.40	0.44	0.95	–	0.54

Targets that are bolded or italicized were applied in the evaluation study (Chapter 7.1). The version applied was the general model trained on the manual selection data set. The 68 case manual selection data set consists of multiple participants evaluating the same 18 patient cases. The 26 case manual selection data set is a five-participant subset of the 68 cases.

BIBLIOGRAPHY

1. Institute of Medicine. *The Learning Healthcare System*. (Olsen L, Aisner D, McGinnis JM, eds.). Washington, D.C.: National Academies Press; 2007. doi:10.17226/11903.

2. Callen J, Georgiou A, Li J, Westbrook JI. The safety implications of missed test results for hospitalised patients: a systematic review. *BMJ Qual Saf*. 2011;20(2):194-199. doi:10.1136/bmjqs.2010.044339.

3. Institute of Medicine. *To Err Is Human*. Washington, D.C.: National Academies Press; 2000. doi:10.17226/9728.

4. Classen DC, Resar R, Griffin F, et al. 'Global trigger tool' shows that adverse events in hospitals may be ten times greater than previously measured. *Health Aff*. 2011;30(4):581-589. doi:10.1377/hlthaff.2011.0190.

5. Makary MA, Daniel M. Medical error—the third leading cause of death in the US. *BMJ*. 2016;353:i2139. doi:10.1136/bmj.i2139.

6. Healey CG, Enns JT. Attention and visual memory in visualization and computer graphics. *IEEE Trans Vis Comput Graph*. 2012;18(7):1170-1188. doi:10.1109/TVCG.2011.127.

7. Kuperman GJ, Teich JM, Bates DW, et al. Detecting alerts, notifying the physician, and offering action items: a comprehensive alerting system. *AMIA Annu Fall Symp Proc*. January 1996:704-708.

8. Ratib O. From multimodality digital imaging to multimedia patient record. *Comput Med Imaging Graph*. 1994;18(2):59-65. doi:10.1016/0895-6111(94)90014-0.

9. Baumann BM, Holmes JH, Chansky ME, Levey H, Kulkarni M, Boudreaux ED. Pain assessments and the provision of analgesia: the effects of a templated chart. *Acad Emerg Med*. 2007;14(1):47-52. doi:10.1197/j.aem.2006.06.057.

10. Rose EA, Deshikachar AM, Schwartz KL, Severson RK. Use of a template to improve documentation and coding. *Fam Med*. 2001;33(7):516-521.

11. Shachak A, Hadas-Dayagi M, Ziv A, Reis S. Primary care physicians' use of an electronic medical record system: a cognitive task analysis. *J Gen Intern Med*. 2009;24(3):341-348. doi:10.1007/s11606-008-0892-6.

12. Kannampallil TG, Jones LK, Patel VL, Buchman TG, Franklin A. Comparing the information seeking strategies of residents, nurse practitioners, and physician assistants in critical care settings. *J Am Med Inform Assoc*. 2014;21(e2). doi:10.1136/amiajnl-2013-002615.

13. Sittig DF, Wright A, Osheroff JA, et al. Grand challenges in clinical decision support. *J Biomed Inform*. 2008;41(2):387-392. doi:10.1016/j.jbi.2007.09.003.

14. Cimino JJ, Johnson SB, Aguirre A, Roderer N, Clayton PD. The MEDLINE button. *Annu Symp Comput Appl Med Care Proc*. January 1992:81-85.

15. Dexter PR, Perkins S, Overhage JM, Maharry K, Kohler RB, McDonald CJ. A computerized reminder system to increase the use of preventive care for hospitalized patients. *N Engl J Med*. 2001;345(13):965-970. doi:10.1056/NEJMsa010181.

16. Duan L, Street WN, Xu E. Healthcare information systems: data mining methods in the creation of a clinical recommender system. *Enterp Inf Syst*. 2011;5(2):169-181. doi:10.1080/17517575.2010.541287.

17. Bates DW, Kuperman GJ, Wang S, et al. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. *J Am Med Informatics Assoc*. 2003;10(6):523-530. doi:10.1197/jamia.M1370.

18. Feblowitz JC, Wright A, Singh H, Samal L, Sittig DF. Summarization of clinical information: a conceptual model. *J Biomed Inform*. 2011;44(4):688-699. doi:10.1016/j.jbi.2011.03.008.

19. Laxmisan A, McCoy AB, Wright A, Sittig DF. Clinical summarization capabilities of commercially-available and internally-developed electronic health records. *Appl Clin Inform*. 2012;3(1):80-93. doi:10.4338/ACI-2011-11-RA-0066.

20. Pivovarov R, Elhadad N. Automated methods for the summarization of electronic health records. *J Am Med Informatics Assoc*. 2015;22(5):938-947. doi:10.1093/jamia/ocv032.

21. Manor-Shulman O, Beyene J, Frndova H, Parshuram CS. Quantifying the volume of documented clinical information in critical illness. *J Crit Care*. 2008;23(2):245-250. doi:10.1016/j.jcrc.2007.06.003.

22. Charles D, King J, Patel V, Furukawa M. *Adoption of Electronic Health Record Systems among US Non-Federal Acute Care Hospitals: 2008-2012.*; 2013. https://www.healthit.gov/sites/default/files/data-brief/2014HospitalAdoptionDataBrief.pdf. Accessed May 19, 2016.

23. Hillestad R, Bigelow J, Bower A, et al. Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. *Health Aff*. 2005;24(5):1103-1117. doi:10.1377/hlthaff.24.5.1103.

24. Moacdieh N, Sarter N. Clutter in electronic medical records. *Hum Factors J Hum Factors Ergon Soc*. 2015;57(4):591-606. doi:10.1177/0018720814564594.

25. Moacdieh N, Ganje T, Sarter N. Electronic health records: effects of clutter and stress on physicians' information search and noticing performance. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol 58. SAGE Publications; 2014:718-722. doi:10.1177/1541931214581167.

26. Singh H, Spitzmueller C, Petersen NJ, Sawhney MK, Sittig DF. Information overload and missed test results in electronic health record–based settings. *JAMA Intern Med*. 2013;173(8):702–704. doi:10.1001/2013.jamainternmed.61.

27. Hall A, Walton G. Information overload within the health care system: a literature review. *Heal Inf Libr J*. 2004;21(2):102-108. doi:10.1111/j.1471-1842.2004.00506.x.

28. Poissant L. The impact of electronic health records on time efficiency of physicians and nurses: a systematic review. *J Am Med Informatics Assoc*. 2005;12(5):505-516. doi:10.1197/jamia.M1700.

29. Zheng K, Padman R, Johnson MP, Diamond HS. An interface-driven analysis of user interactions with an electronic health records system. *J Am Med Informatics Assoc*. 2009;16(2):228-237. doi:10.1197/jamia.M2852.

30. Wahls TL, Cram PM. The frequency of missed test results and associated treatment delays in a highly computerized health system. *BMC Fam Pract*. 2007;8(1):32. doi:10.1186/1471-2296-8-32.

31. American Medical Association. Improving Care: Priorities to Improve Electronic Health Record Usability. 2014. https://www.aace.com/files/ehr-priorities.pdf. Accessed August 26, 2015.

32. Sittig DF. Eight Rights of Safe Electronic Health Record Use. *JAMA*. 2009;302(10):1111. doi:10.1001/jama.2009.1311.

33. McDonald CJ. Toward electronic medical record alerts that consume less physician time. *JAMA Intern Med*. 2013;173(18):1755. doi:10.1001/jamainternmed.2013.9332.

34. Ash JS. Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. *J Am Med Informatics Assoc*. 2003;11(2):104-112. doi:10.1197/jamia.M1471.

35. Prados-Suárez B, Molina C, Peña Yañez C, Prados de Reyes M. Improving electronic health records retrieval using contexts. *Expert Syst Appl*. 2012;39(10):8522-8536. doi:10.1016/j.eswa.2012.01.016.

36. Pickering BW, Gajic O, Ahmed A, Herasevich V, Keegan MT. Data utilization for medical decision making at the time of patient admission to ICU. *Crit Care Med*. 2013;41(6):1502-1510. doi:10.1097/CCM.0b013e318287f0c0.

37. Fischer G. Context-aware systems. In: *International Working Conference on Advanced Visual Interfaces Proceedings*. New York, New York, USA: ACM Press; 2012:287. doi:10.1145/2254556.2254611.

38. Osheroff JA, Teich JM, Levick D, et al. *Improving Outcomes with Clinical Decision Support: An Implementer's Guide*. 2nd ed. Chicago, IL: Healthcare Information and Management Systems Society; 2012.

39. Baysari MT, Tariq A, Day RO, Westbrook JI. Alert override as a habitual behavior - A new perspective on a persistent problem. *J Am Med Informatics Assoc*. 2017;24(2):409-412. doi:10.1093/jamia/ocw072.

40. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. *Acad Med*. 2003;78(8):775-780. doi:10.1097/00001888-200308000-00003.

41. Levinson W, Kallewaard M, Bhatia RS, Wolfson D, Shortt S, Kerr EA. 'Choosing wisely': a growing international campaign. *BMJ Qual Saf*. 2015;24(2):167-174. doi:10.1136/bmjqs-2014-003821.

42. Simons DJ, Levin DT. Change blindness. *Trends Cogn Sci*. 1997;1(7):261-267. doi:10.1016/S1364-6613(97)01080-2.

43. Ball F, Busch NA. Change detection on a hunch: pre-attentive vision allows "sensing" of unique feature changes. *Attention, Perception, Psychophys*. 2015;77(8):2570-2588. doi:10.3758/s13414-015-0963-9.

44. Fitzgerald FS. *The Great Gatsby*. New York, New York: Scribner; 2004.

45. Bryant AD, Fletcher GS, Payne TH. Drug interaction alert override rates in the Meaningful Use era. *Appl Clin Inform*. 2014;05(03):802-813. doi:10.4338/ACI-2013-12-RA-0103.

46. Robertson S, Penzak S. Drug Interactions. In: *Principles of Clinical Pharmacology*. Vol Elsevier; 2007:229-247. doi:10.1016/B978-012369417-1/50055-9.

47. Surowiecki J. *The Wisdom of Crowds*. New York, New York: Anchor Books; 2005.

48. Koch C, Roberts K, Petruccelli C, Morgan DJ. The frequency of unnecessary testing in hospitalized patients. *Am J Med*. 2018;131(5):500-503. doi:10.1016/j.amjmed.2017.11.025.

49. Cassel CK. Choosing wisely. *JAMA*. 2012;307(17):1801. doi:10.1001/jama.2012.476.

50. Studdert DM. Defensive medicine among high-risk specialist physicians in a volatile malpractice environment. *JAMA*. 2005;293(21):2609. doi:10.1001/jama.293.21.2609.

51. Institute of Medicine. *Digital Infrastructure for the Learning Health System*. Washington, D.C.: National Academies Press; 2011. doi:10.17226/12912.

52. Friedman CP, Wong AK, Blumenthal D. Achieving a Nationwide Learning Health System. *Sci Transl Med*. 2010;2(57):57cm29-57cm29. doi:10.1126/scitranslmed.3001456.

53. National Institutes of Health. NIH-wide strategic plan, fiscal years 2016–2020: turning discovery into health. 2016. https://www.nih.gov/sites/default/files/about-nih/strategic-plan-fy2016-2020-508.pdf. Accessed July 19, 2018.

54. Friedman CP, Rubin JC, Sullivan KJ. Toward an information infrastructure for global health improvement. *Yearb Med Inform*. 2017;26(01):16-23. doi:10.15265/IY-2017-004.

55. Office of the National Coordinator for Health Information Technology. Clinical Decision Support. *HealthIT.gov*. https://www.healthit.gov/topic/safety/clinical-decision-support. Accessed May 10, 2018.

56. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. *J Am Med Informatics Assoc*. 2017;24(6):1052-1061. doi:10.1093/jamia/ocx030.

57. Jha AK. Meaningful use of electronic health records. *JAMA*. 2010;304(15):1709. doi:10.1001/jama.2010.1497.

58. Focsa M, Mihalas GI. EHR ecosystem. In: *Pervasive and Mobile Sensing and Computing for Healthcare*. Vol 2. Berlin, Heidelberg: Springer; 2013:251-268. doi:10.1007/978-3-642-32538-0_12.

59. Torda P, Tinoco A. Achieving the promise of electronic health record-enabled quality measurement: a measure developer's perspective. *Gener Evid Methods to Improv Patient Outcomes*. 2013;1(2):1031. doi:10.13063/2327-9214.1031.

60. Holroyd-Leduc JM, Lorenzetti D, Straus SE, Sykes L, Quan H. The impact of the electronic medical record on structure, process, and outcomes within primary care: a systematic review of the evidence. *J Am Med Informatics Assoc*. 2011;18(6):732-737. doi:10.1136/amiajnl-2010-000019.

61. Shachak A, Reis S. The impact of electronic medical records on patient-doctor communication during consultation: a narrative literature review. *J Eval Clin Pract*. 2009;15(4):641-649. doi:10.1111/j.1365-2753.2008.01065.x.

62. DesRoches CM, Campbell EG, Rao SR, et al. Electronic health records in ambulatory care: a national survey of physicians. *N Engl J Med*. 2008;359(1):50-60. doi:10.1056/NEJMsa0802005.

63. Holden RJ. Physicians' beliefs about using EMR and CPOE: In pursuit of a contextualized understanding of health IT use behavior. *Int J Med Inform*. 2010;79(2):71-80. doi:10.1016/j.ijmedinf.2009.12.003.

64. Koppel R, Metlay JP, Cohen A, et al. Role of computerized physician order entry systems in facilitating medication errors. *JAMA*. 2005;293(10):1197. doi:10.1001/jama.293.10.1197.

65. Han H, Lopp L. Writing and reading in the electronic health record: an entirely new world. *Med Educ Online*. 2013;18(1):18634. doi:10.3402/meo.v18i0.18634.

66. Asan O, D. Smith P, Montague E. More screen time, less face time: implications for EHR design. *J Eval Clin Pract*. 2014;20(6):896-901. doi:10.1111/jep.12182.

67. Meeks DW, Smith MW, Taylor L, Sittig DF, Scott JM, Singh H. An analysis of electronic health record-related patient safety concerns. *J Am Med Inform Assoc*. 2014;21(6):1053-1059. doi:10.1136/amiajnl-2013-002578.

68. Grams R. The "new" america electronic medical record (EMR)—design criteria and challenge. *J Med Syst*. 2009;33(6):409-411. doi:10.1007/s10916-009-9319-0.

69. Edwards PJ, Moloney KP, Jacko JA, Sainfort F. Evaluating usability of a commercial electronic health record: A case study. *Int J Hum Comput Stud*. 2008;66(10):718-728. doi:10.1016/j.ijhcs.2008.06.002.

70. Ball MJ, Garets DE, Handler TJ. Leveraging information technology towards enhancing patient care and a culture of safety in the U.S. *Methods Inf Med*. 2003;42(05):503-508. doi:10.1055/s-0038-1634376.

71. Varshney U. Context-awareness in Healthcare. In: *Pervasive Healthcare Computing: EMR/EHR, Wireless and Health Monitoring*. Boston, MA: Springer US; 2009:231-257. doi:10.1007/978-1-4419-0215-3_11.

72. Law AS, Freer Y, Hunter J, Logie RH, Mcintosh N, Quinn J. A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit. *J Clin Monit Comput*. 2005;19(3):183-194. doi:10.1007/s10877-005-0879-3.

73. Monroe M, Rongjian Lan, Hanseung Lee, Plaisant C, Shneiderman B. Temporal event sequence simplification. *IEEE Trans Vis Comput Graph*. 2013;19(12):2227-2236. doi:10.1109/TVCG.2013.200.

74. Koch SH, Weir C, Westenskow D, et al. Evaluation of the effect of information integration in displays for ICU nurses on situation awareness and task completion time: a prospective randomized controlled study. *Int J Med Inform*. 2013;82(8):665-675. doi:10.1016/j.ijmedinf.2012.10.002.

75. Hsu W, Taira RK, El-Saden S, Kangarloo H, Bui AAT. Context-based electronic health record: toward patient specific healthcare. *IEEE Trans Inf Technol Biomed*. 2012;16(2):228-234. doi:10.1109/TITB.2012.2186149.

76. Suermondt HJ, Tang PC, Strong PC, Young CY, Annevelink J. Automated identification of relevant patient information in a physician’s workstation. *Symp Comput Appl Med Care Proc*. 1993:229-232.

77. Zeng Q. Providing concept-oriented views for clinical data using a knowledge-based system: an evaluation. *J Am Med Informatics Assoc*. 2002;9(3):294-305. doi:10.1197/jamia.M1008.

78. Zeng Q, Cimino JJ. A knowledge-based, concept-oriented view generation system for clinical data. *J Biomed Inform*. 2001;34(2):112-128. doi:10.1006/jbin.2001.1013.

79. Senathirajah Y, Bakken S, Kaufman D. The clinician in the Driver’s Seat: Part 1–A drag/drop user-composable electronic health record platform. *J Biomed Inf*. 2014;52:165-176.

80. Plaisant C, Milash B, Rose A, Widoff S, Shneiderman B. LifeLines: visualizing personal histories. *Proc SIGCHI Conf Hum factors Comput Syst common Gr*. 1996:221-227. doi:10.1145/238386.238493.

81. Pickering BW, Herasevich V, Ahmed A, Gajic O. Novel representation of clinical information in the ICU. *Appl Clin Inform*. 2010;01(02):116-131. doi:10.4338/ACI-2009-12-CR-0027.

82. Zeng Q, Cimino JJ. Evaluation of a system to identify relevant patient information and its impact on clinical information retrieval. *AMIA Annu Symp Proc*. 1999:642-646.

83. Pickering BW, Dong Y, Ahmed A, et al. The implementation of clinician designed, human-centered electronic medical record viewer in the intensive care unit: a pilot step-wedge cluster randomized trial. *Int J Med Inform*. 2015;84(5):299-307. doi:10.1016/j.ijmedinf.2015.01.017.

84. Ahmed A, Chandra S, Herasevich V, Gajic O, Pickering BW. The effect of two different electronic health record user interfaces on intensive care provider task load, errors of cognition, and performance. *Crit Care Med*. 2011;39(7):1626-1634. doi:10.1097/CCM.0b013e31821858a0.

85. Tang PC, Annevelink J, Suermondt HJ, Young CY. Semantic integration of information in a physician’s workstation. *Int J Biomed Comput*. 1994;35(1):47-60. doi:10.1016/0020-7101(94)90048-5.

86. Smith SW, Koppel R. Healthcare information technology’s relativity problems: a typology of how patients’ physical reality, clinicians’ mental models, and healthcare information technology differ. *J Am Med Informatics Assoc*. 2014;21(1):117-131. doi:10.1136/amiajnl-2012-001419.

87. Jacob RJK, Karn KS. Eye tracking in human-computer interaction and usability research. In: *The Mind’s Eye*. Vol 1. Elsevier; 2003:573-605. doi:10.1016/B978-044451020-4/50031-1.

88. Wedel M, Pieters R. A review of eye-tracking research in marketing. In: Malhotra NK, ed. *Review of Marketing Research*. Emerald Group Publishing Limited; 2008:123-147. doi:10.1108/S1548-6435(2008)0000004009.

89. Blondon K, Wipfli R, Lovis C. Use of eye-tracking technology in clinical reasoning: a systematic review. *Stud Health Technol Inform*. 2015;210:90-94. doi:10.3233/978-1-61499-512-8-90.

90. Asan O, Yang Y. Using eye trackers for usability evaluation of health information technology: a systematic literature review. *JMIR Hum Factors*. 2015;2(1):e5. doi:10.2196/humanfactors.4062.

91. Henneman PL, Fisher DL, Henneman EA, et al. Providers do not verify patient identity during computer order entry. *Acad Emerg Med*. 2008;15(7):641-648. doi:10.1111/j.1553-2712.2008.00148.x.

92. Rick S, Calvitti A, Agha Z, Weibel N. Eyes on the clinic: accelerating meaningful interface analysis through unobtrusive eye tracking. *9th Int Conf Pervasive Comput Technol Healthc Proc*. May 2015:213-216. doi:10.4108/icst.pervasivehealth.2015.259276.

93. Gold JA, Stephenson LE, Gorsuch A, Parthasarathy K, Mohan V. Feasibility of utilizing a commercial eye tracker to assess electronic health record use during patient simulation. *Health Informatics J*. 2016;22(3):744-757. doi:10.1177/1460458215590250.

94. Doberne JW, He Z, Mohan V, Gold JA, Marquard J, Chiang MF. Using high-fidelity simulation and eye tracking to characterize EHR workflow patterns among hospital physicians. In: *AMIA Annu Symp Proc Symp*. Vol ; 2015:1881-1889.

95. Fong A, Hoffman DJ, Zachary Hettinger A, Fairbanks RJ, Bisantz AM. Identifying visual search patterns in eye gaze data; gaining insights into physician visual workflow. *J Am Med Informatics Assoc*. 2016;23(6):1180-1184. doi:10.1093/jamia/ocv196.

96. Brown PJ, Marquard JL, Amster B, et al. What do physicians read (and ignore) in electronic progress notes? *Appl Clin Inform*. 2014;05(02):430-444. doi:10.4338/ACI-2014-01-RA-0003.

97. Segall N, Taekman JM, Mark JB, Hobbs G, Wright MC. Coding and visualizing eye tracking data in simulated anesthesia care. *Hum Factors Ergon Soc Annu Meet Proc*. 2007;51(11):765-770. doi:10.1177/154193120705101134.

98. Reeder RW, Pirolli P, Card SK. WebEyeMapper and WebLogger: tools for analyzing eye tracking data collected in web-use studies. *CHI '01 Ext Abstr Hum Factors Comput Syst*. 2001:19-20. doi:http://doi.acm.org/10.1145/634067.634082.

99. Beymer D, Russell DM. WebGazeAnalyzer: a system for capturing and analyzing web reading behavior using eye gaze. *CHI '05 Ext Abstr Hum factors Comput Syst*. 2005:1913. doi:10.1145/1056808.1057055.

100. Wright MC, Dunbar S, Moretti EW, Schroeder RA, Taekman J, Segall N. Eye-tracking and retrospective verbal protocol to support information systems design. *Int Symp Hum Factors Ergon Heal Care Proc*. 2013;2(1):30-37. doi:10.1177/2327857913021007.

101. Angus DC, Barnato AE, Linde-Zwirble WT, et al. Use of intensive care at the end of life in the United States: an epidemiologic study. *Crit Care Med*. 2004;32(3):638-643. doi:10.1097/01.CCM.0000114816.62331.08.

102. Bigatello LM, Allain RM. Acute respiratory failure. In: O'Donnell JM, Nácul FE, eds. *Surgical Intensive Care Medicine*. Cham, Switzerland: Springer; 2016:319-334. doi:10.1007/978-3-319-19668-8_24.

103. Weisbord S, Palevsky P. Acute renal failure in the intensive care unit. *Semin Respir Crit Care Med*. 2006;27(3):262-273. doi:10.1055/s-2006-945527.

104. Pino CJ, Humes HD. Stem cell technology for the treatment of acute and chronic renal failure. *Transl Res*. 2010;156(3):161-168. doi:10.1016/j.trsl.2010.07.005.

105. Lieberthal W, Nigam SK. Acute renal failure. I. Relative importance of proximal vs. distal tubular injury. Lieberthal W, Nigam SK, eds. *Am J Physiol Physiol*. 1998;275(5):F623-F632. doi:10.1152/ajprenal.1998.275.5.F623.

106. Visweswaran S, Mezger J, Clermont G, Hauskrecht M, Cooper GF. Identifying deviations from usual medical care using a statistical approach. *AMIA Annu Symp Proceeding*. 2010:827-831.

107. Yount RJ, Vries JK, Councill CD. The medical archival system: an information retrieval system based on distributed parallel processing. *Inf Process Manag*. 1991;27(4):379-389. doi:10.1016/0306-4573(91)90091-Y.

108. Gortmaker SL, Hosmer DW, Lemeshow S. Applied logistic regression. *Contemp Sociol*. 1994;23(1):159. doi:10.2307/2074954.

109. Kotsiantis SB, Zaharakis ID, Pintelas PE. Machine learning: a review of classification and combining techniques. *Artif Intell Rev*. 2006;26(3):159-190. doi:10.1007/s10462-007-9052-3.

110. Safavian S, Landgrebe D. A survey of decision tree classifier methodology. *IEEE Trans Syst Man Cybern*. 1991;21(3):660-674. doi:10.1109/21.97458.

111. Brown S. VistA—U.S. Department of Veterans Affairs national-scale HIS. *Int J Med Inform*. 2003;69(2-3):135-156. doi:10.1016/S1386-5056(02)00131-4.

112. Mamlin BW, Biondich PG, Wolfe BA, et al. Cooking up an open source EMR for developing countries: OpenMRS - a recipe for successful collaboration. *AMIA Annu Symp Proc*. 2006:529-533.

113. Seebregts CJ, Mamlin BW, Biondich PG, et al. The OpenMRS implementers network. *Int J Med Inform*. 2009;78(11):711-720. doi:10.1016/j.ijmedinf.2008.09.005.

114. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825-2830.

115. Django Software Foundation. Django. 2013. https://www.djangoproject.com/.

116. Solomon P. The think aloud method: a practical guide to modelling cognitive processes. *Inf Process Manag*. 1995;31(6):906-907. doi:10.1016/0306-4573(95)90031-4.

117. Brooke J. SUS-a quick and dirty usability scale. In: Jordan PW, Thomas B, McClelland IL, Weerdmeester B, eds. *Usability Evaluation in Industry*. London, England: CRC Press; 1996:189-194.

118. Sauro J. Measuring usability with the System Usability Scale (SUS). *Meas Usability*. 2011. https://measuringu.com/sus/. Accessed May 2, 2015.

119. King AJ, Hochheiser H, Visweswaran S, Clermont G, Cooper GF. Eye-tracking for clinical decision support: a method to capture automatically what physicians are viewing in the EMR. *AMIA Jt Summits Transl Sci Proc*. 2017:512-521.

120. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155-163. doi:10.1016/j.jcm.2016.02.012.

121. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol*. 2011;2(3):27:1--27:27. doi:10.1145/1961189.1961199.

122. Breiman L. Random forests. *Mach Learn*. 2001;45:5-32. doi:10.1023/A:1010933404.

123. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. *MIS Q*. 2003;27(3):425. doi:10.2307/30036540.

124. Calvitti A, Hochheiser H, Ashfaq S, et al. Physician activity during outpatient visits and subjective workload. *J Biomed Inform*. 2017. doi:10.1016/j.jbi.2017.03.011.

125. Wu DTY, Smart N, Ciemins EL, Lanham HJ, Lindberg C, Zheng K. Using EHR audit trail logs to analyze clinical workflow: A case study from community-based ambulatory clinics. *AMIA Annu Symp Proc*. 2017;2017:1820-1827.

126. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. *J Am Med Informatics Assoc*. 2016;23(5):899-908. doi:10.1093/jamia/ocv189.

127. Allenby GM, Rossi PE, McCulloch RE. Hierarchical Bayes Models: A Practitioners Guide. Grover R, Vriens M, eds. *SSRN Electron J*. 2005. doi:10.2139/ssrn.655541.

128. Mandl KD. Ebola in the United States. *JAMA*. 2014;312(23):2499. doi:10.1001/jama.2014.15064.

129. Goddard K, Roudsari A, Wyatt JC. Automation bias: empirical results assessing influencing factors. *Int J Med Inform*. 2014;83(5):368-375. doi:10.1016/j.ijmedinf.2014.01.001.

130. Eghdam A, Forsman J, Falkenhav M, Lind M, Koch S. Combining usability testing with eye-tracking technology: Evaluation of a visualization support for antibiotic use in intensive care. In: *Studies in Health Technology and Informatics*. Vol 169. 2011;945-949. doi:10.3233/978-1-60750-806-9-945.

131. Forsman J, Anani N, Eghdam A, Falkenhav M, Koch S. Integrated information visualization to support decision making for use of antibiotics in intensive care: design and usability evaluation. *Informatics Heal Soc Care*. 2013;38(4):330-353. doi:10.3109/17538157.2013.812649.

132. Nielsona JA, Mamidala RN, Khan J. In-situ eye-tracking of emergency physician result review. *Stud Health Technol Inform*. 2013;192(1-2):1156. doi:10.3233/978-1-61499-289-9-1156.

133. Barkana DE, Açık A. Improvement of design of a surgical interface using an eye tracking device. *Theor Biol Med Model*. 2014;11(Suppl 1):S4. doi:10.1186/1742-4682-11-S1-S4.

134. Weibel N, Rick S, Emmenegger C, Ashfaq S, Calvitti A, Agha Z. LAB-IN-A-BOX: semi-automatic tracking of activity in the medical office. *Pers Ubiquitous Comput*. 2015;19(2):317-334. doi:10.1007/s00779-014-0821-0.

135. Mazur LM, Mosaly PR, Moore C, et al. Toward a better understanding of task demands, workload, and performance during physician-computer interactions. *J Am Med Informatics Assoc*. 2016;23(6):1113-1120. doi:10.1093/jamia/ocw016.