Enhancing Providers' Ability to Follow-up on Abnormal Test Results

Principal Investigator: Lukasz M. Mazur, PhD[1,3]

Co-Investigators: Prithima R. Mosaly, PhD[1], Lawrence Marks, MD[1], Javed Mostafa, PhD[3], Carlton Moore, MD[2]

[1]Division of Healthcare Engineering, University of North Carolina, Chapel Hill, NC
[2]Division of General Medicine, University of North Carolina, Chapel Hill, NC
[3]School of Information and Library Science, University of North Carolina, Chapel Hill, NC

Inclusive Dates of the Project: 8/01/2015 to 7/31/2017

Federal Project Officer: Steve Bernstein, AHRQ, CEPI, Division of Health IT

Grant Number: R21HS024062

Acknowledgment: This study was supported by the grant from the Agency for Healthcare Research and Quality. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality.

Abstract

Purpose: To assess provider task demands, workload, and performance during follow-up on abnormal test results in an electronic medical record (EMR) environment.

Scope: Our focus was on abnormal mammography and Pap smear results, and the baseline and enhanced EHR environment used for the study was Epic®.

Methods: Providers, randomized to regular/cross-coverage and low/high-volume conditions in the baseline/enhanced EMR environments, were assessed using analysis of variance (ANOVA) in order to quantify the impact of these conditions on task demands, workload, and performance.

Results: The high-volume of abnormal test results condition significantly increased task demands ($p<0.01$), providers' perception of workload ($p<0.01$), and time to complete the scenarios ($p<0.01$). We found clinical performance to degrade more in the cross-coverage condition ($p<0.01$), where participants needed to track and follow-up on patients with planned diagnostic evaluation. Participants also experienced more physiological workload ($p=0.04$) and took less time ($p<0.01$) to complete scenarios in the cross-coverage condition. The enhanced EMR environment with results tracking functionality indicated significant improvements in clinical performance ($p=0.03$) and physiological workload (blink rate: $p<0.01$). Overall, fatigue levels affected performance, especially for participants indicating low or high fatigue levels ($p<0.01$).

Research areas: Patient safety, abnormal test results, workload, performance.
Scientific disciplines: Cognitive and Behavioral Sciences; Human Factors.

1. Purpose

To objective of this study was to assess provider task demands, workload, and performance during follow-up on abnormal test results in an electronic medical record (EMR) environment under regular/cross-coverage and low/high-volume conditions (specific aim #1) and in a baseline/enhanced EMR environments (with specific aim #2).

2. Scope

There is evidence that cancer screening reduces morbidity and mortality [1-5]. Specifically, screening mammography and Papanicolaou (Pap) smears can improve patient outcomes in breast cancer and cervical cancer, respectively [1-5]. To achieve the full benefit of cancer screening, appropriate and timely follow-up of abnormal results must occur; however, such follow-up is frequently inadequate [6-10]. Evidence suggests that EHR alerts do not eliminate the problem of incomplete follow-up of abnormal test results [11-16]. The problem is rather a multifaceted safety issue that occurs within the complex "sociotechnical" system of healthcare, involving complex interactions between providers, patients, EMRs, workflows, and organizational factors [17-21]. Our focus was on follow-up of abnormal mammography and Pap smear results, and the baseline and enhanced EMR environment used for the study was Epic®.

Invitations to participate in the research study were sent to all residents and fellows in the school of medicine at one large academic institution, while clearly stating the need for experience with Epic as related to our simulated scenarios. All participants were incentivized to participate with a $100 gift card. Final selections were made based on participants' ability to conduct simulated scenarios (e.g., familiarity with Epic and managing relevant abnormal results) and availability to participate in the study during designated weeks for data collection.

All studies were conducted in our laboratory, 300 square feet dedicated room simulation-based assessments. The laboratory is divided into 2 sections; researcher station and participant station. The research and subject stations are separated via a one-way see through glass and the communication to the subject are made using a two-way microphone. Participant station is equipped with a workstation that closely emulate real clinical environment. Participant's workstation is adjustable includes configurable computer monitor, keyboard and computer mouse (exactly what the subjects use in the real clinic), thus increasing the fidelity of our study. The researcher's workstation allows recording and analyzing the data from the experiments in 'real-time'.

3. Methods

3.1 Specific aim #1: Effect of coverage (regular vs. cross) and volume (low vs. high) of abnormal test results on providers' experienced task demands, workload, and performance

3.1.1 Study participants

Total of 15 residents from the school of medicine at one large academic institution participated in this study, all with sufficient experience with Epic as related to our simulated scenarios (see Table 1 for details).

Specialty	# of Participants	Post Graduate Year (PGY) PGY: count	Gender F: Female M: Male
Internal Medicine	8	1:3 2:1 3:3 4:1	F:4 M:4
Family Medicine	4	1:1 2:1 3:1 4:1	F:2 M:2
Pediatrics	3	2:1 2:2	F:3
Total	**15**	**1:4 2:3 3:6 4:2**	**F:7 M:8**

Table 1. Composition of participants within each experiment.

3.1.2 Data collection

Participants were randomized to a regular- versus cross-coverage condition. Each subject had two sessions. In both sessions participants were asked to recognize and act upon 'abnormal' test results. In second session, to emulate cross-

coverage condition, in addition to recognition and acting upon abnormal test results, participants needed to track and follow-up on patients with planned diagnostic evaluation. The number of abnormal test results were double in the cross- vs. regular-coverage condition, while decreasing the total number of normal results in order to keep the total count of results per condition and session the same (see Table 2). Our total quantity of results used was in line with previous findings that noted providers on average managing approximately 57 abnormal test results per interaction with EMR [18-20]. Given that residents usually manage lower number of patients, our overall total of 35 results was appropriate for the study.

Table 2. Counts of normal vs. abnormal test results per each study condition.

Condition	**Session** (low vs. high volume)	**# of 'normal' results**	**# of 'abnormal' results**	**# of patients with follow-up needed** ('no-show' patients)	**Total results**	**# of partici-pants**
Regular-coverage	Low-volume	27	4	0	35	8
Cross-coverage	Low-volume	27	4	4	35	8
Regular-coverage	High-volume	19	8	0	35	7
Cross-coverage	High-volume	19	5^	8	32	7

^ *3 results did not appear correctly in our Epic playground environment; thus, eliminated were eliminated from the experiment.*

3.1.2.1 Experienced task demands

Task demands were quantified using the total number of clicks that participants needed to complete each session, and further to provide descriptive statistics we sub-categorized it into: i) navigation clicks (e.g., moving from one window to another window on the screen, etc.), ii) decision clicks (e.g., accepting/cancelling a test or medication, etc.), iii) search clicks (e.g., initiating the search option for medications/orders/etc.), and iv) total clicks (sum of navigation, decision, and search clicks).

3.1.2.2 Quantification of perceived workload

The NASA-Task Load Index (NASA-TLX), a widely applied and valid tool, was used to measure perceived workload [22-26].

3.1.2.3 Quantification of physiological workload

Physiological workload was quantified using validated measures and methodologies using data generated from eye tracking and electroencephalography (EEG) equipment collected throughout the simulated sessions.

3.1.2.3.1 Eye tracking

Pupillary measurements were collected using Tobii X2-60, 60Hz remote eye tracker. The raw data was processed based on the validated procedures recommended by experts [27-28]. Three baseline procedures were used to calculate task evoked pupillary response (TEPR) during clinical scenarios: (1) resting pupillary measure calculated as an average pupillary measure during standard blank white screen, (2) average pupillary measure calculated from basic math multiplication task involving 10 trials of multiplications of one-digit by one- or two-digit problems (presented 1-digit/second followed by 8-second calculation time and 10-second response time [typing the answer in the space provided]), and (3) average pupillary measure during reading instructions (while presented on the computer screen). TEPR was then computed as the change in pupillary dilations during clinical scenario from the baselines. Blink frequency was computed by dividing the total number of blinks during the task by the task time [29]. Gaze speed was computed as the distance (measured in degrees) of gaze travelled

between two consecutive samples (data collected at 60 samples per second) divided by the inter-sample time [30].

3.1.2.3.2 EEG

Data collection was done using the X-10 wireless EEG headset system from Advanced Brain Monitoring (ABM). The ABM system including multiple bi-polar sensor sites: Fz, F3, F4, Cz, C3, C4, POz, P3, and P4. In general, ABM software filters the EEG signals with a band-pass filter (0.5Hz-65Hz) before the analog-to-digital conversion. In order to remove environmental artifacts from the power network, sharp notch filters at 50Hz, 60Hz, 100Hz, and 120Hz are applied. The algorithm automatically detects and removes a number of artifacts in the time-domain EEG signal, including spikes caused by tapping or bumping of the sensors, amplifier saturation, and excursions that occur during the onset or recovery of saturations. In addition to the conventional data analysis methodology [difference between power of theta (6-7 Hz) at Fz vs. power of alpha signal at Pz (8-10 Hz) [31], the ABM's algorithm automatically calculated the index of cognitive workload using quadratic and linear discriminant function analyses of model-selected EEG variables derived from the power spectral analysis of the 1-Hz bins from 1-40Hz [32-33], which is different from the conventional analysis. It has been shown that ABM's workload index increases with working memory load, increasing difficulty of cognitive tasks (e.g., arithmetic, problem-solving), and has been validated in variety of 'simple' and 'complex' environments including military, industrial, and educational simulation environments [34-37].

3.1.2.4 Quantification of performance

To standardize our clinical-related performance score, for each participant we summed the counts of i) unacknowledged abnormal test results (identified by a failure to order a referral or additional testing), and ii) unacknowledged patients with abnormal test results that did not 'show up' for their scheduled appointments (identified by not following up with the patient), and divided it by the total number of abnormal test results presented in the corresponding session. Thus, our performance score represented performance degradation and ranged from 0 to 1, with 1 indicating optimal performance. We also quantified the total amount of time that participants took to complete each session.

Fatigue can impact participants' performance [38-41]. Therefore, we asked each participant to evaluate their own state of fatigue right before the experiments using the Crew Status Survey with following assessment scale with following prompt: 1. Fully Alert; Wide Awake; Extremely Peppy; 2. Very Lively; Responsive, But Not at Peak; 3. Okay; Somewhat Fresh; 4. A Little Tired; Less Than Fresh; 5. Moderately Tired; Let Down; 6. Extremely Tired; Very Difficult to Concentrate; 7. Completely Exhausted; Unable to Function Effectively; Ready to Drop. The Crew Status Survey has been previously tested in the 'real' and 'simulated' environments; and has been found to be both reliable and able to discriminate between fatigue levels [42]. For each session, we also recorded the time of day of the experiment.

3.1.3 Data analysis

Before data analyses, we completed tests for normality and equal variance for all study variables using Shapiro-Wilk's and Bartlett test respectively. Results indicated that assumptions were satisfied (normality: all $p>0.05$; equal variance: all $p>0.05$). To assess provider task demands, perceived and physiological workload, fatigue, and performance under regular- vs. cross-coverage condition and under low- vs. high-volume conditions, we conducted multivariable analysis of variance, while treating participants as a random factor. All our data analyses were conducted using JMP 10 software with significance level set at 0.05.

3.1.4 Results

Descriptive statistics of task demands, workloads, and performance for each condition are provided in Table 3.

3.1.4.1 Effect of regular/cross-coverage and low/high-volume conditions on providers' experienced task demands

Analysis of task demand data indicated that the high-volume of abnormal test results condition generated significantly more total clicks when compared to the low-volume of abnormal test results condition ($F(1,29)=29.95$, $p<0.01$; see Table 3).

3.1.4.2 Effect of regular/cross-coverage and low/high-volume conditions on providers' perceived workload (NASA-TLX)

Analysis of NASA-TLX data indicated that the high-volume of abnormal test results generated significantly higher perceived workload when compared to the low-volume of abnormal test results condition ($F(1,29)=14.96$, $p=<0.01$; see Table 3).

Table 3. Descriptive (mean (sd)) statistics of experienced task demands, perceived and physiological workload, and performance measures. Significant results are noted with a † symbol (see notes under the table for details).

Regular-coverage (low-volume)	Regular-coverage (high-volume)	Quantification of Task Demands (clicks), Workloads, Fatigue, and Performance	Cross-coverage (low-volume)	Cross-coverage (high-volume)
		Task Demand		
347 (68)	555 (87)	**Total Clicks** (count) †	407 (92)	458 (100)
174 (55)	282 (62)	**Navigation Clicks** (count)	238 (72)	253 (73)
120 (21)	192 (35)	**Decision Clicks** (count)	123 (24)	144 (23)
44 (14)	80 (11)	**Search Clicks** (count)	46 (19)	61 (16)
		Perceived Workload		
43 (14)	57 (11)	**NASA-TLX** (0=low to 100=high) †	48 (16)	61 (13)
		Psychological Workload		
0.02 (0.3)	0.2 (0.3)	**TEPR** (mm)	-0.03 (.2)	0.08 (.4)
16 (9)	15 (12)	**Blink Rate** (blinks/minute)	15 (6)	14 (9)
183 (103)	201 (159)	**Gaze Speed** (degrees/min)	180 (186)	209 (80)
0.6 (0.1)	0.5 (0.1)	**ABM Metric** (0=low to 1=high)	0.6 (0.1)	0.5 (0.2)
0.6 (0.3)	0.5 (0.6)	**Power of Fz (6-7 Hz) - Pz (8-10 Hz)** (μV2)†	0.8 (0.2)	0.7 (0.6)
		Performance		
		Clinical Performance		
1 (0)	0.9 (0.26)	(1=perfect performance) †	0.84 (.17)	0.82 (.16)
0	1	(counts of missed new abnormal results)	0	1
0	0	(counts of missed to follow-up on 'no-shows')	11	15
		Fatigue		
3 (0.9)	2.4 (1.7)	**Fatigue** (1=fully alert; 7=completely exhausted)	2.4 (1.0)	2.6 (1.4)
28:38 (8:24)	47:09 (6:07)	**Time to Scenario Completion** (min:sec) †	26:10 (8:34)	31:23 (4:25)

†*indicates significant results (p<.05).*

3.1.4.3 Effect of regular/cross-coverage and low/high-volume conditions on providers' physiological workload

Analysis of EEG data indicated the cross-coverage condition, where participants needed to track and follow-up on patients with planned diagnostic evaluation, generated significantly higher psychological workload when compared to the regular-coverage condition ($F(1,25)=3.4$, $p=0.04$; see Table 3; as quantified by normalized difference between power of theta (6-7 Hz) at Fz and power of alpha signal at Pz (8-10 Hz)). EEG data from two participants were not included in the analysis due to high signal-to-noise ratio.

3.1.4.4 Effect of regular/cross-coverage and low/high-volume conditions on providers' performance

Analysis of clinical performance data indicated that the cross-coverage condition, where participants needed to track and follow-up on patients with planned diagnostic evaluation, generated significantly more performance degradation when compared to the regular-coverage condition ($F(1,29)=9.2$, $p<0.01$; see Table 3). Overall, there was a significant effect of fatigue on clinical performance, especially for participants indicating low (=1) or high (5 and 6) fatigue levels ($F(5,29)=13.54$, $p<0.01$; see Table 3; on 6-point Likert scale). For time to complete scenarios, analysis indicated significant longer time to complete scenarios in the high-volume of abnormal test results condition when compared to the low-volume of abnormal test results condition ($F(1,29)=11.74$, $p<0.01$; See Table 3); and in the regular-coverage condition when compared to the cross-coverage condition (($F(1,29)=34.73$, $p<0.01$; see Table 3).

3.1.5 Discussion

The results indicate that the high-volume of abnormal test results significantly increased task demands as quantified by computer mouse clicks, providers' perception of workload as quantified by the NASA-TLX, and time to complete the simulated scenarios, but did not affect physiological workload or clinical performance. We found clinical performance to degrade more in the cross-coverage condition, where participants needed to track and follow-up on patients with planned diagnostic evaluation. Participants also experienced more physiological workload and took less time (possibly 'rushed' by skipping key information retrieval and analysis steps) to complete simulated scenarios in the cross-coverage condition.

The overall rate of failure to appropriately acknowledge abnormal test results was ≈0.5% (2 out 344 total failure opportunities; see Table 3 for breakdown by study condition). This result seems to be lower than findings from real clinical settings that indicated that providers fail to acknowledge abnormal test results in 4% of cases [15].

The overall rate of failure to appropriately follow up on patients with 'no-show' status was ≈30% (26 out of 88 total failure opportunities; see Table 3 for breakdown by study condition). This rate seems to be higher from findings from real clinical settings that indicated that tracking of patient status and timely follow-up was lacking in approximately 7.3% of acknowledged and 9.7% unacknowledged alerts; with the difference most likely being caused by the time allowed for providers to acknowledge alerted abnormal test results (usually 15 days in real clinical settings vs. 1 opportunity in our study) [16].

Results also indicate that clinical performance was affected by fatigue, with no fatigue and relatively high fatigue levels leading to most clinical performance degradation. Specifically, our data revealed that most participants that indicated no fatigue levels came to our laboratory in the morning (before their work shifts) and most participants that indicated high levels of fatigue came to our laboratory in the late afternoons (after their work shifts). This suggests that being either not 'activated' enough (just getting started on the job; or rushing to get to work) or mentally 'tired' after a long day at work can both negatively affect clinical performance. Interestingly, Hysong at colleagues found that 55.5% and 60%

of providers that managed their alerts first thing in the morning and at the end of the day respectively indicated suboptimal performance on timeliness of follow-up on abnormal test results [18].

Overall, these results suggest that challenges exist in ensuring appropriate follow-up of abnormal test results. This underscores the need for longitudinal monitoring and tracking systems within EMRs to ensure appropriate tracking and follow-up of abnormal test results, especially under cross-coverage conditions. In fact, scholars found that 55.5% of providers believe that the EMR systems do not have convenient features for tracking and follow-up on test results; 54.3% do not receive adequate training on system functionality; and 85.6% stay after hours or come in on weekends to address notifications [43]. Poon et al found the most highly desired features of a test result management system were tools to help physicians prioritize their workflows, track test orders to completion, and generate result letters to patients [20].

3.2 Specific aim #2: Effect of EMR environment (baseline vs. enhanced) and volume (low vs. high) of abnormal test results on providers' experienced task demands, workload, and performance

3.2.1 Study participants

Total of 38 residents from the school of medicine at one large academic institution participated in this study, all with sufficient experience with Epic as related to our simulated scenarios (see Table 4 for details).

Specialty	# of Participants	Post Graduate Year (PGY) PGY: count	Gender F: female; M: Male
Internal Medicine	14	1:4 2:2 3:5 4:3	F:9 M:5
Family Medicine	4	1:1 2:1 3:1 4:1	F:2 M:2
Pediatrics	9	1:3 2:2 3:4 4:0	F:7 M:2
Surgery (general, neuro, ortho, head & neck)	5	1:1 2:2 3:0 4:1 5:1	F:3 M:2
Other (cardiology, psychiatry, critical care, ob/gyn)	6	1:1 2:1 3:1 4:2 5:1	F:3 M:3
Total	**38**	**1:10 2:08 3:11 4:06 5:03**	**F:24 M:14**

Table 4. Composition of participants within each experiment.

3.2.2 Data Collection

Each subject had two simulated sessions. The first session was conducted in the baseline EMR environment to allow participants to familiarize themselves with our experimental conditions (e.g., laboratory environment, Epic playground) and practice the simulated scenarios. The second session was treated as the assessment session, where participants were randomized to baseline (EMR without longitudinal monitoring and tracking) vs. enhanced EMR environment (with longitudinal monitoring and tracking), and low versus high volume of abnormal test results. In both sessions participants were asked to recognize and act upon abnormal test results. In assessment session, to emulate cross-coverage condition, in addition to recognition and acting upon abnormal test results, participants needed to track and follow-up on patients with planned diagnostic evaluation. The number of abnormal test results were double in the high versus low volume condition, while decreasing the total number of normal results in order to keep the total count of results per condition and session the same (see Table 2 for details).

3.2.2.1 Experienced task demands: Same as 3.1.2.1

3.2.2.2 Quantification of perceived workload: Same as 3.1.2.2

3.2.2.3 Quantification of physiological workload: Same as 3.1.2.3.

3.2.2.3.1 Eye tracking: Same as 3.1.2.3.1

3.2.2.3.2 EEG: Same as 3.1.2.3.2

3.2.2.4 Quantification of performance: Same as 3.1.2.4

3.2.3 Data analysis: Before data analyses, we completed tests for normality and equal variance for all study variables using Shapiro-Wilk's and Bartlett test respectively. Results indicated that assumptions were satisfied (normality: all $p>0.05$; equal variance: all $p>0.05$). To assess provider task demands, perceived and physiological workload, fatigue, and performance in baseline- vs. enhanced-EMR environment and under low- vs. high-volume conditions, multivariable analysis of variance, while treating participants as a random factor. All our data analyses were conducted using JMP 13 software with significance level set at 0.05.

3.2.4 Results

Descriptive statistics of task demands, workloads, and performance for each condition are provided in Table 5.

Table 5. Descriptive (mean (sd)) statistics of experienced task demands, perceived and physiological workload, and performance measures. Significant results are noted with a † symbol (see notes under the table for details).

Current-EMR (Low-volume)	Current-EMR (High-volume)	Quantification of Task Demands (clicks), Workloads, Fatigue, and Performance	Enhanced -EMR (Low-volume)	Enhanced-EMR (High-volume)
		Task Demand		
390.8 (91.8)	496.0 (110.7)	**Total Clicks** (count) †	396.8 (83.2)	479.5 (118.5)
223.4 (73.4)	276 (76.9)	**Navigation Clicks** (count)	239.7 (75.2)	286.3 (78.1)
120.8 (22.7)	155.9 (29.5)	**Decision Clicks** (count)	106 (25.6)	124 (47.8)
46.6 (17.2)	63.9 (14.9)	**Search Clicks** (count)	51 (18.9)	69 (24.8)
		Perceived Workload		
48.1 (15.5)	58.8 (13.7)	**NASA-TLX** (0=low to 100=high)	49.3 (18.3)	49.3 (13.7)
		Psychological Workload		
0.04 (0.2)	0.1 (0.3)	**TEPR** (mm)	0.02 (0.2)	-0.03 (0.2)
15.1 (9.6)	17.4 (7.5)	**Blink Rate** (blinks/minute)	24.7 (10.4)	22.6 (6.1)
216 (114.6)	192 (113.3)	**Gaze Speed** (degrees/min)	170 (34.7)	157 (32.7)
0.5 (0.1)	0.6 (0.1)	**ABM Metric** (0=low to 1=high)	0.6 (0.1)	0.6 (0.1)
0.7 (0.2)	0.8 (0.5)	**Power of Fz (6-7 Hz) - Pz (8-10 Hz)** (μV2)2	0.6 (0.9)	0.8 (0.8)
		Performance		
0.6 (0.4)	0.8 (0.2)	**Clinical Performance** (1=perfect performance) †	1.0 (0.08)	0.8 (0.2)
2	6	(counts of missed new abnormal results)	0	1
15	17	(counts of missed to follow-up on 'no-shows')	2	4
		Fatigue		
2.7 (1.5)	2.7 (1.5)	**Fatigue** (1=fully Alert; 7=completely exhausted) †	3.0 (1)	2.5 (0.9)
26:12 (7:48)	37:18 (10:24)	**Time to Scenario Completion** (min:sec) †	28:54 (6:12)	34:12 (12.06)

†*indicates significant results ($p<.01$).*

3.2.4.1 Effect of baseline/enhanced-EMR and low/high-volume conditions on providers' experienced task demands

Analysis of task demand data indicated that the high-volume of abnormal test results generated significantly more total clicks when compared to the low-volume of abnormal test results condition ($F(1,35)= 8$, $p<0.01$; see Table 5).

3.1.4.2 Effect of baseline/enhanced-EMR and low/high-volume conditions on providers' perceived workload (NASA-TLX)

Analysis of NASA-TLX scores indicated no significant differences ($p>0.05$).

3.1.4.3 Effect of baseline/enhanced-EMR and low/high-volume conditions on providers' physiological workload

Analysis of eye-based measures indicated that blink rate was significantly less in the baseline-EMR environment ($F(1,36)=7$, $p=.01$; see Table 5), suggesting that mental effort was higher in the baseline-EMR environment compared to the enhanced-EMR environment. Analysis of EEG data indicated no significant differences ($p>0.05$).

3.1.4.4 Effect of baseline/enhanced-EMR and low/high-volume conditions on providers' performance

Analysis of clinical performance indicated a significant improvement in performance in the enhanced-EMR environment when compared to the baseline-EMR environment ($F(1,34)=5$, $p=.03$; see Table 5). Overall, there was a significant effect of fatigue on clinical performance, especially for participants indicating low (=1) or high (5 and 6) fatigue levels ($F(5,38)=4.4$, $p<0.01$; see Table 5; 6-point Likert scale). For time to complete scenarios, analysis indicated that participants took significant longer time to complete scenarios in the high-volume of abnormal test results condition when compared to the low-volume of abnormal test result condition ($F(1,35)=7$, $p<0.01$; see Table 5).

3.2.5 Discussion

The results indicate that the high-volume of abnormal test results significantly increased task demands as quantified by computer mouse clicks, and time to complete the simulated scenarios, but did not affect perceived or physiological workload and clinical performance. We found clinical performance to improve in enhanced-EMR environment, with longitudinal monitoring and tracking functionality for patients with planned diagnostic evaluation.

The overall rate of failure to appropriately acknowledge abnormal test results was ≈2.2% (9 out 399 total failure opportunities; see Table 5 for breakdown by study condition) with 1 occurring in the enhanced EMR-environment (0.5% vs. 3.8% in the baseline-EMR environment vs. 4% from the real clinical settings [15]).

The overall rate of failure to appropriately follow up on patients with a no-show status was ≈16% (38 out of 228 total failure opportunities; see Table 3 for breakdown by study condition), with 6 occurring in the enhanced EMR-environment (5.5% vs. 26.6% in the baseline-EMR environment; vs. 7.3% of acknowledged and 9.7% unacknowledged alerts from real clinical settings [16]).

Results indicate that clinical performance was affected by fatigue, with no fatigue and relatively high fatigue levels leading to most clinical performance degradation.

Overall, these results suggest that longitudinal monitoring and tracking of abnormal test results and patient status can help ensuring appropriate follow-up of abnormal test results. At the same time, this study illustrates that such functionality *alone* does not eliminate all the challenges with acknowledgment and follow-up of abnormal test results. Additional research is needed to quantify effects of innovative

functionality and usability features for ensuring appropriate acknowledgment and follow-up of abnormal test results. There are also opportunities for improvement in policies and procedures, education and training, and audit and performance systems, though these alone are less likely to be effective in addressing the identified challenges, yet might provide the necessary synergy to help protect abnormal test results from 'falling through the cracks' and resulting in patient harm.

While the next generations of EMR systems are being designed there are ample opportunities for this improvement. We propose several potential interventions based on our findings that can be used immediately to improve proper acknowledgment and timely follow-up of abnormal test results. First, there is a need to properly design features of the EMR to focus providers attention on i) abnormal test results, and ii) patients' status (e.g., no-show status), both with enough detail to facilitate (or not facilitate) appropriate follow-up communication. Second, every institution could develop and publicize policies and guidelines regarding work demands (e.g., number of patients/results for review per day, per interaction, per time of the day, per cross-coverage) to ensure appropriate levels of workload and performance, while minimizing negative effects of fatigue on performance. Third, while often less effective, innovative education/training requirements (e.g., simulation based training vs. traditional training) and performance feedback systems for providers on EMRs could be organized and implemented [47-50]).

4. Limitations

There are several limitations to both studies, and thus caution should be exercised in generalizing our findings. First, the results are based on one experiment with relatively small number of participants from one teaching hospital, performed on set of scenarios. Larger studies, while controlling for levels of fatigue, PGY, specialty, training levels could allow for more accurate regression of such factors on the variables of interest (e.g., task demands, workloads, performance). Second, the time between simulated sessions varied from 1 to 3 weeks, which could have unexpectedly bias the study due to some carryover effects between sessions (e.g., learning effect, perceptions of workload). In addition, the day and time of the day to conduct assessments varied, which could have also affected the results. While difficult, future studies could try to control for the amount of time between sessions as well as day and time of the day for assessments. Third, performing the scenarios in the simulated environment, where the subjects knew that their work was going to be assessed, may have affected participants' performance (e.g., more/less attentiveness and vigilance as perceived by being assessed or by possibility of real harm to the patient). Fourth, reporting workload via NASA-TLX is subjective and can be challenging for some participants. Future studies could explore using other instruments for workload quantification. Fifth, our quantification methods of physiological workload, while validated and broadly used, may not fully considered potential confounding factors streaming from cognitive information processing during providers interactions with EMRs, or general cognitive states (e.g., participants' arousal, anxiety, and stress [30-32, 51]). Thus, further research to develop and assess the utility of physiological measures of mental workload during providers' interactions with EHR is needed. Finally, since our laboratory setting was only a simulation of a real clinical environment; some behaviors of the studied scenarios were not easily emulated (e.g., looking up additional information about the patient in alternative software; calling a nurse with a question about particular patient; consultations regarding the use of Epic). Thus, all subjects were informed about the limitations of our laboratory environment before the experiments.

5. References

[1]. Andrae B, Andersson TM, Lambert PC, et al. Screening and cervical cancer cure: population based cohort study. BMJ 2012;344:e900.

[2]. Clarke EA, Anderson TW. Does screening by "Pap" smears help prevent cervical cancer? A case-control study. Lancet 1979;2:1-4.

[3]. Johannesson G, Geirsson G, Day N. The effect of mass screening in Iceland, 1965-74, on the incidence and mortality of cervical carcinoma. Int J Cancer. 1978;21:418-425.

[4]. Independent UK Panel on Brest Cancer Screening. The benefits and harms of breast cancer screening: an independent review. Lancet 2012;380:1778-1786.
[5]. Nelson HD, Tyne K, Naik A, et al. Screening for breast cancer: an update for the U.S. Preventive Services Task Force. Ann Inter Med 2009;151:727-737, W237-742.
[6]. Jones BA, Dailey A, Calvocoressi L, et al. Inadequate follow-up of abnormal screening mammograms: findings from the race differences in screening mammography process study (United States). Cancer Causes Control 2005;16:809-821.
[7]. McCarthy BD, Yood MU, Boohaker EA, et al. Inadequate follow-up of abnormal mammograms. Am J Prev Med 1996;12:282-288.
[8]. Peterson NB, Han J, Freund KM. Inadequate follow-up for abnormal Pap smears in an urban population. J Natl Med Assoc 2003;95:825-832.
[9]. Callen JL, Westbrook JI, Georgiou A, et al. Failure to follow-up test results for ambulatory patients: a systematic review. J Gen Inter Med 2012;27:1334-1348.
[10]. Yabroff KR, Washington KS, Leader A, et al. Is the promise of cancer-screening programs being compromised? Quality of follow-up care after abnormal screening results. Med Care Res Rev 2003;60:294-331.
[11]. Moore C, Saigh O, Trikha A, et al. Timely follow-up of abnormal outpatient test results: Perceived barriers and impact on patient safety. J Patient Saf 2008;4:241-244.
[12]. Kuperman GJ, Teich JM, Tanasijevic MJ, et al. Improving response to critical laboratory results with automation: results of a randomized controlled trial. JAMIA 1999;6:512-522.
[13]. Lin JJ, Moore C. Impact of an electronic health record on follow-up time for markedly elevated serum potassium results. Am J Med Qual 2011;26:308-314.
[14]. Laxmisan A, Sittig DF, Pietz K, et al. Effectiveness of an electronic health record-based intervention to improve follow-up of abnormal pathology results: a retrospective record analysis. Med Care 2012;50:898-904.
[15]. Singh H, Arora HS, Vij MS, et al. Communication outcomes of critical imaging results in a computerized notification system. JAMIA 2007;14:459-466.
[16]. Singh H, Thomas EJ, Mani S, et al. Timely follow-up of abnormal diagnostic imaging test results in an outpatient setting: are electronic medical records achieving their potential? Arch Inter Med 2009;169:1578-1586.
[17]. Zapka J, Taplin SH, Price RA, et al. Factors in quality care--the case of follow-up to abnormal cancer screening tests--problems in the steps and interfaces of care. J Natl Cancer Inst Monogr 2010;2010:58-71.
[18]. Hysong SJ, Sawhney MK, Wilson L, et al. Provider management strategies of abnormal test result alerts: a cognitive task analysis. JAMIA 2010;17:71-77.
[19]. Hysong SJ, Sawhney MK, Wilson L, et al. Understanding the management of electronic test result notifications in the outpatient setting. BMC Med Inform Decis Mak 2011;11:22.
[20]. Poon EG, Gandhi TK, Sequist TD, et al. "I wish I had seen this test result earlier!": Dissatisfaction with test result management systems in primary care. Arch Intern Med 2004;164:2223-2228.
[21]. Mazur LM, Mosaly P, Moore C, et al. Towards a better understanding of workload and performance during physician-computer interactions. JAMIA 2016;23:1113-1120.
[22]. Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: P. A. Hancock and N. Meshkati (Eds.). Human mental workload. Amsterdam: North Holland Press. 1988;139-183.
[23]. Young G, Zavelina L, Hooper V. Assessment of workload using NASA Task Load Index in perianesthesia nursing. J Perianesth Nurs 2008;3:102-110.
[24]. Yurko YY, Scerbo MW, Prabhu AS, et al. Higher mental workload is associated with poorer laparoscopic performance as measured by the NASA-TLX Tool. Simul in Healthc 2010;5:267-271.
[25]. Mazur LM, Mosaly P, Jackson M, et al. Quantitative assessment of workload and stressors in clinical radiation oncology. Int J Radiat Oncol Biol Phys 2012;83:e571-e576.

[26]. Mosaly P, Mazur LM, Jones E, et al. Quantification of physician's workload and performance during cross-coverage in radiation therapy treatment planning. Prac Radiat Oncol 2013;3:e179-e186.
[27]. Beatty J, Lucero-Wagoner B. The pupillary system. Handbook of psychophysiology. 2000:142-162,
[28]. Mosaly P, Mazur LM, Fei Y, et al. Relating task demand, mental effort and task difficulty with physicians' performance during interactions with electronic health records (EHRs). Inter J Hum-Comput Int, 2017 (accepted).
[29]. Poole A, Ball LJ. Eye tracking in HCI and usability research. Encyclopedia of human computer interaction, 2006;1:211-219.
[30]. Holmqvist K, Nyström M, Andersson R, et al. Eye tracking: A comprehensive guide to methods and measures. OUP Oxford, 2011.
[31]. Smith ME, Gevins A, Leong H, et al. Monitoring task loading with multivariate EEG measures during complex forms of human-computer interactions. Hum Factors 2001;43:366-380.
[32]. Berka C, Levendowski DJ, Cvetinovic MM, et al. Real-time analysis of EEG indices of alertness, cognition, and memory with a wireless EEG headset. Inter J Hum-Comput Int 2004:17:151-170.
[33]. Berka C, Levendowski DJ, Lumicao MN. EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. Aviat Space Environmental Medicine 2007;78(5, Suppl.):B231-B244.
[34]. Tremoulet P, Barton J, Craven R, et al. Augmented cognition for tactical tomahawk weapons control system operators. In Schmorrow D, Stanney K, Reeves L, eds. Foundations of Augmented Cognition. Arlington, VA: Strategic Analysis. 2006:313-318.
[35]. Berka C, Levendowski D, Ramsey CK, et al. Evaluation of an EEG-workload model in an Aegis simulation environment. In: Caldwell JA, Wesensten NJ, eds. Proceedings of SPIE Defense and Security Symposium, Biomonitoring for Physiological and Cognitive Performance during Military Operations. Orlando, FL: SPIE: The International Society for Optical Engineering 2005:90-9.
[36]. Berka C, Levendowski D, Westbrook P, et al. EEG Quantification of alertness: methods for early identification of individuals most susceptible to sleep deprivation. In: Caldwell JA, Wesensten NJ, eds. Proceedings of SPIE Defense and Security Symposium, Biomonitoring for Physiological and Cognitive Performance during Military Operations. Orlando, FL: SPIE: The International Society for Optical Engineering 2005:78-89.
[37]. Dorneich MC, Whitlow SD, Mathan S, et al. Supporting real-time cognitive state classification on a mobile participant. J Cogn Eng Decis Making 2007;1:240-270.
[38]. Needleman J, Buerhaus PI, Pankratz VS, et al. Nurse staffing and inpatient hospital mortality. N Eng J Med 2011;364:1037-1045.
[39]. Van den Hombergh P, Kunzi B, Elwyn G, et al. High workload and job stress are associated with lower practice performance in general practice: An observational study in 239 general practices in the Netherlands. BMC Health Serv Res 2009;9:118.
[40]. Weigl M, Müller A, Vincent C, et al. The association of workflow interruptions and hospital doctors' workload: a prospective observational study. BMJ Qual Saf 2012;21:399-407.
[41]. Miller JC, Narveaz AA. A comparison of the two subjective fatigue checklists. Proceedings of the 10th Psychology in the DoD symposium. Colorado Springs, CO; United States Air Force Academy, 514-518, 1986.
[42]. Gawron VJ. Human Performance, Workload, and Situational Awareness Measurement Handbook. Florida, CRC Press, 2008.
[43]. Singh H, Spitzmueller C, Petersen NJ, et al. Primary care practitioners' views on test result management in EHR-enabled health systems: a national survey. J Am Med Inform Assoc 2013;20:727-735.
[44]. Shanteau J, Stewart TR. Why study expert decision making? Some historical perspectives and comments. Organ Behav Hum Decis Process 1992;53:95-106.
[45]. Hollender, N, Hofmann, C, Deneke, M, et al. Integrating cognitive load theory and concepts of human–computer interaction. Comput Human Behav 2010;26:1278-1288.

[46]. Ariza, F, Kalra, D, Potts, HW. How Do Clinical Information Systems Affect the Cognitive Demands of General Practitioners?: Usability Study with a Focus on Cognitive Workload. J Innov Health Infor 2015;22:379-390.
[47]. Steadman RH, Coates WC, Huang YM, et al: Simulation-based-training is superior to problem-based learning for the acquisition of critical assessment and management skills. Crit Care Med 2006; 34:151Y157.
[48]. Mazur LM, Mosaly P, Chera B, et al. Development and Assessment of the Impact of Simulation-based Training on Radiation Oncology Providers' Workload and Performance. Prac Radiat Oncol 2017;7:e309-e316.
[49]. Austin JM, Demski R, Callender T, et al. From Board to Bedside: How the application of financial structures to safety and quality can drive accountability in a large health care system. The Joint Commission Journal on Quality and Patient Safety 2017;43(4):166-175.
[50]. Smith MW, Murphy DR, Laxmisan A, et al. Developing software to track and catch missed follow-up of abnormal test results in a complex sociotechnical environment. Appl Clin Inform In press 2013;4:359-375.
[51]. Mosaly P, Mazur LM, Marks L. Quantification of baseline pupillary response and task-evoked pupillary response during constant and incremental task load. *Ergonomics*. 2017;60:1369-1375.

6. List of Publications

We reported preliminary results in:

Mazur LM, Mosaly P, Falchook A, Eblan M, Hoyle, L, Moore C, Elnahal S, Herman J, Chera B, Marks LB. Towards a Better Understanding of Workload and Performance during Physician-Computer Interactions. *Journal of American Medical Informatics Association.* 23(6):1113-1120, 2016. doi.org/10.1093/jamia/ocw016

Summary: Two experiments were performed in two different electronic medical record (EMR) environments. Each provider (n=29) was instructed to complete set of pre-specified tasks on three routine clinical EMR-based scenarios. Task demands were quantified using behavioral responses (e.g., computer mouse clicks). Performance was quantified based on the maximum severity of omission errors. In both experiments the regression analysis indicated a significant relationship between task demands and performance ($p<0.01$).

Mosaly P, Mazur LM, Fei Y, Guo H, Merck D, Moore C, Marks L, Mostafa J. Relating Task Demand, Mental Effort and Task Difficulty with Physicians' Performance during Interactions with Electronic Health Records (EHRs). *International Journal of Human-Computer Interaction*. Accepted (June 29th, 2017).

Summary: Providers (n=17) performed 3 EMR-based scenarios with varying task demands. Mental workload was measured using eye tracking methods via task evoked pupillary responses (TEPR), blink frequency and gaze speed; task difficulty (or user behavior) was measured using frequent mouse click patterns and task flow; user performance was quantified using omission errors. Mental workload (blink rate) and task difficulty (click patterns) predicted performance ($p<0.01$).

Mosaly P, Mazur LM, Marks L. Quantification of Baseline Pupillary Response and Task-Evoked Pupillary Response During Constant and Incremental Task Load. *Ergonomics*. 60(10):1369-1375, 2017. doi.org/10.1080/00140139.2017.1288930 (methodological contribution).

Summary: The methods employed to quantify the baseline pupil size and task-evoked pupillary response(TEPR) may affect the overall study results. To test this hypothesis, the objective of this study was to assess variability in baseline pupil size and TEPR during two basic working memory tasks:

constant load of 3-letters memorisation-recall (10 trials), and incremental load memorisation-recall (two trials of each load level), using two commonly used methods (1) change from trail/load specific baseline, (2) change from constant baseline. Results indicated that there was a significant shift in baseline between the trails for constant load, and between the load levels for incremental load. The TEPR was independent of shifts in baseline using method 1 only for constant load, and method 2 only for higher levels of incremental load condition. These important findings suggest that the assessment of both the baseline and methods to quantify TEPR are critical in ergonomics application, especially in studies with small number of trials per subject per condition.

Mazur LM, Mosaly P, Moore C, Marks LB. Effect of Coverage and Volume of Abnormal Test Results on Providers' Experienced Task Demands, Workload, and Performance: Results of a Prospective Randomized Trial. To be submitted to *Applied Ergonomics* (summarizes our work from specific aim #1).

Summary: Providers (n=15) were randomized to regular- vs. cross-coverage condition and low- vs. high-volume of abnormal test results in EMR environment requiring their attention. Task demands were quantified using computer mouse clicks. Perceived workload was quantified using the NASA-Task Load Index (NASA-TLX). Physiological workload was quantified using validated measures and methodologies using data generated from eye tracking and electroencephalography (EEG) equipment collected throughout the simulated sessions. Clinical performance was quantified based on acknowledgment/management of patients with abnormal test results. We also measured performance via time-to-session completion. We also quantified participants' fatigue in order to assess its impact on performance. The high-volume of abnormal test results significantly increased task demands ($p<0.01$), providers' perception of workload ($p<0.01$), and time to complete the simulated scenarios ($p<0.01$). We found clinical performance to degrade more in the cross-coverage condition ($p<0.01$), where participants needed to track and follow-up on patients with planned diagnostic evaluation, especially for participants indicating low or high fatigue levels ($p<0.01$). Participants also experienced more physiological workload (EEG: $p=0.04$) and took less time ($p<0.01$) to complete simulated scenarios in the cross-coverage condition.

Mazur LM, Mosaly P, Moore C, Marks LB. Effect of Longitudinal Monitoring and Tracking of Abnormal Test Results on Providers' Experienced Task Demands, Workload, and Performance: Results of a Prospective Randomized Trial. To be submitted to the *Journal of American Medical Informatics Association (JAMIA).*

Summary: Providers (n=38) were randomized to low- vs. high-volume of abnormal test results in the baseline- vs. enhanced EMR with longitudinal monitoring and tracking of abnormal test results and asked to review and act upon patients' test results. Task demands were quantified using computer mouse clicks. Perceived workload was quantified using the NASA-Task Load Index (NASA-TLX). Physiological workload was quantified using validated eye tracking and electroencephalography (EEG) methods. Clinical performance was quantified based on acknowledgment/management of patients with abnormal test results. We also measured performance via time-to-session completion. We also quantified participants' fatigue in order to assess its impact on performance. The high-volume of abnormal test results significantly increased total clicks ($p<0.01$) and time to scenario complete ($p<0.01$). The enhanced EMR environment indicated significantly improvements in clinical performance ($p=0.03$), and physiological workflow (blink rate: $p<0.01$). Overall, fatigue levels affected performance, especially for participants indicating low or high fatigue levels ($p<0.01$).

7. List of Products

Based on our findings, we proposed following recommendations aimed at decreasing providers' burden due to current health IT usability issues. These recommendations were submitted per request of Steve Bernstein (Project Officer, AHRQ, CEPI, Division of Health IT) to HHS Office of National Coordinator (ONC).

- **Recommendations #1.** Health IT interaction-related data (e.g., task demands, workload) should be used as quality/safety metric that is likely representative of providers' performance, and perhaps patient outcomes. This could be operationalized using automated technology (e.g., tracking clicks, click patterns, time-to-task-completion, statistics related to EMR interactions) in the health IT applications. This data should be used to spearhead improvement efforts related to usability, workflows, and policies.

- **Recommendation #2:** All health IT vendors, in collaboration with other research groups, should be required to conduct studies toward the use of health IT as part of a learning health care system. Products of this research should be used to inform the design, testing, and use of health IT within the sociotechnical system. Specific areas of research could include:

    - User-centered design, contextual design, and human factors applied to health IT.
    - Safe implementation, use, and continuous improvement of health IT by all users.
    - Performance of sociotechnical systems associated with health IT.
    - Impact of policy decisions on health IT use in clinical practice.

- **Recommendation #3:** Innovative trainings (e.g., simulation-based training vs. traditional EMR trainings) programs for providers' interactions with health IT should be supported and encouraged by the HHS Office of National Coordinator (ONC), nationally recognized societies/associations and implemented via commercial health IT vendors as part of their education/training package. For example, on the local scale, simulation-based training programs could become part of the education curriculums for residents and students. For example, simulation-based training could be done in naturalistic (e.g., real clinic) settings with residents and students as targeted trainees, and clinical supervisors as mentors/coaches providing evaluations and constructive feedback with 'no blame' mindset on 'fake' plans with purposefully embedded errors into health IT. Such training could enhance providers' performance. This could also help providers to acquire new skills and knowledge to proactively maintain their preoccupation with patient safety during interactions with health IT.

- **Recommendation #4:** There is a need for funding support for further methodological and empirical research to advance our theoretical and practical understanding of the relationships between task demands, task difficulty, mental workload, and performance during providers' interactions with health IT.

Project-Generated Resources: We have developed testable 'fake' patients for Epic playground environment. This allows other researchers to replicate our study. We have uploaded these files to **AHRQ Research Reporting System** under "Attachments".