Metrics for Measuring Internet-based Telemedicine Quality: An Empirical Investigation for Ophthalmology

Bengisu Tulu
Department of Management
Worcester Polytechnic Institute
Worcester, MA, USA

Samir Chatterjee
School of Information Systems and Technology
Claremont Graduate University
Claremont, CA, USA

Abstract— **Our study focuses on the video component of multimedia applications with the goal of understanding the relationship between objective and subjective video quality measures and clinical decision making capability of the medical professionals on the receiving end, with Internet as the transmission channel. An actual telemedicine video sequence is used in this study. The quality of this sequence was degraded using network impairments and 14 degraded sequences plus the original sequence were presented to 15 subjects. Participants were asked to evaluate the quality of the sequences and their decision making capability based on the sequence under consideration. The results indicate that the quality score does not always correlate to the clinical decision making capability score. Same thing is also true for objective quality measures. Further research is needed to better understand the effects of video quality degradation on medical decision making in telemedicine environments.**

Keywords- Telemedicine, video quality measurement, eye exams

I. INTRODUCTION

Telemedicine applications rely on the telecommunication infrastructure which is often chosen carefully to support such applications. Regardless of the transmission technology used in a telemedicine program, there is one single requirement for real-time telemedicine multimedia applications, that is, to provide same quality before and after the transmission of packets over the telecommunication channel. This becomes incredibly hard when Internet is used as the delivery channel. Our study focuses on the video component of multimedia applications with the goal of understanding the relationship between objective and subjective video quality measures and clinical decision making capability of the medical professionals on the receiving end, with Internet as the transmission channel.

Video quality has been an important issue, first for television broadcasting applications. Various measures have been developed for analog video systems to evaluate the effects of transmission on the original video signal. However, today, digital video systems are replacing these analog systems and are becoming an essential part of the world economy [1].

Wolf and Pinson [1] states that, "To be accurate, digital video quality measurements must be based on the perceived quality of the actual video being received by the users of the digital video system rather than the measured quality of traditional video test signals (e.g., color bar)". Especially in telemedicine, perceived video quality plays a critical role in the medical professional's confidence level in decision making. We utilized the outcome of our previous work where a test video was transmitted over an emulated Internet and a large number of degraded video clips were obtained [2]. In the previous study, objective metrics such as Peak Signal to Noise Ratio (PSNR) were calculated for various degradations. In this study, these video clips were later evaluated by human subjects in terms of their perceived quality score and their perceived decision making capability score.

The outline of this paper is as follows. First, a survey of factors affecting video quality and video quality measurement techniques is presented. Second, the experimental design is presented followed by results and discussion. The paper concludes with summary of contributions, limitations, and future research directions.

II. BACKGROUND

The evolution and growth of telemedicine is highly correlated with the developments in communication technology and advances in software applications. Video and image quality are very critical for telemedicine applications and hence quality within the context of telemedicine requires special attention. The ultimate goal of this research is to address the following question:

> *"How can a medical practitioner conducting telemedicine over the Internet assign a metric to the quality of video received?"*

While generic video quality has been reported in other industries (e.g. Broadcast, videoconferencing), very few studies are available for telemedicine context. This section will provide a brief overview of factors affecting quality and measurement techniques developed to study it.

A. Quality Factors

Quite a lot of research has been done over the years on understanding the factors that affect quality within the video broadcasting applications and videoconferencing solutions for

businesses. There are various factors that can affect the video quality of telemedicine applications. At the network level, lack of guarantee in terms of bandwidth, packet loss, delay, and jitter, are some of the challenges that can impact the quality of multimedia content delivered over the Internet as reported in various studies [2-5].

Video transmission is a resource and bandwidth intensive application type [3] that requires the video to be compressed before transmission to utilize the existing resources efficiently without saturating them. The goal of video compression is to remove the redundancy in the original source signal, which will eventually reduce the amount of bandwidth required for transmission [6]. Compression can be done using lossless coding, lossy coding, or a combination of both. At the application level, depending on the coding scheme video quality will vary.

Quality of service (QoS) solution was developed for real-time applications in order to overcome the impairments that occur at the network and application levels. Application-level QoS provides quality improvements without requiring changes of the network infrastructure. For example, adaptive playout techniques were introduced to make real-time applications more tolerant of delays and delay jitter and to dynamically adjust the playback point [4]. Reconstruction methods that compensate for packet loss in real-time applications at the receiver side were also studied. These are solutions that usually are integrated in the application used for transmitting video over the Internet and hence the choice of application has an impact on the quality.

Development of network Quality of Service (QoS) features was partially motivated by the fact that real-time traffic (as well as other applications) may sometimes require priority treatment to achieve good performance on the Internet [7]. QoS can be achieved by managing router queues and by routing traffic around congested parts of the network. Integrated Services (Int-Serv) [8], Differentiated Services (Diff-Serv) [9], and Multiprotocol Label Switching (MPLS) architecture [7] are commonly used network QoS techniques by service providers. In spite of technical advances, such QoS is not available in many portions of the Internet. For rural region and poorer emerging economies, QoS is cost prohibitive for telemedicine.

B. Quality Measurement

Quality measurement can be done either objectively (using electrical measurements) or subjectively (using human viewers) [10]. Below is a brief overview of some of the techniques used for both of these measurements.

1) Objective Measurement: Peak Signal-to-Noise Ratio (PSNR) is the most commonly used metric for measuring video and image quality. It measures how close a sequence is compared to the original one [6]. The PSNR is usually reported in decibels (dB) [3]. An image with a PSNR of 25 dB or below is usually unacceptable. Between 25 dB and 30 dB, perceived quality usually improves and above 30 dB, images are often perceived as good as the original image.

There are other standard and proprietary measurement techniques that have been developed and reported in the literature that are not mentioned here. One commonality between these objective measures, however, is that they require access to both original and processed video sequences. One recent study [6] proposed a new measure, which does not require access to the original video sequence. In this new method, artificial neural networks (ANN) are used to predict perceived voice and video quality using a trained engine based on previous objective and subjective tests.

One other commonly used metric is the "Video Quality Metric (VQM)" [1], which was developed by the Institute for Telecommunication Sciences (ITS). It requires the extraction and classification of features from both the original and processed video sequences similar to the other measurement techniques. Once these features are extracted, the distance between the original and processed video sequences is computed based on these features; and later this distance is mapped to a subjective score [1]. Compared to the PSNR, this metric offers different models for various transmission types, such as videoconferencing or TV models. It is also possible to identify the nature of an impairment using the VQM, which the PSNR does not provide [4].

2) Subjective Measurement: The ITU-R 500 is the standard for subjective assessment of image quality and has evolved over the years to include measures for digital video transmissions as well. This standard provides scales for single and double stimulus methods. The Absolute Category Rating (ACR) is a single stimulus method where test sequences are presented one at a time and are rated on a category scale after they are viewed. Usually a 5-point category scale is used as illustrated in Fig.1.

5-point Quality Scale		**5-point Impairment Scale**	
Estimated Quality	*Score*	*Estimated Impairment Level*	*Score*
Excellent	5	Imperceptible	5
Good	4	Perceptible	4
Fair	3	Slightly Annoying	3
Poor	2	Annoying	2
Bad	1	Very Annoying	1

Figure 1. ITU Video Quality Assessment Scales

The Single Stimulus Continuous Quality Evaluation (SSCQE) is different from the ACR in terms of the scale it uses and the assessment process. Among the double stimulus methods, the Double Stimulus Impairment Scale (DSIS) – also known as the Degradation Category Rating (DCR) - presents pairs of original and impaired video sequences during the test respectively. In this case, subjects are asked to rate the impairment of the second stimulus with respect to the reference (first stimulus) using the 5-point impairment scale.

In the Double Stimulus Continuous Quality Scale (DSCQS) method, the sequences are presented in pairs like in the DSIS and subjects are asked to evaluate the quality of both sequences. The original sequence is included for reference; however, the observers are not told which one is the reference sequence and the order of appearance changes for each test. The scale used in this method is illustrated in Fig. 2. There are other methods where the two sequences are shown

simultaneously and the observers are asked to make a comparison of the two based on stimulus comparison scale.

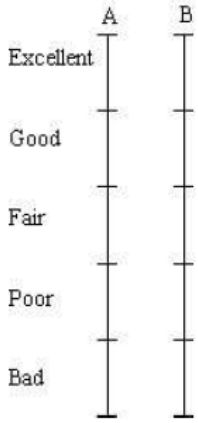

Figure 2. Continuous 5-point Quality Scale for DSCQS

III. EXPERIMENTAL DESIGN

The low cost and ubiquity of the Internet makes it an appealing communication medium for telemedicine services. However, the unreliable connection properties of packet-based systems and their vulnerability to various impairments that affect the physical, network, and application layers hamper the quality of Internet-based telemedicine applications. This research is aimed to better understand the video quality measures and their thresholds for Internet based telemedicine services, specific to general eye examinations. In order to achieve this, an experimental study is planned. This section describes the details of the experimental design.

A. Testbed and Initial Data Collection

An experimental testbed was prepared to emulate the Internet traffic and control the impairment parameters (delay, jitter, and drop) while transmitting an actual telemedicine video obtained from the Regenstrief Institute for Health Care in mpeg video format. This video file is the recording of a general eye examination conducted by a remotely located physician. In a previous study, the original video sequence used in the testbed experiments was generated from the high quality MPEG1 file by converting it into a shorter AVI file. The resulting original sequence was 14 seconds long. The degraded test sequences were generated from this original file by transferring them on the experimental testbed where the selected parameters are manipulated. Each transfer generated a different video sequence. Video quality of impaired sequences is objectively measured using VQM software tool. The details of the experimental design and the objective quality measurement results have been recently reported [2]. The findings of the previous study indicated that the jitter has significant effect on quality compared to packet delay and packet drop.

B. Subjective Measurement Tests

This study builds on the previous one by taking the next step to measure the subjective quality. Subjective quality experiments involve human subjects.

1) Sample Selection: Subjects for this study were selected from different user groups. Since domain expertise is required for making a judgment on the conditions for a medical decision after viewing the eye examination video, 8 optometrists were recruited from public and private optometry clinics. Subjects with no domain knowledge were also included in the study to identify the different perceptions based on the background knowledge. 7 PhD students from an information systems and technology (IST) program were recruited as the second group. The total sample size for this study was 15. ITU-T P.910 recommends having at least 15 participants who are not directly involved in picture quality evaluation as part of their work and are not experienced assessors. The sample in this study satisfies all these recommendations.

2) Selection of Impaired Videos: 14 impaired videos are selected from the pool of videos generated in the previous study. The goal was to generate a sample of videos where all the three impairment factors were represented. At the same time, we tried to have an impaired video pool of varying PSNR values. Table I provides a list of the 14 impaired videos with their PSNR values and the impairment factors used to generate them.

TABLE I. LIST OF EXPERIMENT VIDEOS

Exp#	Delay (ms)	Jitter (ms)	Drop(%)	PSNR(dB)
exp0r	0	0	0	Original
exp2r	100	0	0	41
exp3r	200	0	0	35
exp4r	300	0	0	31
exp26	0	0	5	28
exp28	400	0	5	26
exp29	0	2	5	24
exp20	400	2	0	22
exp30	50	2	5	21
exp19	300	2	0	19
exp12r	100	5	0	18
exp32r	0	5	5	17
exp8	200	10	0	16
exp11	50	5	0	15
exp15	400	5	0	14

3) Experiment Procedures: Each subject was asked to fill out a short questionnaire about their expertise and their previous telemedicine experience before the experiments. Once the experiment procedure is explained by the researchers, they were asked to watch 15 video sequences (see Table I) all generated from the same video source. Viewing conditions were controlled by using the same laptop machine for all the experiments. After viewing each video, the subjects were asked to provide a quality score for the video shown (0-100 continuous scale) and a clinical decision making capability score where they select among (1) I cannot make a clinical decision, (2) I would rather not make a clinical decision, (3) I can make a clinical decision only if this is an emergency case,

(4) I can make a medical decision, up to a reasonable certainty, (5) I can easily make a medical decision.

IV. RESULTS

Initial analysis was focused on the sample characteristics. As reported before, the subjects were asked to fill out a simple questionnaire before the experiments. Results of this survey indicate that out of 7 IST PhD students none of them had any telemedicine experience. On the other hand, 2 out of 8 optometrists had been involved in few telemedicine cases.

Next step in the analysis was to identify the mean and standard deviation values of the quality and decision making capability assessments. The continuous scale used for quality is evaluated as a 100 point scale. Descriptive analysis of the quality score for each video sequence by all, only IST, and only OD subjects are presented in Table II.

The clinical decision making capability scale is evaluated as a 5 point scale where "1" corresponds to "I cannot make a clinical decision" and "5" corresponds to "I can easily make a medical decision". Descriptive analysis of the capability score (cs) for each video sequence by all, only IST, and only OD subjects are presented in Tables III.

Quality scores (qs) which were on a 0-100 scale were later on converted into mean opinion score (MOS) and frequency analysis of MOS values are presented in Table IV. The results show that even the quality of the original video where no degradation was introduced does not reach the excellent score. This is caused by the original recording of the video clip where the camera was zoomed into a patient's eye in order to make the examination. During the zoom effect, the camera had hard time focusing on the eye since the contractions in the eye caused by rapid blinking were very fast. This introduced a blurry effect in the original video which cannot be eliminated. Therefore, the quality scores for the original video were below excellent.

TABLE II. DESCRIPTIVE ANALYSIS OF QUALITY EVALUATIONS

		ALL (n=15)		IST (n=7)		OD(n=8)	
Exp#	PSNR	Mean	St.dev.	Mean	St.dev.	Mean	St.dev.
e0rqs	Orig.	47.1	17.3	52.9	15.2	42.0	18.4
e11qs	15	26.9	17.6	31.3	21.1	23.0	14.3
e12rqs	18	26.3	14.4	31.4	17.4	21.9	10.2
e15qs	14	18.9	12.9	18.4	14.9	19.4	12.0
e19qs	19	34.5	14.9	32.7	16.1	36.1	14.7
e20qs	22	33.8	19.9	45.4	18.9	23.6	15.2
e26qs	28	45.3	18.5	51.3	14.6	40.1	20.8
e28qs	26	47.1	14.4	50.4	15.8	44.1	13.4
e29qs	24	33.9	18.5	35.1	21.3	32.8	17.2
e2rqs	41	50.3	16.6	54.6	18.8	46.5	14.6
e30qs	21	39.0	17.7	43.0	18.0	35.5	18.0
e32rqs	17	35.2	16.3	37.7	13.8	33.0	18.8
e3rqs	35	52.4	14.9	59.6	10.5	46.1	15.9
e4rqs	31	49.8	16.5	50.3	12.2	49.4	20.5
e8qs	16	12.3	11.7	16.1	15.6	9.0	6.1

TABLE III. DESCRIPTIVE ANALYSIS OF CAPABILITY EVALUATIONS

		ALL (n=15)		IST (n=7)		OD(n=8)	
Exp#	PSNR	Mean	St.dev.	Mean	St.dev.	Mean	St.dev.
e0rcs	Original	3.2	0.9	3.4	0.5	3.0	1.1
e11cs	15	2.3	1.1	2.3	1.1	2.3	1.2
e12rcs	18	2.2	1.0	2.3	1.4	2.1	0.6
e15cs	14	1.6	0.7	1.4	0.8	1.8	0.7
e19cs	19	2.7	1.0	2.6	1.0	2.9	1.0
e20cs	22	2.7	1.2	3.0	0.8	2.5	1.4
e26cs	28	3.3	1.0	3.7	0.5	3.0	1.2
e28cs	26	3.5	0.8	3.6	1.0	3.4	0.7
e29cs	24	2.5	1.1	2.4	1.3	2.5	0.9
e2rcs	41	3.3	0.8	3.4	1.0	3.3	0.7
e30cs	21	2.8	1.1	3.0	1.0	2.6	1.3
e32rcs	17	2.4	0.9	2.3	1.0	2.5	0.9
e3rcs	35	3.3	1.0	3.6	0.8	3.1	1.1
e4rcs	31	3.5	0.8	3.6	0.8	3.4	0.9
e8cs	16	1.3	0.6	1.6	0.8	1.1	0.4

Frequency analysis of the capability evaluations are presented in Table V. Only one expert subject assigned the highest confidence score to three videos. Overall, the scores assigned by this subject were high (3 or 4). This implies that the subject felt comfortable enough to make a decision based on this video if this was an emergency case.

After the descriptive analysis, difference between the quality and decision making capability perception of two expert groups (IST and OD) was analyzed using One-Way ANOVA. The results of the analysis indicate that there is no significant difference between the expert groups in terms of their quality and capability perceptions.

TABLE IV. FREQUENCY ANALYSIS OF MOS VALUES (N=15)

Epx#	PSNR	1	2	3	4	5
e0rmos	Original	1	5	5	4	0
e11mos	15	6	4	5	0	0
e12rmos	18	6	7	2	0	0
e15mos	14	10	3	2	0	0
e19mos	19	2	7	6	0	0
e20mos	22	5	5	3	2	0
e26mos	28	2	4	6	3	0
e28mos	26	1	5	7	2	0
e29mos	24	6	1	7	1	0
e2rmos	41	1	2	7	5	0
e30mos	21	4	3	7	1	0
e32rmos	17	4	5	5	1	0
e3rmos	35	1	2	9	3	0
e4rmos	31	1	3	7	4	0
e8mos	16	12	2	1	0	0

TABLE V. FREQUENCY ANALYSIS OF CAPABILITY EVALUATIONS (N=15)

Epx#	PSNR	1	2	3	4	5
e0rcs	Original	1	2	7	6	0
e11cs	15	5	3	5	2	0
e12rcs	18	3	8	3	0	1
e15cs	14	8	5	2	0	0
e19cs	19	2	3	7	3	0
e20cs	22	3	3	4	5	0
e26cs	28	1	2	3	9	0
e28cs	26	0	2	5	7	1
e29cs	24	3	5	4	3	0
e2rcs	41	0	3	4	8	0
e30cs	21	2	5	2	6	0
e32rcs	17	2	7	4	2	0
e3rcs	35	1	2	3	9	0
e4rcs	31	0	2	5	7	1
e8cs	16	11	3	1	0	0

To better understand the relationship between quality and capability values, correlations were analyzed for each experiment. Order of video sequences presented to each subject was based on a randomly generated list. Order in which the video sequence was presented to the subject might have an effect at the final perceived quality as well. Therefore, the correlations included the order variable to understand if there is any correlation between the order of presentation and the perceived scores assigned by the subjects.

The correlation between quality score and the capability score is expected to be high. If this assumption was correct, it would be easy to predict the decision making capability of a medical professional based on the perceived quality scores. In our experiments, out of 15 video sequences, for 10 of them the MOS values were correlated with the capability score values (significant at 0.01 level). MOS and cs for other 3 video sequences were correlated but at a lower significance level (0.05). For 2 of the video sequences (one being the non impaired video – e0r) there was no correlation between the two scores. However, one of these videos (e15) showed correlation between the qs and cs values (significant at 0.01 level). Appendix A presents some of the correlation tables.

V. DISCUSSION

Results presented in the previous section supports our initial hypothesis that the quality score does not necessarily correspond with a medical decision making capability score. Study subjects in their written and verbal comments emphasized the importance of the critical frames that they will base their decision on. They sometimes were able to make decisions with an overall low quality video sequence since the impairment in that sequence occurred at non-critical frames. They also mentioned that if they were able to see that critical frame, they will not need to watch the rest of the video. This finding has important implications on videoconferencing applications that are built to support telemedicine. In a telemedicine environment, it is important for the medical provider to have control on the quality. If the application provides functionality to increase the quality when there is request from the user, the decision making capability will be positively affected.

An interesting case occurred when one subject evaluated experiment 12r (e12r) with a 5 cs which indicates that the subject felt that a medical decision can be made easily based on this video. If we check in Table 6, the PSNR value calculated for this video sequence was 18dB, which indicates bad quality compared to the unimpaired video. Fig. 3 and Fig.4 below present two different frames from the same video.

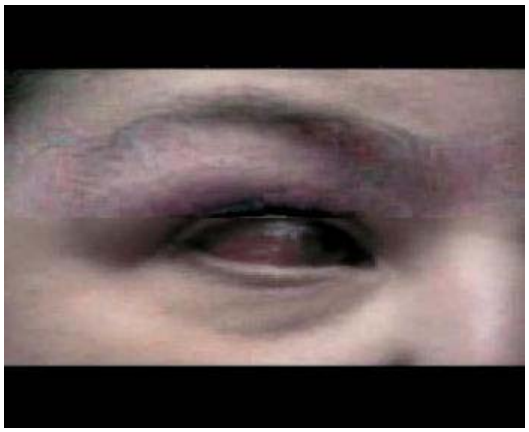

Figure 3. Experiment 12r Critical Frame

Figure 4. Experiment 12r 10 frames before the critical frame

Fig.3 represents the case where the impairments at the network level did not affect the critical frame which is useful for the medical experts during the decision making process. However, as Figure 4 illustrates, the overall quality of the video sequence was quite poor.

VI. CONCLUSION

This study investigated the relationship between objective and subjective quality measures, as well as clinical decision making capability that is affected by the impairments that occur during the transmission of video over the Internet. It was one of the few studies that emphasized differences in quality perception when the video and application is in the medical

context and a medical expert is the user depending on the video being transmitted to make a decision. The findings of this work suggest that further studies that measure quality of video using sequences from real world of telemedicine is necessary to understand the effects of quality on medical decision making. Only then, one can improve the existing applications to serve the needs of medical professionals.

We have stored the results of our study in a quality database, which will later be integrated into a smart video conferencing tool. This database will be utilized to predict quality and decision making capability based on the previously collected data. Identify potential safe zones of operation based on the subjective evaluations from experts is another future research area. If certain minimum thresholds are surpassed, that will prevent the telemedicine session to be effective. No such tool currently exists in the telemedicine market.

There are a few limitations in this study. First of all, quality of the original video sequence was not perfect. Hence, it is not easy to identify the reasons for the degradation in the video quality once it is transmitted over the internet. This is one of the reasons for the original video not scoring high quality values from the subjects of this study. They were complaining more about the blurry effect than the pixelization and smudging effects. Second, the study results are highly dependent on the application area used in this video. It was for general eye examination and the results of this study cannot be generalized for other domains such as telemental health before further research is completed.

Besides all these limitations, this study provides insights from a measurement research specific to telemedicine videos. It is important to study this domain in isolation from others to get better understanding of the user needs which will eventually increase the opportunities around the world to receive medical care via telemedicine.

REFERENCES

[1] S. Wolf and M. Pinson, "Video Quality Measurement Techniques," U.S. DEPARTMENT OF COMMERCE - National Telecommunication and Information Administration NTIA Report 02-392, June 2002.

[2] B. Tulu and S. Chatterjee, "Understanding the Impact of Network Impariments Over Internet-based Telemedicine Video Traffic," presented at Fifthe Annual Workshop on Information Technologies and Systems (WITS'05), Las Vegas, Nevada, USA, 2005.

[3] D. A. Rosenthal, "Analyses of selected variables effecting video streamed over IP," *International Journal of Network Management*, vol. 14, pp. 193-211, 2004.

[4] A. P. Markopoulou, F. A. Tobagi, and M. J. Karam, "Assessing the quality of voice communications over internet backbones," *IEEE/ACM Transactions on Networking*, vol. 11, pp. 747-760, 2003.

[5] M. Hassan, A. Nayandoro, and M. Atiquzzaman, "Internet telephony: services, technical challenges, and products," *IEEE Communications Magazine*, vol. 38, pp. 96 - 103, 2000.

[6] S. Mohamed, "Automatic Evaluation of Real-Time Multimedia Quality: a Neural Network Approach." Rennes: University of Rennes I, 2003.

[7] B. Goode, "Voice Over Internet Protocol (VoIP)," *Proceedings of the IEEE*, vol. 90, pp. 1495-1517, 2002.

[8] S. Shenker and J. Wroclawski, "General Characterization Parameters for Integrated Service Network Elements," Internet Engineering Task Force (IETF), RFC 2215, September 1997.

[9] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W.Weiss, "An Architecture for Differential Services," Internet Engineering Task Force (IETF), RFC 2475, December 1998.

[10] A. Webster, C. Jones, M. Pinson, S. Voran, and S. Wolf, "An Objective Video Quality Assessment System Based on Human Perception," presented at Storage and Retrieval for Image and Video Databases - Human Vision, Visual Processing and Digital Display TV, San Jose, CA, USA, 1993.

APPENDIX A – CORRELATIONS

	e12rorde	e12rqs	e12rcs	e12rmos
e12rorde	1.000			
e12rqs	0.348	1.000		
e12rcs	-0.068	.653(**)	1.000	
e12rmos	0.276	.908(**)	.681(**)	1.000

	e0rcs	e0rqs	e0rorder	e0rmos
e0rcs	1.000			
e0rqs	0.507	1.000		
e0rorder	0.467	0.435	1.000	
e0rmos	0.493	.939(**)	0.442	1.000

	e15order	e15qs	e15cs	e15mos
e15order	1.000			
e15qs	-0.245	1.000		
e15cs	-0.502	.651(**)	1.000	
e15mos	-0.090	.875(**)	0.496	1.000