MEDINFO 2001
V. Patel et al. (Eds)
Amsterdam: IOS Press

The Coming of Age of Computerized ECG Processing: Can it Replace the Cardiologist in Epidemiological Studies and Clinical Trials?

Jan A. Kors, Gerard van Herpen

Department of Medical Informatics, Erasmus University, Rotterdam, The Netherlands

Abstract

In spite of decades of research and widespread use of computer programs for the analysis of electrocardiograms (ECGs), the accuracy and usefulness of computerized ECG processing has been questioned. To determine whether ECG computer programs can replace cardiologists in epidemiological studies and clinical trials, we reviewed the literature for evidence, concentrating on one influential ECG measurement, viz. QT interval duration, and one classification method, the Minnesota Code, which is the de facto standard for ECG coding. We compared interobserver variabilities of cardiologists with differences between computer programs and cardiologists, in order not to prejudice against the computer. Studies that contain this type of information indicate that interobserver variabilities are at least as large as differences between computer and cardiologist. This suggests that ECG computer programs perform at least equally well as human observers in ECG measurement and coding, and can replace the cardiologist in epidemiological studies and clinical trials.

Keywords:

Electrocardiography; Computerized ECG Analysis; Interobserver Variability

Introduction

Automated analysis of electrocardiograms (ECGs) has been among the first applications of computers in medicine. Computer programs for ECG interpretation already appeared in the early sixties of the previous century [1,2]. Since then, spectacular technological advances in combination with increased signal analysis and pattern recognition skills have rendered ECG computer programs one of the most widespread computer applications in health care.

For the past few decades, the role of the ECG has been foremost in the care of the acute coronary patient, but in routine clinical practice it has steadily dwindled in favor of newer diagnostic techniques, e.g., echocardiography. However, in many epidemiological studies and clinical trials ECG parameters were shown to be important risk indicators [3]. In such studies, the ECG is still the test of choice because of its noninvasive character, ease of recording and low cost. Computerized ECG processing has some major advantages over human reading: computer programs do not suffer from fatigue and intraobserver variability, and offer substantial savings in time and money. Also, they allow the exploration of new measurements of potential interest.

Offsetting these advantages, concerns have been raised regarding the accuracy and robustness of ECG computer programs [4-6]. For example, the European Agency for the Evaluation of Medicinal Products states, in a report that discusses QT interval prolongation in relation to testing of drug compounds in clinical studies [6]: “At present, automatic ECG readings are generally not considered sufficiently accurate or reliable.” Are these just criticisms, or can we rather say that ECG computer programs have sufficiently matured to replace human observers in epidemiological studies and clinical trials? To answer these questions, we shall review the evidence that has accumulated in the literature. Since a comprehensive review is beyond the scope of this article, we will concentrate on one classification method, the Minnesota Code (MC) [7], which is the most widely used ECG coding scheme in epidemiological studies. Further we will focus on QT interval duration measurements that are not contained in the MC but have received a great deal of attention in the past decades [8-11]. Also, we will make recommendations to remedy several of the shortcomings in the current practice of algorithm validation, and pay attention to some promising new approaches in computerized ECG processing that may alleviate part of the present problems.

Methods

Criteria for comparison and selection of studies

When results from previous studies are the basis for a comparison between the performance of ECG computer programs and cardiologists, at least two problems arise. One problem is that most studies use their own sets of ECGs. Size and composition of the data sets may show considerable variation, which complicates a comparison across studies. The problem of course is resolved when studies would use the same set of ECGs, but unfortunately very few such sets exist. For waveform recognition, i.e., determination of the onsets and offsets of P, QRS, and T, well-annotated datasets were established in the project Common Standards for Quantitative Electrocardiography (CSE) [12,13]. These ECG databases have since been used in different other studies. For the MC, no generally accepted and accessible database is available.

The other problem is the reference standard against which the computer programs are gauged. Both for ECG measurement and for coding this reference is almost always the human observer (there have been a few notable exceptions to this rule that will be discussed later on). If a standard is perfect, any method being evaluated against it must fall short of the mark, but the method will also appear worse than the reference standard when the latter is imperfect [14], even if it would outperform the imperfect standard when both were gauged against the perfect standard. By definition this prejudices against the computer. However, results can be accredited if information about interobserver variability is provided. Differences between computer and (a consensus opinion of) human observers should then be in the order of the interobserver variability if the computer performs equally well as the observer. Thus, we only considered studies that employed multiple experts to set a reference standard and provide information about interobserver, and possibly also intraobserver, variability.

Computerized ECG measurement

In the past decades, a dozen or so ECG computer programs have been developed by university groups and by industry [15]. Nowadays, a typical ECG computer program processes the simultaneously recorded 12 leads of the standard rest ECG, sampled at a rate of 500 Hz. Most programs compute a single representative P-QRS-T complex for each lead by taking the mean or the median of the individual, characteristic beats in the ECG recording. The representative complexes serve to determine the onsets and offsets of the P wave and the QRS complex, and the offset of the T wave. Once this waveform recognition has been performed, the derivation of wave amplitudes and durations is relatively straightforward.

QT interval is defined as the interval between the onset of the QRS complex and the end of the T wave taken commonly over all 12 leads. The determination of T offset is generally considered the most intricate part of waveform recognition, because the T wave tapers off more or less gradually to the baseline, may have very low amplitude, may be trailed by a U wave, or have a following P wave superimposed on it. Many different algorithms to detect the end of the T wave have been proposed, most of them applying some sort of thresholding or template-matching technique [16-18].

Computerized ECG coding

The MC purports to be a formal, objective description of the morphology of the ECG [7]. It is the most widely used classification system for epidemiological studies and clinical trials. It consists of a set of measurement procedures, coding prescriptions, and exclusion rules of such complexity as to make visual coding cumbersome. Only a limited number of computer programs that automate the coding process are currently operational [19-21].

Review of results

In reviewing results for QT interval measurement and Minnesota Coding, our approach will be to first present interobserver variabilities, and then comparisons of computerized versus manual ECG analysis (both types of results may be included in one study).

ECG measurement

Information about interobserver variability of QT-interval measurement is scanty. In the CSE study, a board of five cardiologists established the two most sizable reference libraries for resting ECG waveform recognition till date [12,13]. However, interobserver variabilities were provided for only one of these libraries (n=310) [12], and only for wave onsets and offsets, not for the QT interval as such. The average interobserver variability proved to be 4.8 ms for QRS onset, and 21.3 ms for T end. This latter value may be considered a lower bound for the interobserver variability of the QT interval. It should be noted that the wave reference points in this study were determined in groups of three simultaneously recorded leads. For the second reference library (n=250) [13], the CSE Party focussed on simultaneous 12-lead ECGs, but interobserver variabilities were not presented. However, waveform recognition results of each of the five cardiologists were separately published for a small subset of ECGs (n=25) [22]. From these results, we were able to compute average interobserver variabilities ourselves. For QRS onset and T end, we found variabilities of 2.8 and 15.0 ms, respectively. These variabilities are somewhat smaller than those from the previous study, denoting that waveform recognition using 12 simultaneous leads is easier and more robust than using only three leads.

With the recent advent of the idea of QT dispersion, a few other studies assessed inter- and intraobserver variabilities for QT intervals measured in individual leads [23-26]. QT measurement in single leads, however, should be considered a step backwards, not only because it unnecessarily discards information from multiple leads, thus increasing measuement error, but also because the notion of different T-wave offsets in different leads is a physical misconception [27]. We therefore did not bother to further review these results.

For the second CSE waveform library [13], the reference was compared with 11 ECG computer programs for different interval measurements, including the QT interval. The standard deviation of the differences in QT interval between reference and programs varied between 9.8 and 18.6 ms [28]. These values show a quite wide variation among the various programs, but also suggest, when compared with the interobserver variability of 15 ms, that the best programs performed at least as well as the cardiologists.

Again, several other studies [17,25,26,29,30] that focussed on QT dispersion compared manual and computerized measurement of QT intervals, but always in single leads and often involving only one human observer who set the reference. For the same reasons as mentioned above, we elected to leave these studies out of consideration.

ECG coding

Data about interobserver variability in Minnesota coding are also sparse. We culled only three studies. The reference in these studies was the consensus opinion of two or more expert coders. We concentrated on one important category in the MC, code 1, which deals with the classification of Q and QS waves that are important for the assessment of myocardial infarction. Table 1 gives sensitivities and specificities, where sensivity is defined as the probability that an individual human coder issues a code 1 when the reference is code 1, and specificity is the probability that no code 1 is given when the reference indicates no code 1 either. Rautaharja et al. [31] mention a sensitivity as low as 70% for technician coders; for senior coders the sensitivity is 90% or higher. The other two studies present comparable sensitivities, with high specificities.

Table 1. Sensitivity and specificity (in %) of visual coding for Minnesota Code 1 (Q and QS patterns)

Study	Sensitivity	Specificity
Rautaharju et al. [31]	70-≥90	
Tuinstra et al. [32]	87.5-95.5	93.2-99.6
Kors et al. [21]	88.3	98.2

In two of these studies [21,32], computer programs were compared with human performance (Table 2). In the oldest study by Tuinstra et al. [32], three computer programs were assessed. Sensitivities proved comparable, but specificities lagged behind those of the human observers. In a more recent study of one particular program [21], we found that the computer had a specificity equal to the human coders, but better sensitivity. Recently, we undertook yet another study [33] in which two ECG computer programs were compared with a visual standard. The results lie more or less in between those of the two previous studies. Unfortunately, interobserver variability that would allow a direct comparison with human performance was not assessed in this study.

This last study [33] presents an interesting alternative for use of a human reference standard, because it also compared both computerized and visual coding with respect to their prognostic associations with coronary heart disease (CHD) events in a population-based, prospective cohort study. It was demonstrated that computer-detected MC better predicts new CHD events than the human coder.

Table 2. Sensitivity and specificity (in %) of computer coding for Minnesota Code 1 (Q and QS patterns)

Study	Sensitivity	Specificity
Tuinstra et al. [32]	84.0-95.5	87.0-92.9
Kors et al. [21]	97.8	98.2
Kors et al. [33]	86.4-94.3	91.1-95.3

Discussion

We reviewed pertinent results from the literature to determine whether ECG computer methods have sufficiently matured to replace cardiologists for ECG analysis in epidemiological studies and clinical trials. For the measurement and classification that were studied, computer algorithms have an accuracy level comparable to or better than humans. Thus, computerized ECG processing is a viable alternative for human involvement.

Can our results be generalized to other ECG (compound) measurements as well? An answer must necessarily be speculative since little information about the accuracy of measurements other than QT interval is available. However, the measurement of ECG amplitudes and durations is a relatively straightforward task once correct waveform recognition points are available. As we have seen, results from the CSE study demonstrate that the waveform recognition of the better ECG computer programs is as good as that of expert cardiologists. This suggests that ECG computer measurements in general will be at least as accurate as those of cardiologists.

This is not to say that computer programs cannot make serious waveform recognition errors. Computers lack the “common sense” that is natural to humans when performing pattern-recognition tasks. Thus, computers may occasionally go astray in a way a human never would. It has therefore been suggested to visually verify all waveform recognition results of the computer, correcting faulty wave onsets and offsets manually [20]. Whether the increased workload involved in this approach pays off against the correction of mistakes that a good computer program will rarely make, is an open question.

We explicitly limited our review to ECG measurement and coding, not addressing diagnostic ECG interpretation, which carries its own intricacies. Computerized ECG interpretation, however, is mainly utilized in clinical settings; its relevance for epidemiological studies and clinical trials is not great.

Comparison between computer and human, and between different computer algorithms is greatly facilitated by the availability of well-validated databases. The CSE project established several such databases, but of relatively small size. Initiatives have been started to compile datasets of

ECGs that have been collected in past studies [34], but the establishment of a reliable human reference standard appears to be the bottleneck in these efforts, mainly because such a validation would require heavy involvement of a group of cardiologists. Also, considering the scarcity of information on interobserver variabilities, studies should publish this information to allow a fair comparison of automatic and manual methods.

Another approach is to assess the performance of both computer programs and cardiologists against an non-ECG reference standard, e.g., fatal cardiac events or cardiac disorders that were validated on the basis of non-ECG evidence. This approach, which previously was followed in the CSE study to assess the diagnostic performance of computer programs and cardiologists [35], and more recently to compare computerized and visual Minnesota coding [33], does not require interobserver variabilities for a proper comparison. However, the establishment of a golden clinical reference standard is likely to prove a formidable task.

Finally, we want to call attention to several new, promising directions of research that circumvent some of the measurement problems or elaborate on the existing approaches. First, measurements have recently been proposed that capture aspects of T-wave morphology without having to rely on accurate waveform recognition [36-38]. Several of these parameters have been shown to have high prognostic value [36,37]. Second, different ECGs of the same individual, recorded at different time instants, can be compared. Serial changes in selected ECG measurements and in the MC classification have been reported to have strong predictive value for CHD events [39-41], and warrant further investigation. Also, alternative ECG classification schemes have been proposed [39] that solve some problematic aspects of the MC, especially its instability for small measurement changes. Comparitive studies are needed to validate these newer methods.

Conclusion

Results from the literature indicate that interobserver variabilities of cardiologists are at least as large as differences between ECG computer programs and cardiologists. This implies that computer programs perform at least equally well as human observers in ECG measurement and coding, and can replace the cardiologist in epidemiological studies and clinical trials.

References

[1] Pipberger HV, Arms RJ, Stallmann FW. Automatic screening of normal and abnormal electrocardiograms by means of a digital electronic computer. *Proc Soc Exp Biol Med* 1961;106:130-2.

[2] Caceres CA, Steinberg CA, Abraham S, Carbery WJ, McBride JM, Tollens WE, et al. Computer extraction of electrocardiographic parameters. *Circulation* 1962;25:356-62.

[3] Rautaharju PM. Electrocardiography in epidemiology and clinical trials. In: Macfarlane PW, Lawrie TDV, eds. *Comprehensive Electrocardiology*. New York: Pergamon Press; 1989. p. 1219-66.

[4] Hurst JW, Treasure CB, Sathavorn CS. Computer errors in electrocardiography. *Clin Cardiol* 1996;19:580-6.

[5] RuDusky BM. Errors of computer electrocardiography. *Angiology* 1997;48:1045-50.

[6] Committee for Proprietary Medicinal Products. *Points to consider: the assessment of the potential for QT interval prolongation by non-cardiovascular medicinal products*. Report No.: CPMP/986/96. London: European Agency for the Evaluation of Medicinal Products; 1997. Available from: URL: http://www.eudra.org/en_home.htm.

[7] Prineas RJ, Crow RS, Blackburn H. *The Minnesota Code Manual of Electrocardiographic Findings.* Boston: John Wright PSB; 1982.

[8] Schwartz PJ, Wolf S. QT interval prolongation as predictor of sudden death in patients with myocardial infarction. *Circulation* 1978;57:1074-7.

[9] Bellavere F, Ferri M, Guarini L, et al. Prolonged QT period in diabetic autonomic neuropathy: a possible role in sudden cardiac death? *Br Heart J* 1988;59:379-83.

[10] Goldberg RJ, Bengtson J, Chen ZY, Anderson KM, Locati E, Levy D. Duration of the QT interval and total and cardiovascular mortality in healthy persons (the Framingham Heart Study experience). *Am J Cardiol* 1991;67:55-8.

[11] De Bruyne MC, Hoes AW, Kors JA, Hofman A, Van Bemmel JH, Grobbee DE. Prolonged QT interval predicts cardiac and all-cause mortality in the elderly: the Rotterdam study. *Eur Heart J* 1999;20:278-84.

[12] Willems JL, Arnaud P, Van Bemmel JH, et al. Establishment of a reference library for evaluating ECG measurement program. *Comput Biomed Res* 1985;18:439-57.

[13] Willems JL, Arnaud P, Van Bemmel JH, Bourdillon PJ, Degani R, Denis B, et al. A reference data base for multilead electrocardiographic computer measurement programs. *J Am Coll Cardiol* 1987;10:1313-21.

[14] Rautaharju PM, Smets P. Evaluation of computer ECG programs. The strange case of the golden standard. *Comput Biomed Res* 1979;12:39-45.

[15] Willems JL. Computer analysis of the electrocardiogram. In: Macfarlane PW, Lawrie TDV, eds. *Comprehensive Electrocardiology*. New York: Pergamon Press; 1989. p. 1139-76.

[16] Van Bemmel JH, Zywietz C, Kors JA. Signal analysis for ECG interpretation. *Methods Inf Med* 1990;29:317-29.

[17] McLaughlin NB, Campbell RWF, Murray A. Comparison of automatic QT measurement techniques in the normal 12 lead electrocardiogram. *Br Heart J* 1995;74:84-9.

[18] Xue Q, Reddy S. Algorithms for computerized QT analysis. *J Electrocardiol* 1998;30 Suppl:181-6.

[19] Macfarlane PW, Devine B, Latif S, McLaughlin S, Shoat DB, Watts MP. Methodology of ECG interpretation in the Glasgow program. *Methods Inf Med* 1990;29:354-61.

[20] Rautaharju PM, MacInnis PJ, Warren JW, et al. Methodology of ECG interpretation in the Dalhousie program; NOVACODE ECG classification procedures for clinical trials and population health surveys. *Methods Inf Med* 1990;29:362-74.

[21] Kors JA, Van Herpen G, Wu J, Zhang Z, Prineas RJ, Van Bemmel JH. Validation of a new computer program for Minnesota coding. *J Electrocardiol* 1996;29 Suppl:83-8.

[22] Willems JL, ed. *CSE multilead atlas*. Report No.: CSE 88-04-15. Leuven: ACCO Publ; 1988.

[23] Kautzer J, Yi G, Camm J, Malik M. Short- and long-term reproducibility of QT, QTc, and QT dispersion measurement in healthy subjects. *PACE* 1994;17:928-37.

[24] Murray A, McLaughlin NB, Bourke JP, Doig JC, Furniss SS, Campbell RWF. Errors in manual measurement of QT intervals. *Br Heart J* 1994;71:386-90.

[25] Glancy JM, Weston PJ, Bhullar HK, Garratt CJ, Woods KL, De Bono DP. Reproducibility and automatic measurement of QT dispersion. *Eur Heart J* 1996;17:1035-9.

[26] Kors JA, Van Herpen G. Measurement error as a source of QT dispersion: a computerised analysis. *Heart* 1998;80:453-8.

[27] Kors JA, Van Herpen G, Van Bemmel JH. QT dispersion as an attribute of T-loop morphology. *Circulation* 1999;99:1458-63.

[28] Willems JL. *Common Standards for Quantitative Electrocardiography*. *CSE 6th Progress Report*. Report No.: CSE 86-12-08. Leuven: ACCO Publ; 1986. p. 139.

[29] Murray A, McLaughlnin NB, Campbell RWF. Measuring QT dispersion: man versus machine. *Heart* 1997;77:539-42.

[30] Savelieva I, Yi G, Guo X, Hnatkova K, Malik M. Agreement and reproducibility of automatic versus manual measurement of QT interval and QT dispersion. *Am J Cardiol* 1998;81:471-7.

[31] Rautaharju P, Warren J, Prineas RJ, Smets P. Optimal coding of electrocardiograms for epidemiological studies. The performance of human coders–a statistical model. *J Electrocardiol* 1979;13:55-9.

[32] Tuinstra CL, Rautaharju PM, Prineas RJ, et al. The performance of three visual coding procedures and three computer programs in classification of electrocardiograms according to the Minnesota Code. *J Electrocardiol* 1982;15:345-9.

[33] Kors JA, Crow RS, Hannah PJ, Rautaharju PM, Folsom AR. Comparison of computer-assigned Minnesota Codes with the visual standard method for new coronary heart disease events. *Am J Epidemiol* 2000;151:790-7.

[34] Norman JE, Bailey JJ, Berson AS, Haisty WK, Levy D, Macfarlane PW, Rautaharju PM. NHBLI workshop on the utilization of ECG databases. *J Electrocardiol* 1998;31:83-9.

[35] Willems JL, Abreu-Lima C, Arnaud P, et al. The diagnostic performance of computer programs for the interpretation of electrocardiograms. *N Engl J Med* 1991;325:1767-73.

[36] Priori SG, Mortara DW, Napolitano C, Diehl L, Paganini V, Cantu F, Cantu G, Schwartz PJ. Evaluation of the spatial aspects of T-wave complexity in long-QT syndrome. *Circulation* 1997;96:3006-12.

[37] Kors JA, De Bruyne MC, Hoes AW, Van Herpen G, Hofman A, Van Bemmel JH, Grobbee DE. T axis: a new risk indicator for cardiac events in the elderly. *Lancet* 1998;352:601-5.

[38] Acar B, Yi G, Hnatkova K, Malik M. Spatial, temporal and wavefront direction characteristics of 12-lead T-wave morphology. *Med Biol Eng Comput* 1999;37:574-84.

[39] Rautaharju PM, Calhoun HP, Chaitman BR. NOVACODE serial ECG classification system for clinical trials and epidemiologic studies. *J Electrocardiol* 1991;24 Suppl:179-87.

[40] Michaelis J, Lippold R, Nafe B, Scheidt E. Risk assessment of future myocardial infarction from automated serial ECG analysis. *J Electrocardiol* 1992;25 Suppl:20-5.

[41] Crow RS, Prineas RJ, Hannan PJ, Grandits G, Blackburn H. Prognostic associations of Minnesota Code serial electrocardiographic change classification with coronary heart disease mortality in the Multiple Risk Factor Intervention Trial. *Am J Cardiol* 1997;80:138-44.

Address for correspondence

J.A. Kors, PhD
Department of Medical Informatics
Faculty of Medicine and Health Sciences
Erasmus University
P.O. Box 1738
3000 DR Rotterdam
The Netherlands
E-mail: kors@mi.fgg.eur.nl