COMPUTERS AND BIOMEDICAL RESEARCH 18, 439–457 (1985)

Establishment of a Reference Library for Evaluating Computer ECG Measurement Programs

JOS L. WILLEMS,* PIERRE ARNAUD, JAN H. VAN BEMMEL, PETER J. BOURDILLON, ROSANNA DEGANI, BERNARD DENIS, FRITS M. A. HARMS, PETER W. MACFARLANE, GIANFRANCO MAZZOCCA, JÜRGEN MEYER, HENK J. RITSEMA VAN ECK, ETIENNE O. ROBLES DE MEDINA, AND CHRISTOPH ZYWIETZ

*University Hospital St. Rafael-Gasthuisberg, Herestraat 49, 3000 Leuven, Belgium†

Received January 2, 1985

As a result of an international cooperative project entitled "Common Standards for Quantitative Electrocardiography" (CSE), an ECG reference data base has been established with the aim of standardizing computer-derived ECG measurements. The objective of the project is to reduce the wide variation in wave measurements currently obtained by ECG analysis programs. A library of 250 ECGs with selective ECG abnormalities was established and a comprehensive reviewing scheme was devised for the visual determination of the onsets and offsets of P, QRS, and T. This task was performed by a board of cardiologists on highly amplified, selected complexes from the library. A subset was examined in order to study beat-to-beat and intraobserver variability. By using a modified Delphi approach, individual outlying point estimates were eliminated in four successive rounds. In this way final referee estimates were obtained which proved to be highly reproducible and precise. A reference library has thereby been developed which allows testing of the performance of ECG measurement programs and is a useful instrument in establishing recommendations for more precise measurement rules and definitions.

INTRODUCTION

The standardization of computer-derived ECG measurements has been recommended on many occasions over the last decade (*1–5*). Cooperative efforts have been strongly recommended (i) for testing the performance of various ECG computer programs using well-established test libraries and evaluation procedures and (ii) for developing more consistent and precise ECG measurement rules, as well as definitions and terminology for implementation in computer programs. However, concerted actions aimed at meeting these goals have been very difficult to establish. The need for such efforts has been underlined in studies demonstrating substantial differences in the reporting of small Q and R

† See Appendix for other academic affiliations.

0010-4809/85 $3.00

waves and in time measurements by different computer programs analyzing identical digital ECG recordings (*6, 7*). A situation has thereby been created whereby the exchange of diagnostic criteria and the transfer of scientific results is hampered. It is evident that diagnostic criteria developed by one particular program cannot be transferred to another program without critical review and a common yardstick.

To overcome some of the problems, an international project entitled Common Standards for Quantitative Electrocardiography (CSE), was initiated in the European Communities with the aim of developing common standards for computer-derived ECG measurements (*8–13*). The principal objectives of CSE are to establish recommendations for the standardization of computer-derived ECG measurements and to obtain agreement on definitions of waves and of references for the on- and offsets of P, QRS, and T waves. When the same data are given as input to programs A, B, or C, the ultimate goal is to obtain the same measurement results, for example, of Q durations. Only then can diagnostic ECG criteria for infarction and other diagnostic categories be exchanged and possibly standardized. Means and variances of measurement results obtained by various programs analyzing a common data base should fall within acceptable ranges.

The purpose of this paper is to describe the establishment of such a reference library and the methods used to ensure maximum quality and reproducibility of the reference.

Methods

Protocol and Organization

The CSE Working Party consists of active participants from 21 institutes of the European Communities. In addition, investigators from 6 North American and 1 Japanese center also collaborated in the project, by processing data or as consultants (see Appendix). The data collection, various procedures for human and computer analysis of the common CSE library, as well as formats for transferring information to and from the coordinating center were established during a pilot study, and have already been described in detail (*8–13*). The visual analysis of the first phase CSE library (vide infra) has been performed by a board of cardiologists from five different countries. These referees had experience of computer-assisted ECG interpretation, but had not been involved in program development. Computer analysis of the first phase library has been performed in each of the processing centers.

CSE First Phase Library

At the start of the project in 1979, the newer computer compatible recorders which record all leads simultaneously were not yet on the market. The first stage CSE library has therefore been confined to Frank XYZ and standard 12-lead ECGs recorded in lead groups with a minimum of 3 simultaneous leads.

Seventy ECGs, analyzed by the referees, were recorded with six simultaneous leads, i.e., the six peripheral and the six precordial leads.

Five centers submitted digitized ECGs to the coordinating center. From this set an enriched pathological sample of 250 ECGs with different morphologies and common rhythm disturbances seen in daily cardiological practice was selected. All recordings were made in patients above age 14. The library encompasses 32 complete normal recordings (normal P, QRS, and ST-T), 11 with incomplete right bundle branch block (RBBB), 18 with complete RBBB, 16 left anterior fascicular block, 9 complete left bundle branch block, 24 acute myocardial infarction (MI), 19 old anterior MI, 21 old posterodiaphragmatic MI, 12 old lateral MI, 11 MI with conduction defects, 22 left ventricular hypertrophy, 17 right ventricular hypertrophy, 6 pulmonary emphysema, 10 Wolff-Parkinson-White, 42 cases with atrial fibrillation or flutter, and 32 with ventricular extrasystoles. These diagnoses were based only on ECG findings and not on independent clinico-pathological data. In total 67 patients had strict normal QRS complexes, 91 myocardial infarction patterns, and 48 conduction disturbances with QRS duration greater than 120 msec. The sampling rate was 500 Hz and a quantization level of 5 μV or less was required. Further technical details can be found elsewhere (*8–13*).

In view of the different techniques applied in various computer programs, particularly the use of different beats for measurement, a so-called artificial ECG library was constructed, in addition to the original basic library. By selecting one beat from each of the lead groups of the original ECG recordings, strings of identical beats were composed with constant RR interval. A variable segment was interlaced between the beats in order to correct for possible offset artifacts. The selected beats were chosen by eye in such a way as to be close to the dominant beat with the least possible baseline shift, noise, and artifact. These beats were analyzed by the referees.

Another group of 60 artificial ECGs was composed from the beats additionally selected for a study of beat-to-beat variation so that the artificial library was composed of 310 recordings. Two beats adjacent to the one first selected were chosen for the beat-to-beat variability study. In order to study the intraobserver variability, the referees were offered the same beats of 26 ECGs on two further occasions, randomly over a period of 1 year. The total number of ECGs analyzed by each referee was thus 310 plus 52 and by the programs 310 artificial as well as the 250 original recordings.

The original and corresponding artificial recordings were evenly divided into two data sets, containing almost equal samples of each pathological entity. The ECGs of data set 1 and 2 were interspersed and could not be identified by the referees.

Analysis by the Referees

An overview of the analysis by the referees is presented in Fig. 1.

In view of the well-known inter- and intraobserver variability in determining wave recognition points, an elaborate reviewing scheme has been devised in

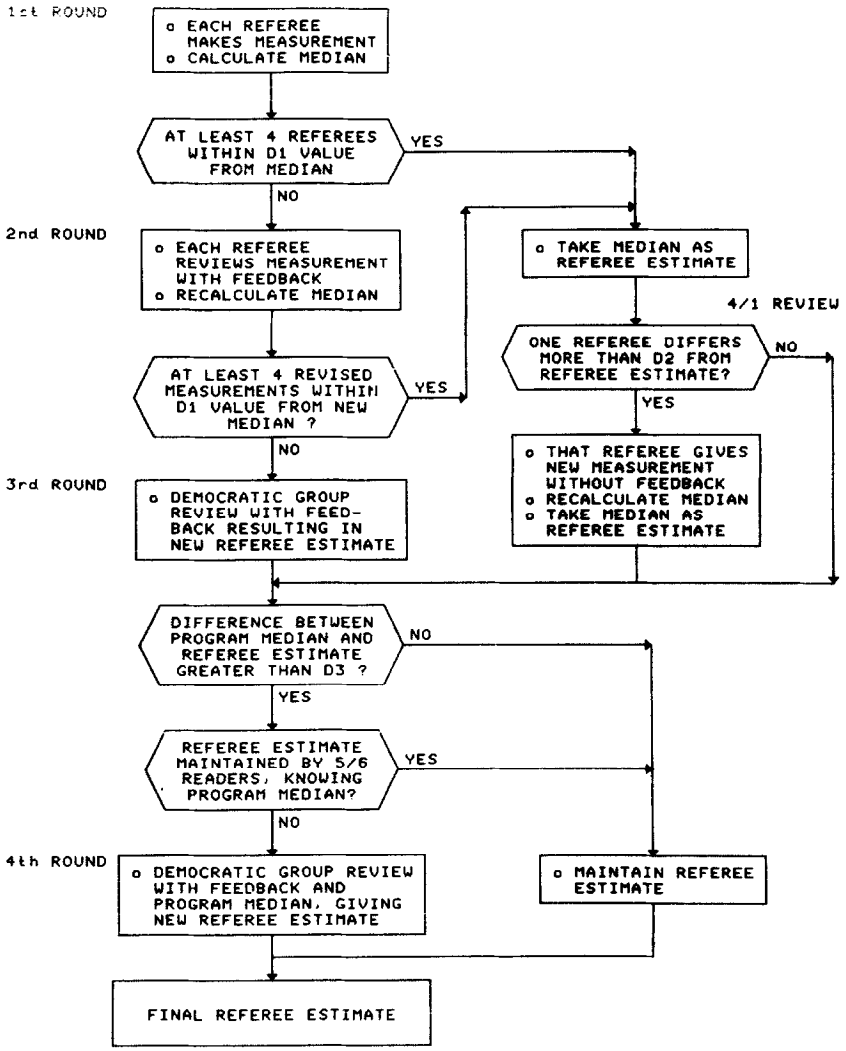

FIG. 1. Summary of the different reviewing rounds and the final determination of P, QRS, and T onsets and offsets by the group of referees. The limits D1 to D3 used to estimate deviations of individual versus median referee results were respectively as follows for P onset and offset, QRS onset and offset, and end of T: for D1—10,10,6,10, and 26 msec; for D2—12,14,8,14, and 36 msec; and for D3—12,12,6,10, and 28 msec. (Reproduced in modified form with permission of the American Heart Association from *Circulation* **71**:523–534, 1985.)

CSE. By using a modified Delphi approach (*14*), individual referee outliers have been eliminated in successive steps, an outlier being a point estimate which differs considerably from the median referee result. The ultimate goal was to obtain a final estimate, which would be as precise as possible and should serve as a standard for computer ECG measurement.

The Delphi method is an iterative procedure in which experts are supposed

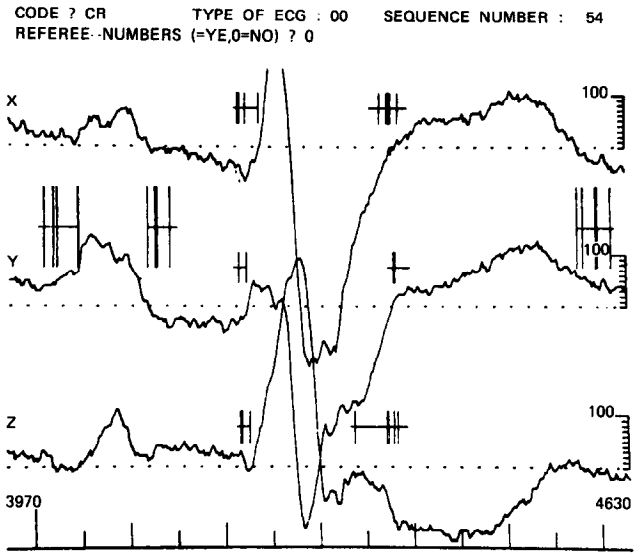

FIG. 2. Example of an enlarged beat (amplification ×10) and first round results given as feedback for second round analysis. Vertical lines denote the five individual point estimates, horizontal ones the ranges of confidence intervals. Note that individual referee estimates may overlap.

to judge or measure a certain event, using an adjusted assessment scale (*14*). During the iterative process of judging or measuring rounds, the experts are provided with identified previous results. In our approach, the original method has been modified in such a way that, for certain rounds, no feedback was given to the referees while in other rounds previous results were available, but without identification. The methods used were as follows.

First round analysis. Each referee received three recordings per ECG written out on a 4-channel Mingograph recorder. The first was a 25-mm/sec, 10-mm/mV tracing with 5 sec of the four standard lead groups and 10 sec of XYZ data. The fourth channel on the trace indicated which complex of a lead set the referee was asked to analyze. In the second recording these individual complexes were written out separately at 500 mm/sec and 100 mm/mV gain. On these recordings the referees were asked to indicate point estimates with corresponding upper and lower confidence limits for the group on- and offsets of the P wave and for the group end of the T wave, as well as for the individual on- and offsets of the QRS complexes in each lead (see Fig. 2). The earliest onset and latest offset of QRS in any lead was taken as the group onset and offset, respectively. In addition, they had to provide, per lead, a wave morphology description, e.g., P+QRSR'T+, i.e., positive P and T wave and an R' after a QRS complex. Third, they were sent a low pass filtered output of the selected beat (3 dB point at 15 Hz and zero output at 35 Hz) at a paper speed of 250 mm/sec of the selected beats. The purpose of this trace was to assist the referees in identifying the localization of P and T reference points.

All the markings of the referees (45 per ECG × 362 × 5 referees) had to be made on the 500 mm/sec recordings, where sample indications on the 4th

channel were 1 mm (i.e., 2 msec) apart. The time locations of the 81450 markings that the five referees made, were manually read in the coordinating center and entered into a computer for statistical analysis.

If at least four of the five referees agreed within a delta (or tolerance) value for the point estimate, the median of the five measurements was accepted into the data base. The delta values were empirically derived from a pilot study (*9*). They were fixed in such a way that approximately 10 to 15% of the measurements would have to be reviewed. This number was acceptable from a workload standpoint.

Second and third round analysis. Programs were developed in the coordinating center for an interactive analysis of certain measurements on a Tektronix 4010 graphics display terminal. When at least two out of the five referees differed by more than delta msec from the median of the whole group, all referees had to review the point estimates. The delta limits were respectively 10 msec for the on- and offsets of P and for the end of QRS, 6 msec for QRS onset, and 26 msec for the end of T. The second round was performed by each referee individually at the coordinating center. The point estimates made by each of the five referees in the first round were displayed without reader identification (Fig. 2) and a new estimate had to be made using the terminal's cursor. In the original Delphi procedure (*14*) feedback is given in the second round with identification of the reader's previous result. This was not performed in CSE in order to reduce personal bias. Each referee came to the coordinating center for five second round sessions each lasting 1 day. The second round reduced the cardiologists scatter considerably but delta limits were still exceeded in about 3% of the measurements. Those measurements were reviewed jointly by all referees during four 1-day third round sessions at the coordinating center. In addition to having the second round point estimates displayed, the median of the second round was also shown (see Fig. 3). A consensus or majority decision was reached for these difficult cases.

The ECGs were read in five batches over a period of $1\frac{1}{2}$ years. For the 26 ECGs which were randomly read three times by the referees, the median of the final estimates of these readings was computed for each measurement, and was used as the final measurement. However, if one of the measurements of these recordings had been submitted to a third round reading, then that result had priority and was finally taken into the data base. In case a QRS measurement was involved, QRS readings of the other leads were also given priority.

4/1 review. Each referee was given the opportunity to correct measurement errors interactively in cases where he deviated as single referee (4/1 review) by more than delta milliseconds from the group median (delta being 8 msec for QRS onset, 12 msec for P onset, 14 msec for the offset of P and QRS, and 36 msec for T end). These new delta values were again empirically established in order to eliminate as many remaining outliers as possible, while keeping the workload of the referees at an acceptable level. No feedback of previous readings was given for this exercise.

Fourth round. The cases where median results of the computer programs and referee estimates differed widely (see Figs. 1 and 4) were looked at again by the

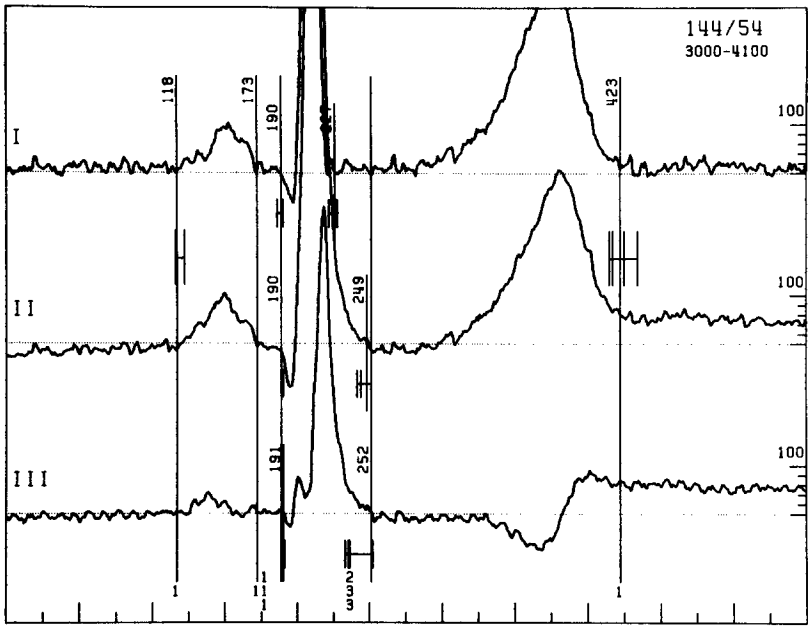

FIG. 3. Example of feedback for third round discussion. The long vertical lines denote the median estimates of five referees for the P, QRS, and T fiducials, obtained after the second round analysis. In this case, QRS end had to be discussed in lead II and III. Note the 50-msec-wide isoelectric segment at QRS end in lead I.

board of cardiologists and the project leader in order to ensure that obvious mistakes in the visual analysis had not occurred. Only lead group onsets and offsets of P, QRS, and T have been analyzed in this step. The measurements were firstly screened at home using enlarged Tektronix plots displaying both referee point estimates and median program results. In this offline analysis referee estimates were maintained in 84% of the reviewed cases by a minimum of five readers out of six, the referees plus the project leader. They interactively screened the remaining problem cases during a last common reviewing round in the coordinating center.

Statistical Data Analysis

Differences between individual referee results and final group estimates were calculated and parametric statistics were used to compute the significance of mean differences and variances.

Reproducibility of the median and individual referee results was examined in the 26 ECGs which were read three times over the study period. For the reproducibility per referee, the average of the three readings was first calculated. Absolute deviations from the corresponding average were computed next for each measurement.

Measurement precision was assessed in the ECGs, where six leads were

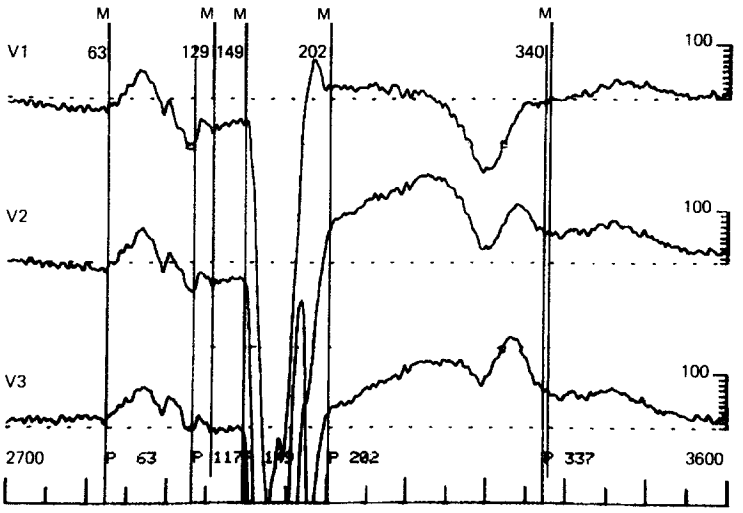

FIG. 4. Example of a case which has been submitted for a fourth round analysis. Final referee results are indicated with vertical lines labeled with M; median program results are identified by P at the foot. The adjacent numbers denote time locations in sample points. In this case P end had to be verified.

recorded simultaneously, but which were analyzed in sets of three. Differences between results from the bipolar and unipolar limb leads were compared to those of the precordial leads using paired Student's t and chi-square tests. From a theoretical point of view, wave onsets and offsets should be the same in simultaneously recorded unipolar and bipolar limb leads, since these leads are mathematically interrelated. This is not the case for the precordial leads.

P and QRS durations as well as QT and PR intervals were derived from the final point estimates. The earliest onset and latest offset in any of the three leads was used to calculate the QRS duration in the respective lead groups. These leadgroup onsets and offsets were also used to compute so-called isoelectric segments at the beginning and end of QRS of each lead, by measuring the distance to QRS onsets and offsets determined in the respective single leads.

RESULTS

Number of Measurements Reviewed after First Round

The percentage of measurements reviewed during the second round amounted to 9.5% (1548 out of 16290) in total. For P onset, P offset, T end, QRS onset, and QRS offset it equaled respectively 8.0, 12.6, 14.6, 7.8, and 9.0%. The percentage for all leads decreased significantly with time from 12.5% in the first batch of ECGs to 10.7, 8.5, 8.9, and 7.6% in the second to fifth batches respectively (see Table 1). This downward trend with time was mainly caused by the fall in the number of P onset measurements that had to be reviewed (from 13.3% in the first batch to 3.5% in the fifth batch). The overall results were not significantly different between both data sets.

TABLE 1

PERCENTAGE OF MEASUREMENTS REVIEWED DURING SECOND ROUND

Batch	1	2	3	4	5	Total
ECGs (N)	60	72	72	83	75	362
P onset	13.3	10.3	7.8	6.3	3.5	8.0
P end	13.7	14.4	11.4	10.6	13.3	12.6
T end	13.7	16.7	17.8	14.9	9.9	14.6
QRS onset	10.9	8.3	6.4	5.9	8.4	7.8
QRS end	13.0	10.1	6.7	10.3	5.4	9.0
Average	12.5	10.7	8.5	8.9	7.6	9.5

The number of measurements reviewed in the third round remained almost constant and averaged 3.0% (N = 486). The percentages equaled 3.3, 3.6, 2.2, 3.1, and 2.7% in the first to the fifth batch, respectively. Each referee reviewed between 1.6 and 3.5% of his measurements during the so-called 4/1 review. For the five readers combined, this amounted to 1975 measurements or 12.1% of the grand total.

The fourth round analysis was performed for 340 and 363 measurements of data set 1 and data set 2, respectively. Modifications to the third round estimates were made for 66 measurements in both sets combined.

Interobserver Variability

When individual referee results were compared to the final group estimates, minor but systematic differences were observed (see Table 2). Results obtained in data set 2 were concordant with those of data set 1 (see Fig. 5), indicating that no significant bias occurred in the selection and analysis of both data sets.

TABLE 2

MEAN DIFFERENCES (IN msec) OF INDIVIDUAL REFEREE RESULTS, OBTAINED AFTER THE SECOND ROUND, COMPARED WITH FINAL REFEREE ESTIMATES FOR DATA SETS 1 AND 2 COMBINED (N = 310 × 5 LEADGROUPS)

	A		B		C		D		E	
Referee	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
P onset[a]	0.4	5.2	−0.2	5.4	−0.4	5.4	0.8	4.8	−1.2	5.6
P offset[a]	0.4	6.8	−2.0	6.0	2.2	6.8	−2.4	6.8	0.8	6.4
QRS onset	0.6	3.0	0.6	3.2	0.2	3.4	−0.2	3.2	1.2	3.6
QRS offset	−0.8	5.0	1.4	6.0	−0.4	6.0	−1.4	5.2	−2.6	7.0
T end	−6.6	15.2	−4.4	16.0	10.6	17.8	5.6	16.4	−1.2	12.6

[a] N = 261 for P wave results; cases with atrial fibrillation and some other arrhythmias were excluded.

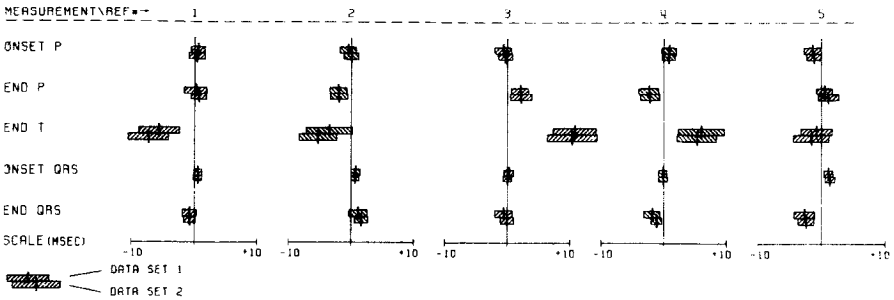

FIG. 5. CSE Data set 1 + 2 (11/12/1984). Data set 1 compared to data set 2. Bar graph of differences (in msec) between individual referee estimates and final group results. Mean differences are depicted by small vertical lines and 99% confidence intervals by horizontal bars. The long vertical lines denote zero difference. Composite leadgroup results are presented for data set 1 and 2 separately.

Mean differences were smallest for QRS and P onset, whereas they were largest for T end. One referee had systematically put the offset of the T wave 10 to 12 msec later than the other readers. The standard deviation (SD) of the differences varied around 3 msec for QRS onset. It equaled 5 to 6 msec for the end of QRS as well as for P on- and offset, whereas for T end it varied between 12 and 20 msec. Standard deviations of best and median reader results are listed in Tables 3a and b, respectively, for first and second round analysis data, the best reader being defined as having the smallest variance. It can be seen that the

TABLE 3a

FIRST ROUND RESULTS—4/1 CORRECTIONS NOT INCLUDED

Leadgroup	P wave		QRS wave		T wave
	Onset	Offset	Onset	Offset	End
Standard deviation of differences (in msec) of "best" reader with respect to final referee estimates					
I–III	5.8	7.2	5.2	7.6	17.8
aVR–aVF	5.8	7.2	5.0	7.0	16.2
V1–V3	8.0	9.2	4.0	6.0	16.6
V4–V6	7.8	7.2	3.2	6.0	17.8
XYZ	6.0	6.8	4.0	5.2	17.0
Average	6.7	7.5	4.3	6.4	17.1
Standard deviation of differences (in msec) of "median" reader with respect to final referee estimates					
I–III	6.2	7.4	5.6	9.0	20.4
aVR–aVF	6.8	8.0	5.8	8.2	23.4
V1–V3	8.2	10.8	4.2	6.4	20.6
V4–V6	8.4	9.0	3.6	7.0	21.2
XYZ	7.6	7.2	4.6	6.4	20.8
Average	7.4	8.5	4.8	7.4	21.3

TABLE 3b

SECOND ROUND RESULTS—WITH 4/1 CORRECTIONS

Leadgroup	P wave Onset	P wave Offset	QRS wave Onset	QRS wave Offset	T wave End
Standard deviation of differences (in msec) of "best" reader with respect to final referee estimates					
I–III	3.4	6.2	3.8	5.6	12.0
aVR–aVF	3.6	5.6	3.0	4.6	11.6
V1–V3	6.0	6.2	2.2	3.8	12.4
V4–V6	5.0	5.6	2.4	4.6	12.2
XYZ	4.0	4.8	3.0	4.4	14.2
Average	4.4	5.7	2.9	4.6	12.5
Standard deviation of differences (in msec) of "median" reader with respect to final referee estimates					
I–III	4.0	6.4	3.8	6.2	16.4
aVR–aVF	4.6	6.0	4.0	6.6	13.8
V1–V3	6.2	7.2	2.6	4.6	14.4
V4–V6	6.4	6.8	2.6	6.0	14.4
XYZ	4.2	5.4	3.4	5.4	17.6
Average	5.1	6.4	3.3	5.8	15.3

interactive and iterative analysis reduced the initial variance of individual referee results by 25 to 35%. Variance figures for P onset and offset determination were smallest in the peripheral leads, whereas for QRS this occurred in the precordial leads (see Tables 3a and b).

Reproducibility of Referee Results

Table 4 shows an estimate of the reproducibility of the final group estimates for the 26 ECGs which were analyzed three times during the study period. Maximal differences between any pair of the three repeat readings are listed. It can be seen that in 347 cases out of 390 (89.0%) the final estimates of QRS onset were within 4 msec. For QRS offset, P onset and P offset this number equaled respectively 76.4, 72.4, and 67.6%. The repeat readings of T end were within 20 msec 80% of the time.

With respect to the reproducibility of individual referee results, no significantly different results were obtained. Average deviations and corresponding SD were of the same order of magnitude for each referee.

Comparison of Results from 6 Channel Recordings

Reproducibility and precision of referee results could also be derived from the ECGs in which 6 channels were recorded simultaneously. Table 5 shows the differences between the final group onsets and offsets of P, QRS, and T waves in the six simultaneously recorded peripheral and precordial leads. QRS

TABLE 4

REPRODUCIBILITY OF MEDIAN REFEREE RESULTS

MAX[a] DIF	QRS onset	QRS offset	P[b] onset	P[b] offset	MAX[a] DIF	T end
0	141	121	24	10	0–2	23
2	167	138	33	45	4–6	32
4	39	39	19	16	8–10	16
6	13	38	7	8	12–14	17
8	15	12	6	6	16–18	16
≥10	15	42	16	20	≥20	26
Total	390	390	105	105	Total	130

[a] Maximum differences (in msec) between medians of three repeat readings in 26 cases.
[b] Five cases with atrial fibrillation excluded.

onset could most reliably be determined. The onset of QRS in lead group I–III differed by no more than 4 msec from the time location obtained in the simultaneously recorded, but separately analyzed, lead group aVR–aVF in 64 out of 70 cases (91.4%). For QRS offset, P on- and offset a difference of less than or equal to 4 msec was observed in 87.1, 88.2, and 80.4%, respectively. The difference for T end was less than 20 msec in 90% of the cases.

The differences between the point estimates derived from the bipolar and unipolar limb leads varied symmetrically around zero and were significantly less ($P < 0.01$) than the differences obtained between the precordial lead groups (see Table 5). P onset, P end, as well as QRS onset were often determined

TABLE 5

FREQUENCY DISTRIBUTION OF DIFFERENCES BETWEEN FINAL REFEREE ESTIMATES DERIVED FROM LEADS I TO III VERSUS aVR TO aVF (A) AND FROM V1 TO V3 VERSUS V4 TO V6 (B), WHERE THE SIX LEADS WERE RECORDED SIMULTANEOUSLY ($N = 70$)

Difference[a] (msec)	P onset[b]		P end[b]		QRS onset		QRS end		Difference[a] (msec)	T end	
	A	B	A	B	A	B	A	B		A	B
≤ −8	0	14	1	15	2	7	1	6	≤ −16	2	19
−6 −4	5	5	4	9	4	16	3	10	−14 − 8	4	8
−2 +2	40	24	33	25	54	42	56	30	− 6 + 6	49	28
+4 +6	4	4	8	—	9	5	7	15	+ 8 +14	7	7
≥ +8	2	5	5	3	1	—	3	9	≥ +16	8	8

[a] Negative differences indicate that point estimates in leadgroups aVR–aVF (A) and V4–V6 (B) were located later in time than in leadgroups I–III (A) and V1–V3 (B), respectively.
[b] ECGs with atrial fibrillation, flutter, or AV-junctional rhythm are excluded.

earlier, whereas QRS offset was often located later, in lead group V1–V3 than in V4–V6 (see Table 5).

Derived Interval Measurements

The QRS duration was slightly longer (2 to 3 msec) in the Frank XYZ leads (mean 114.1 msec, SD 28.7) and in lead group V1–V3 (mean 115.7 msec, SD 29.6) compared to lead groups I–III, aVR–aVF, and V4–V6 (mean between 111.7 and 112.7 msec) (P = NS). The P duration averaged 109.1 msec (SD 19.1) in lead group V1 to V3 and 113.2 msec (SD 18.0) in lead group I–III, whereas the results for the other lead groups were in between these values. The ranges for P duration were from 52 to 180 msec, whereas for QRS duration they ranged from 56 to 210 msec.

Figure 6 demonstrates that so-called isoelectric segments of 10 msec and more were not uncommon at the beginning as well as at the end of QRS, especially in leads I, aVR, aVL, and lead X, where it occurred in 17 to 23% of the cases. An example is illustrated in Fig. 7.

Discussion

The objectives of the CSE project are aimed at reducing the variation of measurements made by computer programs for interpreting the electrocardio-

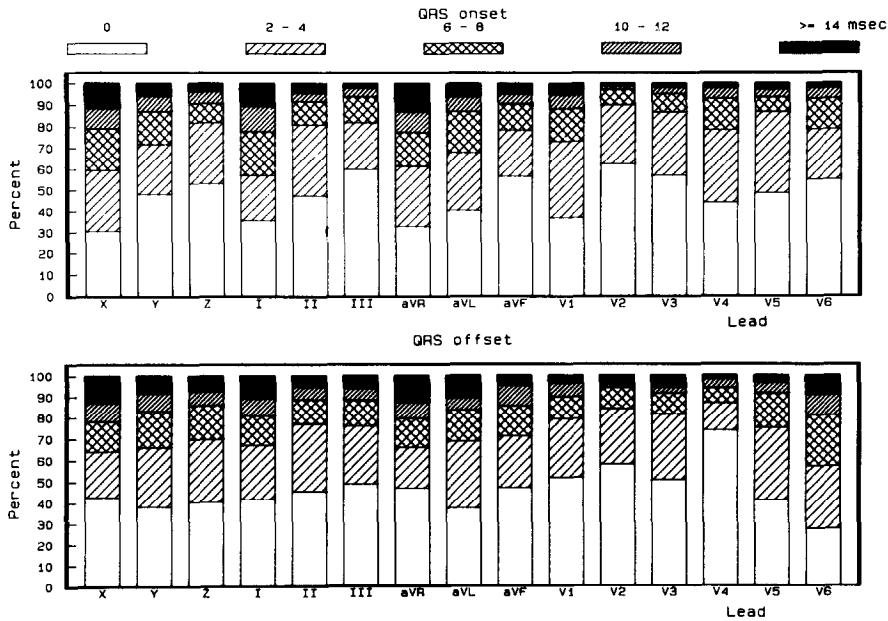

FIG. 6. Bar chart of frequency distribution of isoelectric segments found at QRS onset and offset in the 15 leads analyzed by the referees. Results are derived from the final referee estimates. N = 310 observations for each lead. (Reproduced in modified form with permission of the American Heart Association from *Circulation* **71**:523–534, 1985.)

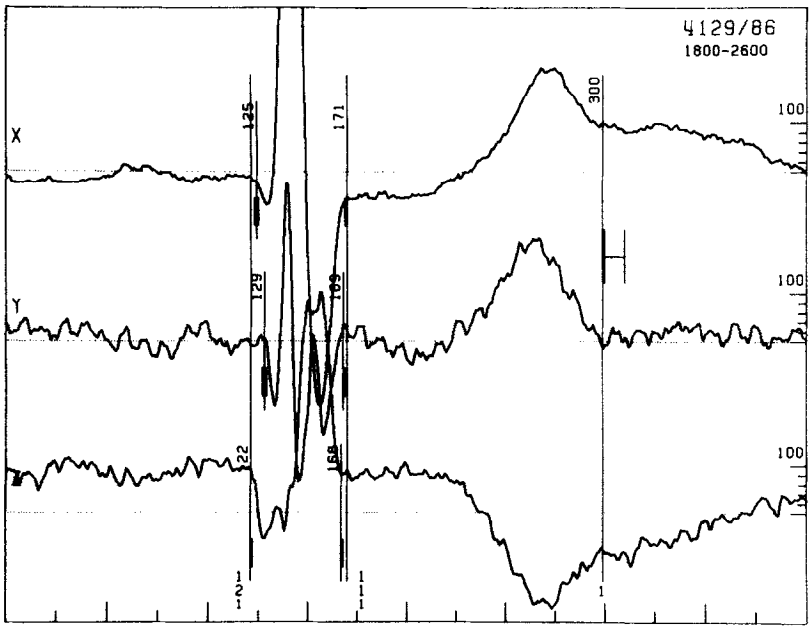

FIG. 7. Enlarged selected XYZ complex from a patient with atrial fibrillation (scale 1 cm = 100 μV; time 100 msec = 2 cm). Small vertical lines denote final point estimates determined by the referees in each lead whereas long ones refer to group onsets and offsets. The figures close to these lines refer to sample points relative to the onset of the selected beat. Note that QRS onset in lead Y apparently starts 14 msec later than in lead Z. Inclusion of this isoelectric segment results in a Q duration in lead Y of 28, exclusion of 14 msec. This is a clinically important difference for the diagnosis of inferior infarction.

gram. The elaborate reviewing process described in this paper has resulted in the establishment of a data bank with well-defined wave reference points. As might be expected, individual referee results demonstrated a certain inter- and intraobserver variability. However, this variability was lower than in former studies (*15*) due to the interactive reviewing process. Indeed, each of the four rounds of the Delphi-type reviewing process led to smaller variances. When acting as a group, the final results of the referees proved to be very stable and can be supposed to be a valid standard reference. Results for each recording of half the library (the so-called training set) have been published in a CSE Atlas (*16*) and are available on magnetic tape. These results can be used to test or refine wave recognition results of ECG analysis programs which use three simultaneously recorded leads.

A number of compromises were required for an effective implementation of the procedures used for choosing the standard reference time points for ECG wave onsets and ends (*12*). One such compromise was the use of median values of the referees after the first and second review rounds as the "correct" reference points for evaluating interval measurements by the programs. It is conceivable that occasionally the median value does not correspond to the most

accurate reference point. However, the choice of the median values was considered necessary in order to cope with the problems caused by "outliers," i.e., sporadic erratic measurements in certain difficult or noisy records. The tolerance limits for detecting outliers were different in the different rounds in order to cope with the workload. Although the referees were given filtered recordings in order to assist them in localizing the onset and offset of the P wave and the end of the T wave, they had to indicate all the fiducial points on the unfiltered high-gain recordings. Averaging techniques and signal transformations, such as spatial velocity or magnitude curves, were not provided since none of the cardiologists used such signals in routine electrocardiography. In addition, this could have biased the analysis results toward certain algorithms. The standard 12-lead ECGs of the current data base were acquired with the conventional 3-channel sequencing, which is known to be suboptimal, due to lack of orthogonality of the lead groups. However, this recording technique is used all over the world and all routinely used 12-lead computer programs, in existence at the start of the project, required such data. Of course, lack of orthogonality cannot be offered as a criticism of the analysis of XYZ data in the present study. It can be assumed that the interactive reviewing process has resulted in a learned response. However, such an effect on the readers was at most limited to 12% of the measurements, which went beyond the first round analysis. This percentage only slightly decreased over the study period, indicating that the five readers remained independent.

There were several reasons why the establishment of a data base with well-defined onsets and offsets of P, QRS, and T waves was given high priority in the CSE project. Experiences of investigators working in pattern recognition have demonstrated that several mathematical algorithms may lead to similar solutions in the average case. Some methods, however, may perform better under different conditions than others and vice versa. The use of a data base for the development of algorithms is standard practice in various fields, from automated character reading to computer-assisted chromosome and leucocyte typing. For this a local data bank and human wave recognition, usually by a single reader, has mostly been used. Furthermore, discussions with cardiologists had revealed an unwillingness of the medical community to accept strict mathematical definitions, if they had not been tested against wave recognition results derived by human reading.

From the intra- and interobserver reproducibility tests in the present study it is apparent that QRS onset is the measurement that can be made most reliably. A precision of less than 6 msec (three sample points at the 500-Hz sampling rate used in the present investigation) is attainable for QRS onset at high amplification in relatively noise-free records. Based on the results of the present study for P onset and P offset, as well as for QRS offset, a difference of 10 msec is tolerable, whereas for T end, this may be increased to 26 msec. These empirical findings are in accordance with electrophysiology. The onset of ventricular depolarization is usually a well-defined entity. QRS offset, on the contrary, is a rather arbitrary fiducial point where the final echoes of depolarization merge

imperceptibly with the early signs of repolarization. The same is true for the end of P. The T wave recovery forces move slowly and are of small magnitude. The end of T is therefore inherently less well defined. Nonetheless, in practical electrocardiology the end of QRS, as well as of the P and T waves, needs to be determined as accurately as possible. Furthermore, the interpretation of ST segment shifts is essential for the ECG interpretation of ischemia, injury, and so-called early repolarization. The P duration is a basic measurement for the diagnosis of atrial abnormalities and the QT interval is widely used for different purposes. Unless clinicians abandon the use of these primary variables, computer ECG programs must also measure them.

Data from the present study (see Table 3) demonstrate that the onset and offset determination of the P wave can most reproducibly be performed in the peripheral leads, whereas for QRS this occurs in the precordial leads.

The construction of the current data base from simultaneously recorded three lead ECGs is a primary step in the process of standardization of computer ECG measurement programs, as was recommended at the first IFIP Conference (*1*) and at the Tenth Bethesda Conference (*3*) on computer-assisted electrocardiography. While the data base cannot be guaranteed to be a representative sample of the ECG universe (the collection of all conceivable ECGs) and the number of ECGs in the data base has been constrained by practical considerations, it is probable that conclusions reached by evaluating program performance on the present data base may be generalized to cover program performance in daily routine practice. Actual results from an analysis by 9 VCG and 10 standard 12-lead ECG computer programs will be reported elsewhere (*17*).

In addition, the data base proved to be an instrument in establishing recommendations for more precise measurement rules and definitions (*18*). For example, data from the present study demonstrate that isoelectric segments of 10 msec and more are not uncommon at the beginning and end of QRS, especially in leads I, aVR, aVL, and X. From an electrophysiological point of view this is not an unexpected finding, as can be understood from the projection of the initial and terminal activation vectors on these leads. Importantly, however, these results illustrate the need for recommendations with respect to the inclusion or exclusion of these segments in the duration of the initial or terminal QRS waves, something which at present is still undefined in the literature.

Similarly, analysis of 6-channel multilead recordings demonstrated that the lead group onset of QRS, as well as of P, occurred more often slightly earlier in leads V1 to V3 than in V4 to V6. This may be explained by the fact that the initial QRS vectors, representing septal activation, are usually directed anteriorly and somewhat to the right, which is closely perpendicular to the lead axes of the lateral leads. Other factors also play a role in the recognition of wave onsets and offsets. For example, the transition between the QRS complex and the ST–T segment is often more gentle in leads V1 to V3 than in other leads, especially in cases with anterior infarction or left ventricular conduction disturbances. This may explain the fact that QRS end was often located later in the right than in the left precordial leads. The presence of a U wave can have an

important influence on the determination of T end. U waves are usually most prominent in lead group V1 to V3, which may partially explain the earlier determination of T end in these leads in some cases.

These conclusions are based on a restricted number of 6-channel recordings. The influence on ECG measurements from recordings in which all 12 (8 independent) standard leads are analyzed simultaneously, certainly merits further investigation. Studies in this direction have recently been performed by some investigators (*19*) and have also been initiated within the CSE project (*20*).

ACKNOWLEDGMENTS

This research was supported in part by the Commission of the European Communities, within the frame of its Medical and Public Health Research program under Project 82/616/EEC II.2.2 and by local and national research funding to different institutes in nine EEC Member States, including NFWO Grant 3.0050.83 of the Belgian Government.

APPENDIX

Organizational Structure: CSE Committees and Participants

CSE Steering Committee

Arnaud, P. (F), Degani, R. (I), Macfarlane, P. W. (UK), van Bemmel, J. H. (NL), Willems, J. L. (B) (Project Leader), Zywietz, Chr. (D).

CSE Board of Referees

Bourdillon, P. J. (UK), Mazzocca, G. (I), Denis, B. (F), Meyer, J. (D), Robles de Medina, E. O., and Harms, F. M. A. (NL), acting as a team with one vote and Ritsema van Eck, H. J., consultant (NL).

CSE European Working Party

B: Brohet, Chr. (Univ. of Louvain), Demeester, M. (Univ. of Brussels), Pardaens, J., De Schreye, D., and Willems, J. L. (Univ. of Leuven)

D: Dudeck, J. (Univ. of Giessen), Meyer, J., and Michaelis, J. (Univ. of Mainz), Pöppl, S. J. (Institute Medical Data Processing, Munchen), Zywietz, Chr. (Univ. of Hannover)

DK: Damgaard Andersen, J. (Univ. of Copenhagen)

F: Arnaud, P. (INSERM U121 Lyon), Denis, B. (Univ. of Grenoble), Rubel, P. (INSA, Lyon)

G: Moulopoulos, S. (Univ. of Athens), Skordalakis, E. (NCR Democritos, Attiki)

I: Dalla Volta, S. (Univ. of Padova), Degani, R. (Ladseb CNR, Padova), Mazzocca, G. (Univ. of Pisa)

Ire: Graham, I., and Reardon, B. C. (Univ. of Dublin)

NL: van Bemmel, J. H., and Talmon, J. L. (Univ. of Amsterdam), Harms, F. M. A., and Robles de Medina, E. O. (Univ. of Utrecht), Ritsema van Eck, H. J. (Univ. of Rotterdam)

UK: Bourdillon, P. J. (Univ. of London), Macfarlane, P. W. (Univ. of Glasgow).

Consultants

Bailey, J. J. (NIH) and Pipberger, H. V. (George Washington Univ.–USA), Rautaharju, P. M. (Univ. of Dalhousie–Canada).

Non-European Participants

USA: Bonner, R. (IBM), Doue, J. (Hewlett-Packard), Michler, K. (Telemed)
Canada: Rautaharju, P. M., Macinnis, P. (Univ. of Dalhousie)
Japan: Okajima, M., Okamoto, N., Yokoi, M. (Univ. of Nagoya), Ohsawa, M. (Fukuda Denshi).

CSE Coordinating Center

Division Medical Informatics, University of Leuven, Belgium.

REFERENCES

1. ZYWIETZ, CHR., AND SCHNEIDER, B. "Computer Applications in ECG and VCG Analysis," p. 271. North-Holland, Amsterdam, 1973.
2. VAN BEMMEL, J. H., AND WILLEMS, J. L. "Trends in Computer Processed Electrocardiograms," p. 437. North-Holland, Amsterdam, 1977.
3. RAUTAHARJU, P. M., ARIET, M., PRYOR, T. A., ARZBAECHER, R. C., BAILEY, J. J., BONNER, R., *et al.* Task Force III: Computers in diagnostic electrocardiography. *Amer. J. Cardiol.* **41,** 158 (1978).
4. BAILEY, J. J. The future of gold standards and computerized electrocardiography. *In* "Computerized Interpretation of the Electrocardiogram" (G. D. Tolan and T. A. Pryor, Eds.), p. 229. Engineering Foundation, New York, 1980.
5. WOLF, H. K., AND MACFARLANE, P. W. "Optimization of Computer ECG Processing," p. 346. North-Holland, Amsterdam, 1980.
6. WILLEMS, J. L., AND PARDAENS, J. Differences in measurement results obtained by four different ECG computer programs. *In* "Computers in Cardiology" (H. G. Ostrow and K. L. Ripley, Eds.), p. 115. IEEE Computer Society, Long Beach, Calif., 1977.
7. WILLEMS, J. L. A plea for common standards in computer aided ECG analysis. *Comput. Biomed. Res.* **13,** 120 (1980).
8. WILLEMS, J. L., ARNAUD, P., DEGANI, R., MACFARLANE, P. W., VAN BEMMEL, J. H., AND ZYWIETZ, CHR. Protocol for the concerted action project "Common Standards for Quantitative Electrocardiography," 2nd R&D programme in the field of Medical and Public Health Research of the EEC (80/344/EEC), CSE Ref. 80-06-00, p. 152. ACCO Publ., Leuven, 1980.
9. The CSE European Working Party (Willems, J. L., Arnaud, P., van Bemmel, J. H., *et al.*). Common Standards for Quantitative Electrocardiography. The CSE Pilot Study. *In* "Proceedings of Medical Informatics Europe 81" (F. Gremy, P. Degoulet, B. Barber, R. Salamon, Eds.), p. 319. Springer-Verlag, Heidelberg, 1981.
10. WILLEMS, J. L. Common Standards for Quantitative Electrocardiography, 2nd Progress Report. CSE Ref. 82-11-20, p. 246. ACCO Publ., Leuven, 1982.
11. The CSE European Working Party (Willems, J. L., Arnaud, P., van Bemmel, J. H., *et al.*). Common Standards for Quantitative Electrocardiography. CSE Project Phase One. *In* "Computers in Cardiology" (K. L. Ripley, Ed.), p. 69. IEEE Computer Society, Long Beach, Calif. 1982.
12. BOURDILLON, P. J., DENIS, B., HARMS, F. M. A., MAZZOCCA, G., MEYER, J., ROBLES DE

MEDINA, E. O., RITSEMA VAN ECK, H. J., AND WILLEMS, J. L. European experience in the standardization of measurements and of definitions of the electrocardiogram. *In* "Computerized Interpretation of Electrocardiograms VII" (M. Laks, Ed.), p. 9. Engineering Foundation, New York, 1982.

13. MACFARLANE, P. W., AND WILLEMS, J. L., on behalf of the CSE Working Party. The CSE Project: Progress as viewed by the cooperating centers. *In* "Computer Interpretation of Electrocardiograms VIII" (R. Selvester, Ed.), p. 293. Engineering Foundation, New York, 1984.
14. DALKEY, N. Analysis from a group opinion study. Rand McNally, Chicago, p. 541. Futures, Dec. 1969.
15. FISCHMANN, E., COSMA, J., AND PIPBERGER, H. V. Beat to beat and observer variation of the electrocardiogram. *Amer. Heart J.* **75,** 465 (1968).
16. WILLEMS, J. L. CSE Atlas—Referee Results First Phase Library Data Set One, CSE Ref. 83-05-13, p. 655. ACCO Publ., Leuven, 1983.
17. WILLEMS, J. L., ARNAUD, P., VAN BEMMEL, J. H., *et al.* Assessment of the performance of electrocardiographic computer programs with the use of a reference data base. *Circulation* **71,** 523 (1985).
18. The CSE Working Party. Common Standards for Quantitative Electrocardiography. Recommendations for measurement standards (to be published).
19. BORTOLAN, G., CAVAGGION, C., AND DEGANI, R. T. A comparison of ECG measurements derived from 3, 6 and 12 simultaneous leads. *In* "Computers in Cardiology" (K. L. Ripley, Ed.), p. 269. IEEE Computer Society, Long Beach, Calif., 1983.
20. WILLEMS, J. L. Common Standards for Quantitative Electrocardiography, 4th Progress Report, p. 277. ACCO Publ., Leuven, 1984.