ORIGINAL CONTRIBUTION
cardiac arrest
data interpretation, statistical
emergency medical services
emergency medical technicians
ventricular fibrillation

Inter-Rater Agreement of Paramedic Rhythm Labeling

Ronald G Pirrallo, MD, MHSA*
Robert A Swor, DO, FACEP*
Ronald F Maio, DO, MS, FACEP†

From the Department of Emergency Medicine, William Beaumont Hospital, Royal Oak, Michigan; and Section of Emergency Medicine, Department of Surgery, University of Michigan, Ann Arbor.†*

Received for publication August 18, 1992. Revision received December 14, 1992. Accepted for publication January 7, 1993.

Presented at the Society for Academic Emergency Medicine Annual Meeting in Toronto, Ontario, Canada, May 1992.

This study was supported by the SAEM Physio-Control EMS Fellowship.

Study hypothesis: Substantial inter-rater agreement is present in the labeling by paramedics of ventricular fibrillation and asystolic rhythms.

Design: Prospective, cross-sectional study.

Type of participants: One hundred five practicing paramedics from nonvolunteer agencies who are advanced cardiac life support certified.

Methods: Five static cardiac arrest rhythm strips, classified by Cummins' average peak amplitude method, were arranged into five different orders of presentation and placed into five booklets. The paramedics were instructed to label each rhythm ventricular fibrillation or asystole based on rhythm recognition, not on treatment plan.

Results: The overall κ value for labeling the five rhythms was .63, indicating a moderate degree of inter-rater agreement. However, as the rhythm's amplitude decreased, the amount of inter-rater agreement also decreased. When the amplitude was approximately 1 mm, agreement was no different than chance; the proportion of paramedics labeling the rhythm ventricular fibrillation was .46 (95% confidence interval, .36, .56). Only a flat line (0 mm) demonstrated perfect inter-rater agreement, with no paramedic labeling the rhythm ventricular fibrillation.

Conclusion: Inter-rater agreement of ventricular fibrillation rhythm labeling by paramedics in this emergency medical services system was amplitude dependent. An analysis of ventricular fibrillation rhythm data that does not address the degree of inter-rater agreement of rhythm labeling cannot ensure uniform reporting of out-of-hospital cardiac arrest data.

[Pirrallo RG, Swor RA, Maio RF: Inter-rater agreement of paramedic rhythm labeling. *Ann Emerg Med* November 1993;22:1684-1687.]

INTRODUCTION

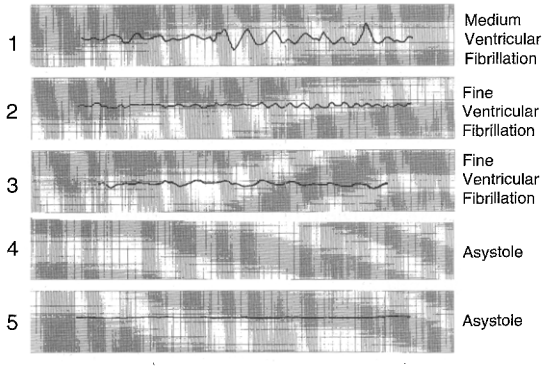
Many emergency medical services (EMS) systems collect, analyze, and report cardiac arrest data based on the label of the initial arrest rhythm.[1-3] However, no published cardiac arrest study has addressed the degree of inter-rater agreement of that rhythm label. In the Oakland County (Michigan) EMS system, the paramedics label the patient's cardiac rhythm based on their advanced cardiac life support (ACLS) training and indicate their interpretation on the prehospital care runsheet. Cardiac arrest information then is abstracted from the runsheet and entered into a data base. We believe this is a common convention in EMS data collection.

If inter-rater agreement is not present in the labeling by a paramedic of ventricular fibrillation and asystolic rhythms, a significant misclassification error is introduced into the reporting of out-of-hospital cardiac arrest data. Kelsey and colleagues stated, "In the conduct of epidemiologic research, measurement error is potentially a major problem that may invalidate the results of otherwise well-designed studies."[4] The purpose of this study was to describe the amount of inter-rater agreement in rhythm labeling by paramedics of ventricular fibrillation and asystole; we hypothesized that substantial inter-rater agreement was present.

MATERIALS AND METHODS

One hundred five practicing paramedics from three nonvolunteer advanced life support (ALS) agencies in Oakland County, Michigan, participated in the study. All paramedics have a minimum of 600 hours of paramedic training and are ACLS certified.

Figure.
The five study rhythms



Five examples of cardiac arrest rhythm strips were used (Figure and Table 1).[5,6] Each rhythm strip was 3.5 to 4.5 seconds in length. The rhythms were classified by Cummins' average peak amplitude method into three categories.[7] Each strip was photocopied, enlarged to the dimensions of 18.5 × 2.5 cm, and centered on a single piece of standard white 8 1/2 × 11 in. office paper.

The rhythm strips were arranged into five different orders of presentation and placed into five booklets marked A through E. The booklets were stacked in a repeated, alphabetical order (ie, A, B, C, D, E, A, B, C, D, E, ...) and distributed to the paramedics on entering the conference room. The paramedics entered the conference room in no predetermined fashion and were not allowed to sit next to anyone with a similarly marked booklet. Each booklet contained the following instructions: The following rhythms are of five patients who are in cardiac arrest, pulseless, and apneic. Assume that all equipment is appropriately connected and functioning properly. The rhythm is identical in all leads. You may choose only ONE of the following responses: asystole or ventricular fibrillation. Circle only one response. Once you have identified the rhythm, do not return to that patient.

The paramedics were instructed to choose their responses based on rhythm recognition, not on treatment plan. They were expected to label the rhythm using their current, working definitions of ventricular fibrillation and asystole. They were told that this project was a survey, not a test.

The survey was administered at each ALS agency's regularly scheduled fall quarterly meeting and was unannounced. Each paramedic was given ten minutes to complete the survey. The survey was proctored by the agency's supervisor.

Table 1
Classification of rhythms

Strip	Term for the Rhythm*	Average Peak Amplitude (mm)*	Source of Rhythm Strip (Reference)
1	Medium ventricular fibrillation	3 to < 7	5
2	Fine ventricular fibrillation	1 to < 3	5
3	Fine ventricular fibrillation	1 to < 3	6
4	Asystole	< 1	5
5	Asystole	0	5

*Based on Cummins' average peak amplitude method.[7]

An overall κ for multiple ratings per subject was performed on the five rhythm strips.[8] κ also was calculated for selected pairs of these strips. The proportion of the number of rhythm strips labeled ventricular fibrillation and the 95% confidence interval were calculated for each rhythm strip.

RESULTS

One hundred five paramedics were surveyed, and all of the paramedics answered all of the questions.

The overall κ value for labeling of the five rhythm strips was .63, indicating a moderately good degree of inter-rater agreement. The κ value corresponds to the amount of agreement between the raters, not whether their answer agrees with an external gold standard. When two raters agree only at the chance level, κ = 0; when two raters agree perfectly, κ = 1. Landis and Koch have categorized κ values and suggest the following corresponding degrees of agreement.[8] A κ value of less than .40 suggests poor agreement, a κ value of .40 to .75 suggests fair to good agreement, and a κ value of more than .75 suggests excellent agreement. In our study design, κ characterized the variability of the multiple raters' answers only when comparing two or more rhythm strips. Therefore, only the most similar appearing rhythm strips were selected for κ analysis. A κ value cannot be calculated on a single rhythm strip or if the numerator is 0, as in strip 5.

Almost-perfect inter-rater agreement was present for strips 1 and 2 (Table 2). However, as the rhythm's amplitude decreased, the amount of inter-rater agreement also decreased. For strips 3 and 4, κ equaled .13, indicating poor agreement. When the amplitude was approximately 1 mm (strip 4), the probability of a paramedic labeling the rhythm ventricular fibrillation was no different than chance. Only a flat line, 0 mm (strip 5), demonstrated perfect agreement, with no paramedic labeling the rhythm ventricular fibrillation.

Table 2
Results of paramedic rhythm labeling

Strip	Average Peak Amplitude (mm)	No. Labeled Ventricular Fibrillation	No. Labeled Asystole	Proportion Labeled Ventricular Fibrillation (95% CI)	κ
1	3 to < 7	104	1*	0.99 (.97, 1.0)	≈1
2	1 to < 3	104	1*	0.99 (.97, 1.0)	
3	1 to < 3	85	20	0.81 (.73, .89)	.13
4	< 1	48	57	0.46 (.36, .56)	
5	0	0	105	0 (0, .04)	NA
All rhythms					.63

*Not the same individual.

DISCUSSION

The problem of inter-rater agreement, often discussed as observer error, has been recognized and well described in clinical medicine. Thirty years ago, Garland[9] and Knox[10] summarized 12 investigations that examined observer error in the interpretation of clinical and laboratory procedures. These investigations ranged from the taking of medical histories to the counting of erythrocytes to the interpretation of ECGs. Pozen and colleagues[11] and Jarmon and Yesalis[12] described the potential effect of this problem during prehospital care in the identification and treatment of arrhythmias. Our study documents that observer error is present in paramedics' labeling of low-amplitude rhythms.

Moderately good agreement existed in the overall labeling of these rhythms (κ = .63). This occurred because the κ statistic is a weighted average of the positive ratings from each strip. The paramedics' extremely high degree of inter-rater agreement on strips 1, 2, and 5 was averaged with their poor agreement on strips 3 and 4. This summation drove the overall value of κ.

This study was not intended to nor does it document a lack of expertise in the paramedics' treatment of cardiac arrest patients. Likewise, this study's design and statistic did not compare the paramedics' interpretation of cardiac rhythms with external gold standard definitions such as those of Cummins and colleagues.[7,13] This study documents a common pitfall in all research: lack of inter-rater agreement.

Valenzuela and colleagues found a sixfold difference in survival rate from out-of-hospital cardiac arrest depending on the combination of case and survival definition selected.[14] Our study suggests that this number could be even larger. The actual effect that the lack of inter-rater agreement in rhythm labeling has on cardiac arrest outcome reporting depends on the number of near-1-mm-amplitude rhythms that occur in that EMS system. If 1-mm-amplitude rhythms are common presenting cardiac arrest rhythms, the denominator, the number of cases identified as ventricular fibrillation, can vary by as much as 50%. If 1-mm-amplitude rhythms are rare, this effect may be minimal. The rate of low-amplitude ventricular fibrillation cardiac arrest rhythms is unknown. Our paramedics were allowed only a single response, ventricular fibrillation or asystole. This was intended to minimize the variability of the paramedics' answers to best fit the Utstein style of reporting out-of-hospital cardiac arrest data and increase their likelihood of agreement.[13]

Maio and Burney have shown that standard definitions alone do not limit inconsistencies in abstracting runsheet

data.[15] The use of decision rules in addition to standard definitions is needed to enhance the agreement between individual abstractors. This implies that even the Utstein style definitions of ventricular fibrillation and asystole alone will not eliminate inconsistencies in reporting survival rates from out-of-hospital cardiac arrest.[13] The validity of the results of any retrospective analysis of ventricular fibrillation data that did not address the degree of inter-rater agreement of rhythm labeling should be questioned.

This study had several limitations. The inherent loss of detail in enlargement and photocopying of these rhythms must be considered. An attempt to control for this was made by using the original rhythm strips for all copies. Although all paramedics were ACLS certified, controlling for further paramedic education and experience was not done. Also, this study used static rhythms rather than dynamic monitor readings. The use of dynamic monitor rhythms may better approximate prehospital conditions.

CONCLUSION

Inter-rater agreement of ventricular fibrillation rhythm labeling by paramedics in this EMS system was amplitude dependent. An analysis of ventricular fibrillation rhythm data that does not address the degree of inter-rater agreement of rhythm labeling cannot ensure uniform reporting of out-of-hospital cardiac arrest data. Studies that use paramedic ventricular fibrillation rhythm labeling alone, without verification mechanisms, may be nonreproducible and therefore invalid.

REFERENCES

1. Eisenberg MS, Horwood BT, Cummins RO, et al: Cardiac arrest and resuscitation: A tale of 29 cities. *Ann Emerg Med* 1990;19:179-186.

2. Hargarten KM, Stueven HA, Waite EW, et al: Prehospital experience with defibrillation of coarse ventricular fibrillation: A ten-year review. *Ann Emerg Med* 1990;19:157-162.

3. Becker LB, Ostrander MP, Barrett J, et al: Outcome of CPR in a large metropolitan area—Where are the survivors? *Ann Emerg Med* 1991;20:355-361.

4. Kelsey JL, Thompson WD, Evans AS: Measurement error, in *Methods in Observational Epidemiology.* New York, Oxford University Press, 1986, p 285.

5. Marriott HJ, Myerburg RJ: Recognition of cardiac arrhythmias and conduction disturbances, in Hurst JW (ed): *The Heart,* ed 7. New York, McGraw Hill, 1990, p 529, fig 31-56.

6. American Heart Association: *Textbook of Advanced Cardiac Life Support,* ed 2. Dallas, Texas, AHA, 1987, p 78, fig 53.

7. Cummins RO, Stults KR, Haggar B, et al: A new rhythm library for testing automatic external defibrillators: Performance of three devices. *J Am Coll Cardiol* 1988;11:596-602.

8. Fleiss JL: *Statistical Methods for Rates and Proportions,* ed 2. New York, John Wiley & Sons, chap 13, 1981, p 212-236.

9. Garland LH: The problem of observer error. *Bull NY Acad Med* 1960;36:570-584.

10. Knox GS: How often are we wrong? *Oklahoma State Med Assoc* 1964;57:494-500.

11. Pozen MW, D'Agostino RB, Sytkowski PA: Effectiveness of a prehospital medical control system: An analysis of the interaction between emergency room physician and paramedic. *Circulation* 1981;63,2:442-447.

12. Jarmon RG, Yesalis CE: Provider performance in the recognition and treatment of telemetered electrocardiogram patterns. *JACEP* 1976;5:971-974.

13. Cummins RO, Chamberlain DA, Abramson NS: Recommended guidelines for uniform reporting of data from out-of-hospital cardiac arrest: The Utstein style. *Ann Emerg Med* 1991;20:861-874.

14. Valenzuela TD, Spaite DW, Meislin HW, et al: Case and survival definitions in out-of-hospital cardiac arrest: Effect on survival rate calculation. *JAMA* 1992;267:272-274.

15. Maio RF, Burney RE: Improving reliability of abstracted prehospital care data: Use of decision rules. *Prehosp Disaster Med* 1991;6:15-20.

The authors thank the paramedics of the Oakland County EMS system, the William Beaumont Hospital library staff, Dr Raymond Jackson, MD, FACEP, and Dr Judith Tintinalli, MD, MS, FACEP, for their efforts in completing this study.

Address for reprints:

Ronald G Pirrallo, MD, MHSA
Medical College of Wisconsin
Department of Emergency Medicine
Box #204
8700 West Wisconsin Avenue
Milwaukee, Wisconsin 53226