Bias in a prehospital esophageal detector device trial: lessons learned

P. Richard Verbeek, MD; Glen Bandiera, MD; Brian Morris; Dennis St. Pierre

ABSTRACT

Objectives: Our goals were to determine whether selection bias occurred in a prehospital study comparing an esophageal detector device (EDD) to a disposable capnometer for detecting esophageal intubation, and to determine whether such a bias would have changed the study's conclusions about EDD effectiveness.

Methods: In a study of patients requiring prehospital intubation, we determined the sensitivity, specificity and predictive values of the EDD for detecting esophageal intubation. We then compared intubation success rate in patients who were enrolled in the study (n = 129) to that in eligible patients who were excluded from it (n = 107). After finding that the incidence of failed intubation was higher in the "excluded" group, we used sensitivity and specificity parameters derived from the study population to assess whether EDD test characteristics would differ in studied vs. excluded patients.

Results: The first intubation attempt was successful in 125 of 129 study patients and 76 of 107 excluded patients (97% vs. 71%, p = 0.03), confirming the presence of selection bias. The negative predictive value of the EDD for esophageal intubation was 98% in the study cohort and would have been 77% in patients like those excluded (i.e., difficult intubation cases).

Conclusion: The high "first attempt" intubation success rate seen in this study was due to selective exclusion of failed intubations. This selection bias led to a clinically important overestimation of the EDD's negative predictive value. Bias may substantially alter the estimations of test accuracy reported in scientific studies. To reduce the chance of unrecognized selection bias in studies of diagnostic tests, investigators must determine whether recruited subjects resemble patients in whom the test will ultimately be used.

RÉSUMÉ

Objectifs : Établir la présence d'un préjugé de sélection lors d'une étude en milieu préhospitalier comparant un vérificateur de position oesophagienne (VPO) avec un capnographe jetable pour la détection d'une intubation oesophagienne et déterminer si un tel préjugé aurait modifié les conclusions de l'étude au sujet de l'efficacité du VPO.

Méthodes : Lors d'une étude auprès de patients nécessitant une intubation préhospitalière, nous avons déterminé la sensibilité, la spécificité et les valeurs prédictives du VPO pour déceler les intubations oesophagiennes. Nous avons ensuite comparé le taux de succès des intubations chez des patients inclus dans l'étude (n = 129) à celui chez des patients éligibles exclus de l'étude (n = 107). Après avoir constaté que l'incidence des intubations ratées était plus élevée parmi le groupe «exclus», nous avons eu recours aux paramètres de sensibilité et de spécificité dérivés de la popu-

Dr. Verbeek is with the Department of Emergency Services, Sunnybrook and Women's College Health Science Centre, University of Toronto; Dr. Bandiera is with the Department of Emergency Services, St. Michael's Hospital, University of Toronto; and Mr. Morris and Mr. St. Pierre are Level III Paramedics with the Toronto Ambulance Service, Toronto, Ont.

Received: Nov. 9, 1999; final submission received: Apr. 1, 2000; accepted: Apr. 9, 2000

This article has been peer reviewed.

lation à l'étude pour évaluer si les caractéristiques du test du VPO seraient différentes entre les patients à l'étude et les patients exclus.
Résultats : La première tentative d'intubation fut couronnée de succès chez 125 des 129 patients à l'étude et chez 76 des 107 patients exclus (97% vs 71%, p = 0,03), confirmant la présence d'un préjugé de sélection. La valeur prédictive négative du VPO pour l'intubation oesophagienne était de 98% pour la cohorte à l'étude et aurait été de 77% chez les patients qui avaient été exclus (i.e., les cas d'intubation difficile).
Conclusions : Le taux élevé de succès de l'intubation lors de la «première tentative» constaté lors de cette étude était attribuable à l'exclusion sélective des intubations ratées. Ce préjugé de sélection entraîna une surestimation cliniquement importante de la valeur prédictive négative du VPO. Les préjugés peuvent modifier substantiellement l'estimation de l'exactitude des tests décrits dans des études scientifiques. Afin de diminuer le risque d'un préjugé de sélection non identifié lors de l'étude d'épreuves diagnostiques, les chercheurs doivent déterminer si les sujets recrutés ont les mêmes caractéristiques que les patients chez qui cette épreuve sera éventuellement utilisée.

Key words: emergency medical services; clinical trial; intubation; selection bias

Introduction

Diagnostic test results can greatly influence patient care. Whereas emergency physicians have relied on diagnostic tests to guide clinical and therapeutic decisions for years, less emphasis has been placed on the use of diagnostic tests in the prehospital care environment. However, as the sophistication of paramedicine evolves, it is inevitable that prehospital diagnostic test use will increase.

One vexing problem in prehospital care relates to the confirmation of endotracheal tube (ETT) placement. The traditional practice of visualizing the ETT passing through the vocal cords, then performing other physical assessments to confirm proper placement, has been labelled a "fool's gold standard."[1] At a recent prehospital care consensus conference,[2] waveform or colorimetric capnometry, complemented by an endotracheal detector device (EDD), was proposed as the new gold standard for verifying ETT placement. But while capnometry is a widely accepted standard,[1–4] the EDD's role remains controversial.[5–12]

Diagnostic tests are often introduced to clinical practice before adequate appraisal,[3,13] and studies that evaluate test utility can be affected by several types of bias, including verification bias, interpretation bias, selection bias, and the absence of a gold standard.[4,13,14] These and other forms of bias generate flawed test performance data and limit the validity of study conclusions about test utility.

In a recent prehospital study intended to assess the usefulness of the EDD in detecting esophageal intubation, we noted an unexpectedly high rate of "first attempt" intubation success. We suspected that selection bias might be responsible for this observation and suspended the study to investigate this possibility.

Methods

The EDD study

Between May 1995 and June 1996, in a large urban emergency medical services (EMS) system, we prospectively compared the effectiveness of an EDD (60-mL syringe attached to a 7.0-mm ETT adapter) to that of a disposable colorimetric capnometer (Easy Cap™, Nellcor).[8] At the time, the latter device was felt to be the best prehospital reference ("gold") standard[15,16] for detecting esophageal intubation of non-cardiac arrest patients. Cardiac arrest patients were excluded from the study due to the lack of a valid out-of-hospital reference standard to ascertain ETT placement.[5] The trial was approved by our institutional research ethics board.

Following the intubation of an eligible patient, the non-intubating paramedic used an EDD and attempted to aspirate 30 mL of air, after which the plunger was released and observed for rebound. Impeded aspiration with plunger rebound was interpreted to indicate esophageal tube placement. Without knowledge of the EDD result, the intubating paramedic then attached the capnometer and recorded its result. Any colour change of the capnometer filter paper indicated tracheal placement of the ETT, whereas no colour change indicated esophageal placement. The EDD and capnometry results were recorded on separate data sheets, and each paramedic was required to sign a statement asserting that they had no knowledge of the other test findings when interpreting their own test.

Assuming that the "first attempt" intubation success rate would be 70% and that EDD sensitivity, specificity and negative predictive value would be at least 95%,[17,18] we calculated that 200 subjects were required to complete the trial.

An analysis of the initial 129 subjects showed that "first attempt" intubation success was 97% rather than the expected 70%. We recognized that many patients had been excluded from the study and considered the possibility of selection bias; therefore, we designed a retrospective survey to determine whether recruited subjects were different from those who fulfilled study eligibility criteria but were excluded.

The retrospective survey

Three paramedics who were trained to abstract data from ambulance call reports explicitly, reviewed all paramedic patient encounters between September and December 1995, and documented the reason for intubation and the "first attempt" intubation success rate. Reason for intubation was dichotomized into primary respiratory failure (e.g., pulmonary edema, chronic obstructive pulmonary disease, pneumonia) or airway protection (e.g., trauma, overdose, stroke). Our paramedics are taught to document an intubation attempt each time a laryngoscope is passed beyond the teeth or an ETT is passed into the nares with the intent of intubating; therefore, "first attempt" success was recorded when only one intubation attempt was documented on the ambulance call report.

Results

The retrospective survey identified 143 patients who fulfilled EDD study eligibility criteria. Of these, 36 (25%) had been enrolled in the study and 107 (75%) excluded from it. Table 1 shows that excluded subjects (n = 107) were similar to enrolled subjects (n = 129) with respect to demographics, reason for intubation and route of intubation, but that "first attempt" intubation success rate and overall intubation success rate were significantly lower in excluded patients. These differences suggested there was a bias against enrolling subjects who could not be intubated on the first attempt (or at all), and that the diagnostic parameters calculated in this study might be different if EDDs are used to evaluate "real world" prehospital intubations.

Table 2 illustrates actual EDD performance characteristics in the study cohort (n = 129), where the "first attempt" intubation success rate was 97%. In this cohort, the EDD identified one of 4 esophageal intubations identified by capnometry, thus had a sensitivity of 25%. Table 3 shows the expected EDD performance characteristics in a hypothetical cohort of 129 subjects where the "first attempt" intubation success is 71% (i.e., patients like those in the excluded group). Because it is generally accepted[3] that test sensitivity and specificity remain constant when tests are used in populations with differing prevalence of the target condition (e.g., esophageal intubation), we used sensitivity and specificity figures from our EDD study to calculate the predictive values illustrated in Table 3. The key difference shown in Table 3 is that the EDD's negative predictive value falls from 98% in the study cohort (where the prevalence of failed intubation was 3%) to 77% in the "excluded" cohort (where the prevalence of failed intubation was 29%). The implications of this are outlined below.

Discussion

Despite our attempts to perform a careful analysis of EDD effectiveness, an unexpected selection bias compromised our results. Selection bias occurs when the subjects studied are different from those targeted by the trial, hence are not representative of subjects to whom the test will be applied in the clinical setting.[13,14] In this study, paramedics selectively enrolled patients whom they were able to intubate easily. In the real world, EDD accuracy is most important in patients who are difficult to intubate and in whom there is doubt about tube position. This means that in our study the subjects who would have been most important to assess EDD accuracy were least likely to be recruited into the trial.

Table 1. Comparison of control and esophageal detector device (EDD) groups

	Control group (n = 107)	EDD group (n = 129)	p value
Reason for intubation			
Respiratory failure	60 (56)	64 (50)	0.49
Airway protection	47 (44)	65 (50)	0.49
Intubation route			
Nasal	70 (65)	96 (74)	0.83
Oral	21 (20)	33 (26)	0.83
Unable to intubate	16 (15)	0	<0.01
"First attempt" success	76 (71)	125 (97)	0.03
"Overall" success	91 (85)	129 (100)	<0.01

Results are shown as no. (and %).

Table 2. Capnography vs. EDD results in 129 patients (intubation success rate = 97%)

Esophageal intubation? (EDD result)	Esophageal intubation? (capnography result)		
	Present	Absent	Total
Present	1	0	1
Absent	3	125	128
Total	4	125	129

Sensitivity = 25% (95% confidence interval [CI]), 0%–67%; positive predictive value (PPV) = 100%; specificity = 100%; negative predictive value = 98%

https://doi.org/10.1017/S1481803500004875 Published online by Cambridge University Press

Bias in EDD research

Other forms of bias are common in trials evaluating diagnostic tests,[4,13,14,19,20] and two such biases are illustrated by recent EDD studies.[9,10] Interpretation bias occurs when knowledge of one diagnostic test influences the interpretation of another.[8,20,21] For example, in the current study, if the paramedic applying the EDD was aware of the capnometry result, this could alter his or her interpretation of the EDD result. Interpretation bias can be avoided by blinding both evaluators to the results of the alternate test.[22] Another type of bias may occur when an inadequate reference standard is used (which allows patient misclassification) or if the experimental test is used to help establish the "true" diagnosis and thereby acts as its own reference standard.[19]

Two large prospective prehospital EDD trials have been previously published.[9,10] In both, the reference standard used to determine "true" ETT placement was clinical evaluation — an inadequate standard[1,2,5] that may have allowed misclassification errors. In one of the studies,[9] Marley used the results of clinical evaluation and the EDD to determine "true" ETT placement. This may have introduced an interpretation bias if knowledge of the EDD result influenced the interpretation of clinical findings, or if knowledge of clinical findings influenced interpretation of the EDD. In the other study,[10] paramedics were told not to alter tube placement based on EDD results, but it is likely that knowledge of the EDD result could have biased the interpretation of the paramedics' clinical evaluation regarding ETT placement.

Perhaps because of study biases (Table 4), these investigators drew very different conclusions regarding the prehospital performance of the EDD. Marley[9] concluded that the EDD was 100% sensitive and 78% specific (identifying 17 of 17 esophageal intubations and 75 of 88 tracheal intubations). In contrast, Pelucio[10] concluded that the EDD was 50% sensitive and 99% specific (identifying 5 of 10 esophageal intubations and 156 of 158 tracheal intubations). It is difficult to know how to interpret these conflicting results, and they leave us uncertain about the true value of EDD.

Table 3. Expected* capnography results vs. EDD results if intubation success rate = 71%

Esophageal intubation? (EDD result)	Esophageal intubation? (capnography result)		
	Present	Absent	Total
Present	9	0	9
Absent	28	92	120
Total	37	92	129

*Table 3 is based on the following assumptions, derived from Table 2: n = 129; sensitivity = 25%; specificity = 100%. Calculated PPV = 100%; calculated negative predictive value = 77%.

Predictive value and prevalence

When basing treatment decisions on diagnostic tests, it is important to understand the concept of predictive value.[19,23] Positive predictive value (PPV) answers the question: "If the test is positive, what is the probability that the patient has the condition of interest (e.g., esophageal intubation)?" Negative predictive value (NPV) answers the question: "If the test is negative, what is the probability that the patient is free from the condition of interest?" Excellent PPV is important when a test is used to initiate treatments that are associated with morbidity or mortality (e.g., thrombolysis). Excellent NPV is important when a test is used to rule out conditions that are associated with morbidity or mortality (e.g., esophageal intubation).

In this situation, NPV answers the question, "If the EDD result indicates tracheal intubation, how likely is a tracheal intubation to have occurred?" NPV therefore defines the ability of a paramedic using an EDD to rule out esophageal

Table 4. The potential role of bias in prehospital EDD trials

Trial	Selection bias	Inadequate reference standard	Interpretation bias
Marley[9]	Yes: Convenience sample	Yes: Clinical exam used as reference standard	Yes: EDD result influenced reference standard determination
Pelucio[10]	Yes: Only 57% of eligible patients enrolled	Yes: Clinical exam used as reference standard	??: Unclear whether paramedics blinded to alternate test result
Current study	Yes: Only 25% of eligible patients enrolled	No: Capnography used as reference standard	Possible: Attempts made to blind paramedics to alternate test result

https://doi.org/10.1017/S1481803500004875 Published online by Cambridge University Press

intubation. In our study, the negative predictive value of the EDD for esophageal intubation was 97%. Consequently, a paramedic acting on a "negative" EDD study would be correctly reassured 32 out of 33 times, and would "miss" esophageal intubation only 3% of the time. However, when we recalculated EDD predictive value in a sample where "first attempt" intubation success was 71% rather than 97%, NPV fell to 77%. This means that, in a more realistic patient sample, a paramedic acting on "negative" EDD study would be correctly reassured 3 out of 4 times and would "miss" a quarter of esophageal intubations (Tables 2 and 3). These data illustrate the profound effect that prevalence has on the predictive value of diagnostic tests and demonstrates that when the prevalence of a condition is high, negative tests are more likely to be false negative and misleading.

A limitation of our study is the wide 95% confidence interval of the original estimate of sensitivity (Table 1). However, even if we assumed a maximal predicted sensitivity for the EDD of 67% in our hypothetical cohort, the NPV would still have been a disappointing 88%.

Lessons learned

This trial taught us a great deal about the difficulties of conducting prehospital research. We were surprised to observe that, in several instances, paramedics deviated from our study protocol and used a capnometer without applying the EDD. Perhaps we failed to engender a sense of ownership and hence our paramedics had no vested interest in completing the study successfully. In addition, they were unaware that this behaviour compromised patient enrollment and could bias the study results. These problems could have been mitigated by involving the paramedics during the trial design and implementation phases.

Our second mistake was the failure to establish a surveillance log. A surveillance log would track demographic characteristics of eligible subjects, allowing comparison of patients enrolled vs. patients excluded. Such a log would have quickly showed us that we were studying a skewed patient sample rather than the consecutive sample intended. With this knowledge we would have detected selection bias at an earlier stage and intervened to improve the recruitment process.

Our third problem was a conflict of interest involving our principal investigator, who was also the EMS medical director responsible for quality assurance and monitoring paramedic clinical skills. Based on this, it is conceivable that paramedics might have been reluctant to report failed intubations within the study.

These concerns led to several changes in our prehospital research program. First, we designated an emergency physician, who has no role in evaluating paramedic competence, as the director of prehospital care research. In addition, we no longer use research data for quality assurance purposes. Second, we formed a prehospital research committee with broad representation from the base hospital, paramedics and ambulance operational managers. All research protocols must now be presented to and approved by this group. All aspects of trial design and implementation are discussed in an open forum with the intent to troubleshoot problems in the early development stage. Third, we directly involve paramedics in research. Now, all projects must have a paramedic representative who is willing to liaise with their peers during study development and assist with trial implementation. Finally, we have added a research section to our paramedic newsletter and instituted regular sessions where prehospital research initiatives are presented. These keep paramedics informed and involved and allow us to address their questions and concerns directly.

Conclusion

The high "first attempt" intubation success rate seen in our study was due to selective exclusion of failed intubations. This selection bias led to a clinically important overestimation of the EDD's negative predictive value. Bias may substantially alter the estimations of test accuracy reported in scientific studies. To reduce the chance of unrecognized selection bias, investigators must determine whether recruited subjects resemble patients in whom the test will ultimately be used.

Acknowledgments: The authors would like to thank the late Dr. Christopher Rubes for his leadership in establishing the Toronto paramedic program and for originally proposing the EDD trial. We also acknowledge Toronto paramedics, who have taught us how to improve our EMS research program. This trial was partially funded by an Ontario Ministry of Health, Emergency Health Services grant (#09894S).

References

1. White SJ, Slovis CM. Inadvertent esophageal intubation in the field: reliance on a fool's "gold standard." Acad Emerg Med 1997;4:89-91.
2. Falk JL, Sayre MR. Confirmation of airway placement. Prehosp Emerg Care 1999;3:273-8.
3. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research: getting better but still not good. JAMA 1995;274:645-51.
4. Pearl WS. A hierarchical outcomes approach to test assessment. Ann Emerg Med 1999;33:77-84.
5. Bozeman WP, Hexter D, Liang HK, Kelen GD. Esophageal detector device versus detection of end-tidal carbon dioxide level in emergency intubation. Ann Emerg Med 1996;27:595-9.

6. Morgan D, Trompler V. Concerns about intubation placement aids [letter]. Acad Emerg Med 1997;4:928-9.
7. Jenkins WA, Verdile VP, Paris PM. The syringe aspiration technique to verify endotracheal tube position. Am J Emerg Med 1994;12:413-6.
8. Cardoso MMSC, Banner MJ, Melker RJ, Bjoraker DG. Portable devices used to detect endotracheal intubation during emergency situation: a review. Crit Care Med 1998;26:957-64.
9. Marley CD, Eitel DR, Koch MF, Hess DR, Taigman MA. Prehospital use of a prototype esophageal detection device: A word of caution! Prehosp Disaster Med 1996;11:223-7.
10. Pelucio M, Halligan L, Dhindsa H. Out-of-hospital experience with the syringe esophageal detector device. Acad Emerg Med 1997;4:563-8.
11. Ardagh M, Moodie K. The esophageal detector device can give false positives for tracheal intubation. J Emerg Med 1998;16:747-9.
12. Davis DP, Stephen KAC, Vilke GM. Inaccuracy in endotracheal tube verification using a toomey syringe. J Emerg Med 1999;17:35-8.
13. Sheps SB, Schechter MT. The assessment of diagnostic tests: a survey of current medical research. JAMA 1984;252:2418-22.
14. Mower WR. Evaluating bias and variability in diagnostic test reports. Ann Emerg Med 1999;33:85-91.
15. Anton WR, Gordon RW, Jordan TM, Posner KL, Cheney FW. A disposable end-tidal CO_2 detector to verify endotracheal intubation. Ann Emerg Med 1991;20:271-5.
16. MacLeod BA, Heller MB, Gerard J, Yealy DM, Menegazzi JJ. Verification of endotracheal tube placement with colorimetric end-tidal CO_2 detection. Ann Emerg Med 1991;20:267-70.
17. OLeary JJ, Pollare BJ, Ryan MJ. A method of detecting oesophageal intubation of confirming tracheal intubation. Anaesth Intens Care 1988;16:299-301.
18. Wee MYK. The oesophageal detector device: assessment of a new method to distinguish oesophageal from tracheal intubation. Anaesthesia 1988;43:27-9.
19. Altman DG, Bland JM. Diagnostic tests 1: sensitivity and specificity. BMJ 1994;308:1552.
20. Deyo RA, Haselkorn J, Hoffman R, Kent DL. Designing studies of diagnostic tests for low back pain or radiculopathy. Spine 1994;19:2057S-65S.
21. Doubilet P, Herman PG. Interpretation of radiographs: effect of clinical history. AJR 1981;137:1055-8.
22. Sox HC. The evaluation of diagnostic tests: principles, problems and new developments. Annu Rev Med 1996;47:463-71.
23. Altman DG, Bland JM. Diagnostic tests 2: predictive values. BMJ 1994;309:102.

Correspondence to: Dr. Richard Verbeek, BG-15, Emergency Services, 2075 Bayview Ave., Toronto ON M4N 3M5; 416 392-3885; fax 416 397-9060; e.rverbeek@basehospital.on.ca

https://doi.org/10.1017/S1481803500004875 Published online by Cambridge University Press