Mystery shopping in health service evaluation

Helen Moriarty, Deborah McLeod and Anthony Dowell

SUMMARY

Background: *Over the last 5 years, primary care telephone triage systems have been introduced in the United Kingdom, United States, Australia, and most recently in New Zealand. Evaluation of the clinical safety of such systems poses a challenge for health planners and researchers.*

Aim: *To evaluate the use of simulated patients in the assessment of aspects of clinical safety in a pilot New Zealand primary care telephone triage service.*

Design of study: *'Mystery shopping', an evaluation strategy commonly used in market research, was adapted by using simulated patients for telephone triage service evaluation.*

Setting: *New Zealand.*

Methods: *Four scripted clinical scenarios were developed by academic general practitioners, validated in student teaching situations, and then used by simulated patients to make 101 telephone calls. The scenarios were designed to necessitate a referral to a medical practitioner for further investigation. The documentation kept by the callers was compared with the call records from the telephone triage company, and both were analysed for capture and handling of the clinical safety features of each scenario. In cases where the endpoint was not a medical assessment, possible reasons for this were explored.*

Results: *Records were retrieved for 85 telephone calls. Considerable triage variability was discovered. There were discrepancies between expected and actual triage outcomes with 51% of analysed calls resulting in a self-care recommendation. A number of reasons were identified both for the triage variability and the unpredicted outcomes. Audiotaping of consultations would have enhanced the credibility of the evaluation but it would have carried ethical constraints.*

Conclusion: *Simulated patients can be used to evaluate the limitations of health services and to identify areas that could be addressed to improve patient safety. Evaluation of patient satisfaction with services is not sufficient alone to evaluate safety.*

Keywords: *clinical competence; consultation and referral; delivery of health care; health services evaluation; outcome assessment; patient simulation; remote consultation; telephone hotline; triage.*

H Moriarty, DPH, MGP, FAChAM, FRNZCGP, senior lecturer; D McLeod, DPH, PhD, research director; A Dowell, FRCGP, professor, Department of General Practice, Wellington School of Medicine and Health Science, University of Otago, New Zealand.

Address for correspondence

Dr H J Moriarty, Department of General Practice, Wellington School of Medicine and Health Science, University of Otago, PO Box 7343 Wellington South, New Zealand.

Submitted: 17 February 2003; Editor's response: 30 May 2003; final acceptance: 6 August 2003.

Introduction

TELEPHONE triage is a traditional role for primary care nurses, but the use of dedicated call centres using decision-support systems is a more recent development. These services have been introduced progressively in the United States (US), the United Kingdom (UK), Australia[1] and New Zealand over the past 5–6 years. The pilot New Zealand telephone triage service used McKesson's CareEnhance™ Systems decision-support program, designed in the USA. Binary chain logic algorithms (successive clinical questions with yes or no answers) guide the triage nurse consultations. The system allows nurse override of the endpoint as a safety feature; the nurse may exercise clinical judgement to correct clinical inconsistencies and counterbalance any vagaries of the triage algorithm if necessary.

The New Zealand pilot triage service was introduced to provide a safe, effective, and flexible 24-hour service to triage health enquiries, inform safe patient self-care, and facilitate referral where appropriate. Specific objectives were to help 'improve quality, increase cost-effectiveness, and reduce unnecessary demands on other health services' and to address inequalities in the use of health services by New Zealand Maori, who have higher risks of some chronic health conditions.[2] The pilot service was implemented in four regions, selected by socioeconomic and ethnicity profile, and covered a population of nearly 630 000.

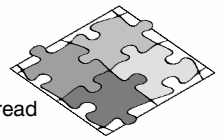
When the UK National Audit Office recently audited a similar telephone triage service in the UK's NHS Direct service,[3] the need to explore, not only call statistics, customer satisfaction, and changes in caller intent, but also the clinical appropriateness of the triage service was noted.[4] A number of possible approaches were mooted,[3] including the use of standardised scenarios. A small study found considerable clinician variability of advice given to mystery callers to NHS Direct, raising questions about triage consistency.[5]

The New Zealand pilot triage service evaluation included audits of client satisfaction and referral patterns, but these outcomes did not take into account caller medical conditions, health outcomes, or the utility of the algorithms, thereby limiting their value in determining clinical safety.

Mystery shopper methodology is a well-established and evaluated market research tool, with international health service evaluation applications.[6,7] Simulated patients have been used extensively in medical teaching and assessment in New Zealand[8] and overseas,[9] and in health services outcomes research both in New Zealand[10,11] and overseas. This has been shown to be a valid and reliable method of assessment of the quality of clinical care.[12,13]

A research protocol was developed and evaluated jointly by the Department of General Practice, Wellington School of Medicine and Health Science, University of Otago and BRC

HOW THIS FITS IN



What do we know?

Telephone triage is becoming a widespread and acceptable practice in primary care. Assessment of clinical safety is an unresolved dilemma especially in the context of known variation in clinical practice.

What does this paper add?

This paper describes the use of simulated patients to explore aspects of clinical safety. Simulated patients can be used to identify safety concerns that could be targeted in quality improvement exercises.

Marketing & Social Research (a market research company), using simulated patients to assess the clinical safety of the telephone triage pilot.

Methods

The pilot telephone triage service was notified of the intention to run simulated patient telephone calls from each of the four pilot regions in October and November 2001. Ethics committee approval was obtained from the Wellington Ethics Committee accredited by the Health Research Council of New Zealand.

Senior academic general practitioners (GPs) from the Wellington School of Medicine and Health Science developed four standardised scenarios. These were based on case studies that were validated through their use in teaching and evaluation of medical consultation skills. Scenario developers were unaware of the nature or order of algorithm questions. Each case incorporated clinical elements with safety implications strongly suggestive of need for a practitioner referral. The cases were as follows:

- a male patient with a past history of rheumatic fever calling about a sore throat,
- a female patient with venous thromboembolism risk factors calling about chest pain,
- a young woman with sexually transmitted disease risk factors calling about genitourinary symptoms,
- and a female care-giver calling about a systemically unwell child.

In the last case, the child had a fever, rash, vomiting, and loose stools, and simple self-care measures had failed. Three paediatric scenario versions were formulated, with fever, rash, or vomiting and loose stools given as the presenting symptom. The simulated patient callers were instructed to reveal certain details only if specifically asked. These details were as follows: sexual risk-taking in the adult dysuria scenario, past history of rheumatic fever in the sore throat scenario, recent travel and oral contraceptive use in the chest pain scenario, and behaviour change and appetite loss in the child scenario.

Scripted variations in caller name, location, and other nonessential details were incorporated so that each caller would appear to be unique, despite multiple calls for each scenario. Detailed briefing notes and scenario notes were developed to support the callers in their fieldwork. The writers discussed scenarios in depth with the market research company to highlight the important elements of each case. A checklist was designed for completion by standardised simulated patients during each call.

The callers (four women and two men) were experienced market research interviewers, subcontracted by BRC Marketing & Social Research. Callers were briefed verbally and in writing by the market research company to emphasise adherence to the script. After the first week, a teleconference was held between the simulated patients and one author to ensure that the instructions were being followed correctly, that scenarios were working as intended, and to discuss queries. At this stage, the chest pain scenario was modified to alter the nature and site of the chest pain. This case had triggered advice to seek assistance urgently and offers to call out an ambulance, which placed some callers in the difficult position of evading urgent ambulance despatch while remaining in character over the telephone. Other scenarios worked satisfactorily without revisions.

The simulated patients documented their interactions on forms specially designed to capture the nature of the nurse's greeting, the overall impression of the consultation, and some case-specific details of triage questioning.

Computerised call records, obtained from the telephone triage service at completion, were matched with the caller documentation and examined for capture of the patient history, clinical safety elements, and appropriateness of advice. If the endpoint of the consultation was not a medical assessment, possible reasons for this were sought by in-depth analysis of the records. The findings were presented to the telephone triage service providers and reasons why a referral to a medical practitioner had not occurred were discussed.

Results

A total of 101 simulated patient calls were made (Table 1). Computerised call records were recovered for 85% of the calls (85/101). No specific reasons could be identified for the inability to recover the remaining 15%. The recovered calls had been handled by 24 different triage nurses throughout the 2-month period, consequently most individual nurses had responded to only one call for each scenario (with a maximum of two same-scenario triages per nurse).

In 51% (43/85) of analysed calls, the triage ended with a self-care recommendation. The callers were very complimentary about the nature of their telephone interaction with the triage nurses and were generally satisfied with the advice offered. In only two calls did the triage nurse override the algorithm endpoint to upgrade the advice — in one case the nurse cited 'cannot exclude need for medical assessment, mother's concern about child's unusual behaviour' and the other was overridden because it was a call on behalf of a third party.

Some contributory factors in triage variability were readily identified (Table 2); for example, in some sore throat calls, where the past history of rheumatic fever had been documented but not as a response to any algorithm question, the triage failed to recommend medical assessment. At times the responses to algorithm questions contradicted informa-

Table 1. Analysis of mystery caller records.

Scenario	Calls per scenario (*n*)	Calls analysed (*n* [%])	Provider referral recommended (*n* [%])
Adult sore throat	24	19 (79)	8 (42)
Chest pain			
All cases	17	16 (94)	12 (63)
Initial version	*6*	*6 (100)*	*6 (100)*
Amended version	*11*	*10 (91)*	*6 (60)*
Sick child			
All cases	38	30 (79)	3 (10)
Fever presentation	*13*	*10 (78)*	*3 (23)*
Rash presentation	*12*	*11 (83)*	*0*
Vomiting and loose stools presentation	*13*	*9 (69)*	*0*
Adult dysuria	22	20 (91)	20 (100)
Total	101	85 (85)	43 (51)

tion captured in free comment fields. Examples included the freehand documentation of a past history of rheumatic fever but a negative response was entered for the rheumatic fever algorithm question, and for the child scenario 'spots on body' were noted but a negative response was given for the algorithm question on the presence of a rash. Some triage questions were subject to variable interpretation by the nurses with, for example, differing opinions on presence of 'diaphoresis' or a 'high fever'. In paediatric calls, triage provider referral resulted only when fever was given as the first symptom.

Discussion

Main findings

This study was designed to explore the use of simulated patient methodology to identify potential safety issues in a telephone triage service. Our evaluation has demonstrated apparent failures of triage when the system is artificially put under pressure. The intent, when designing cases to be delivered by simulated patients, was not to portray a specific diagnosis but to deliver a cluster of presenting clinical symptoms warranting further investigation, with the expectation that triage referral to medical assessment would be recommended. However, in half of the consultations carried out in this study there was no recommendation for medical assessment.

Factors with the potential to influence this outcome included the unpredictability of human interactions, the limitations imposed on the consultations by the algorithm structure, and human error, as well as the chosen research methodology itself. Our analysis revealed some triage problems that were directly attributable to clinician error, system failure or both factors in combination. Clinical information discussed by the patient and documented by the nurse was not reflected in the programmed triage recommendation unless captured correctly by the algorithm questions. In the cases of chest pain, paediatric, and sore throat scenarios especially, information inaccurately entered or missed in the algorithm led to unintended self-care recommendations. Inconsistencies between patients' free comments and the information history captured in the responses to algorithm questions had passed unnoticed by the nurses, as prior responses scrolled off the screen. Thus, triage could be misdirected through data capture and documentation problems.

Although the program permitted nurses to override the algorithm endpoint to correct for clinical nuance, ambiguities or discrepancies, this occurred infrequently, even when the documented history raised questions about the appropriateness of the triage endpoint. Reasons for this were unclear, but personal accountability may be one reason for nurse reticence to override endpoints.

Study strengths and limitations

The use of simulated patient methodology was both a key strength of and a limitation to this study. The direction and outcome of each telephone triage encounter is potentially

Table 2. Analysis of mystery caller records.

Scenario	Factors influencing triage	Calls affected (*n* [%])
Adult sore throat (19 calls analysed)	No documentation that rheumatic fever history was sought or volunteered	4 (21)
	Negative response to rheumatic fever history question entered into algorithm	11 (58)
	Rheumatic fever past history documented in free comment field, but negative response entered to algorithm question	2 (10)
Chest pain (all cases combined) (16 calls analysed)[a]	Variable nurse interpretation of presence of 'true diaphoresis'	3 (19)
	Recent air travel history not sought/not volunteered	3 (19)
	Oral contraceptive use not sought/not volunteered	4 (25)
	Answers to the fixed algorithm questions contradict history recorded in free comment fields	4 (25)
Sick child (all versions combined) (30 calls analysed)[a]	Variable nurse interpretation of presence of temperature, 'high' or 'very high' temperature	15 (50)
	Variable nurse interpretation of 'attempted self care' or 'self-care failure'	7 (23)
	Other relevant history documented but not captured by answers to questions of the algorithm	17 (57)
Adult dysuria (20 calls analysed)	Caller confidentiality concerns were documented but the reasons for that were not explored	4 (20)
	Caller confidentiality concerns were not documented	16 (80)

[a]More than one factor per call in some instances.

influenced by a number of factors relating to the behaviour of the caller, the triage nurse, and the triage system itself, as well as the interaction of all three. Although it was possible to standardise the clinical scenario under study, and the triage algorithm questions were fixed, it was not possible to standardise the nature of the interaction between the caller and the triage nurse. In discussing the study findings, the service provider pointed out that the transmission of human emotions such as fear, anxiety, and concern would not have been easily captured in our evaluation. The nuances of the interpersonal interactions could not be explored within this evaluation since the telephone interactions were not tape recorded. Confidentiality considerations required that only mystery caller telephone calls, and not genuine patient calls, were recorded, placing responsibility for recording with the caller. An element of deception is intrinsic with use of simulated patients, but tape recording could have repercussions for both the nurse and simulated patient if their respective performances were called into question by recorded evidence. Therefore, as with similar research,[14,15] written documentation was considered preferable to tape recordings of calls. The documentation demonstrated that, although our callers did deliver specific clinical clues, as concerned real patients would do, this information did not influence the triage recommendation unless it was entered in response to a fixed algorithm question.

The service provider had also suggested that apparent triage failures could result from the failure of simulated callers to follow the scripted scenario word-for-word. Available documentation showed that the key features of the case scenarios that the simulated patients had been instructed to deliver, were in fact recorded somewhere, although not necessarily in the most appropriate place.

A limitation of the study was the absence of a suitable primary care service for comparison (algorithm- or non-algorithm-driven nurse telephone triage, internet or face-to-face GP consultations). The choice of a comparative service is problematic because assessment criteria perform differently in these settings (Chris Salisbury, personal communication, 2003). Absence of a comparative setting has limited the extent to which any conclusions can be made regarding the impact of clinical practice variation on the findings of the evaluation. Practitioner variability[16,17] is a well accepted but poorly understood phenomenon, that has previously been identified in other nurse triage services.[5,12,14]

Related literature

Nurse-led telephone triage systems are becoming widely used.[1,3,14] To date, most evaluation has focused on surveys of patient satisfaction and summaries of the endpoints of triage.[4,14] However, these do not measure the quality of the advice provided to patients and there is a need to also examine the clinical safety of these systems. Concerns about the quality of advice was raised by research using a similar methodology of simulated patient calls to NHS Direct.[15] Paediatric calls, in particular, have been shown to be susceptible to inadequate advice under telephone triage conditions.[14]

A study of GP practice management systems in New Zealand concluded that the nature and format of information systems programming could impose a rigid framework upon professional consultation style.[18] This was true from our evaluation, where the triage algorithm, although intended to support clinical decision making, had directed the practice nurse through its set order of fixed algorithm questions, an inverse of the usual control in clinical consultations.

Future implications

The practitioner variability seen in simulated consultations warrants further evaluation, as do systems factors that may influence the accuracy of capture of the patient's agenda. Callers are unaware that such problems could potentially influence their telephone triage and safety, raising questions of the validity of caller satisfaction as a quality outcome measure. This may present significant implications for the future evaluation of health services.

The New Zealand government has stated its intention to develop a Health Care Quality Improvement Strategy, with parallel objectives to the UK National Patient Safety Agency,[19] and safety will be a key component of this. The UK National Audit Office audited a similar telephone triage service, NHS Direct.[3] To quote from that report 'No referral system can ever be perfect. What counts is that errors are reduced to a minimum'. If primary care telephone triage is to become widespread, attention to both actual and potential clinical safety issues should become a priority.

This evaluation has raised more questions about algorithm-assisted telephone triage consultations than it has answered.

References

1. Roland M. Nurse-led telephone advice. A useful additional service? [Editorial]. *Med J Aust* 2001; **176(3):** 96.
2. BRC Marketing & Social Research. *The evaluation of the Healthline Service: Final evaluative report*. New Zealand: Ministry of Health, 2002. http://www.moh.govt.nz/moh.nsf/wpg_Index/News+and+Issues-Healthline (accessed 11 Nov 2003).
3. National Audit Office. *NHS Direct in England*. London: The Stationery Office, 2002.
4. Medical Care Research Unit, University of Sheffield. *Evaluation of NHS Direct first wave sites: Final report of the Phase 1 research.* Sheffield: Medical Care Research Unit, University of Sheffield, 2001.
5. Williams S. NHS Direct investigated. *Health Which?* **2000 Aug:** 12-16.
6. European Society for Social Opinion and Marketing Research http://www.esomar.org (accessed 27 Oct 2003).
7. Devon Hill Associates. Mystery shopping http://www.devonhillassociates.com/home/ (accessed 11 Nov 2003).
8. Pullon S, McBain L. All the world's a stage: simulated patients in consultation skills teaching. *N Z Fam Physician* 1998; **25(3):** 53-56.
9. Barrows HS. An overview of the uses of standardised patients for teaching and evaluating clinical skills. *Acad Med* 1993; **68:** 443-453.
10. Consumer's Institute of New Zealand (Inc). dot DOC or dot QUACK? *Consumer* **2001 Mar:** 18-21.
11. Norris P. Which sorts of pharmacies provide more patient counselling? *J Health Serv Res Policy* 2002; **7(Suppl 1):** S23-S28.
12. Grant C, Nicholas R, Moore L, Salisbury C. An observational study comparing quality of care in walk-in centres with general practice and NHS Direct using standardised patients. *BMJ* 2002; **324:** 1556-1561.
13. Rethans JJ, Saebu L. Do general practitioners act consistently in real practice when they meet the same patient twice? Examination of intradoctor variation using standardised (simulated) patients. *BMJ* 1997; **314:** 1170-1176.
14. Aitken M, Carey M, Kool B. Telephone advice about an infant given by after-hours clinics and emergency departments. *N Z Med J* 1995; 108: 315-317.
15. Ratcliffe N. NHS Direct-help or hindrance? *Health Which?* **2003**

Jul: 10-13.
16. Audit Commission. *A prescription for improvement: towards more rational prescribing in general practice.* London: HMSO, 1994.
17. Walker J, Mathers N. The Impact of a general practice group intervention on prescribing costs and patterns. *Br J Gen Pract* 2002; **52:** 181-186.
18. Cornford E. Circuits of power. *A study of the development of computer software and its use in general medical practice* [MA thesis]. Palmerston North, New Zealand: Massey University, Department of General Practice Research, 2000.
19. National Health Committee. *Safe systems supporting safe care. Final report on health care quality improvement in New Zealand.* Wellington, New Zealand: National Health Committee, 2002.

Acknowledgements

This evaluation was funded by the New Zealand Ministry of Health. The researchers would like to thank the simulated patients and to acknowledge the input into the evaluation by BRC, staff at the telephone triage service, and academics at the Department of General Practice, Wellington School of Medicine and Health Science.