Medication Administration Evaluation and Feedback Tool:
Intra-rater and Inter-rater Reliability Using Simulation

Karen Davies[1][2], Ian Coombes[1][2], Samantha Keogh[1][3], Karen Hay[4], Cameron Hurst[4], Karen Whitfield[1][2]

[1]Royal Brisbane and Women's Hospital, [2]University of Queensland, [3]Queensland University of Technology, [4]QIMR Berghofer Medical Research Institute

Abstract

Purpose

The purpose of this study was to assess the reliability of an evaluation tool to observe and provide feedback on medication administration performance for nurses, based on the recommended content by an expert panel review.

Methods

The study was an observational fully crossed design of nurses administering medications in a digitally recorded simulated clinical environment. There were: a) 8 recorded scenarios; b) 3 nurse educators viewed the recordings and evaluated the nurse administering medications using the designed tool; c) The same 3 nurse educators viewed the recordings again 7 days later.

There were 22 criteria in the tool with ratings of yes, no and not applicable. A total of 176 ratings for each observer. The intra-rater and inter-rater consistency of agreement for each rater and between raters was determined using Fleiss' kappa and average percentage agreement.

Results

Results showed the intra-rater agreement percentage for each observer and overall comparison ratings for both time points ranged between 81.25% and 84.28%. The inter-rater average percentage agreement for both time points and overall ranged between 79.35% and 84.47%. The overall Fleiss' Kappa statistic for intra-rater reliability was 0.72 and inter-rater reliability was 0.68. The evaluation criteria for kappa agreement of 0.60-0.74 rates the tool as good. Therefore, overall intra-rater and inter-rater reliability of the tool was good.

Conclusions

The study shows that the designed medication administration evaluation and feedback tool is reliable in a simulated clinical environment. Further studies are being conducted to test the reliability of the tool in the clinical setting.

Aim

- To use **simulation** as a method to assess the **reliability** of the
- "Medication Administration Evaluation and Feedback Tool" (MAEFT)
  - » designed in Phase 1. based on the recommended content by an expert panel review (K. M. Davies, Coombes, Keogh, & Whitfield, 2018)

Objectives

- To test the **inter-rater and intra-rater reliability, agreement and accuracy** of the developed MAEFT
  - by Nurse Educators observing a nurse administering medications in **scripted** digitally recorded **scenarios in a simulated clinical environment**

Method

Study Design/ Setting

Observational fully crossed design for testing of reliability and agreement of the MAEFT
- Using scripted digitally recorded scenarios of a nurse administering medications
- In a clinical skills laboratory simulated environment in November 2017
- The setting was a simulated clinical ward bed environment with an adjoining medication preparation room fitted with a controlled drug safe.

To determine intra-rater and inter-rater reliability of the MAEFT, there were:
- **8 patient scenarios** digitally recorded of a nurse administering medications
- **3 nurse educators** viewed the digitally recorded scenarios in December 2017 and using the MAEFT evaluated the nurse's medication administration practice
- Same **3 nurse educators independently viewed** the **8** digitally recorded scenarios **again 7 days later**

Data Collection / Statistical Analysis

- Categorical rating scale MAEFT "yes", "no" or "not applicable" coded on a nominal scale
- Each observer viewed 8 simulations with 22 questions equalling 176 ratings per rater
- The consistency of agreement was calculated for:
  - intra-rater (for each rater)
  - inter-rater (between raters)
  - Fleiss' Kappa coefficient multiple raters (Fleiss, Levin, & Paik, 2004)
    » Corrects for agreement between ratings due to chance
  - Fully crossed two-way random effects design treating variation due to rater and scenario as random
  - Research Electronic Data Capture (REDcap) version 8.5.0. (Paul A. Harris, 2009)

Fleiss' Kappa coefficient Agreement Evaluation Criteria

- Poor= κ<0.40; fair= κ 0.40-0.59; good= κ 0.60-0.74; excellent= κ >0.74 (Polit & Beck, 2006)
- Average percentage agreement Davies and Fleiss (Davies M, 1982)
- Subgroup analysis for each individual question calculated for agreement using Fleiss' Kappa coefficient
- Where there was perfect agreement - average percentage agreement was calculated
- Average percentage correct answer was also calculated to determine accuracy Statistical Analysis was conducted using:
  - Stata (v15, StataCorp, College Station, 2017) and
  - R statistical package (v3.4.4, R Core Team, 2018)
  - R package irr used to calculate Fleiss' Kappa coefficient (Gamer, Lemon, Fellows, & Singh, 2012)

Results

Table 1: **Intra-rater reliability** Fleiss' Kappa coefficient statistic

Comparison	Observed Agreement	Expected Agreement	Fleiss' Kappa coefficient	95% CI	Evaluation
Overall comparisons of ratings between time1 and 2	84.28%	43.13%	0.7236	(0.66,0.79)	Good
Comparison for observer 1	84.09%	42.69%	0.7224	(0.63,0.82)	Good
Comparison for observer 2	87.50%	41.75%	0.7854	(0.70,0.87)	Excellent
Comparison for observer 3	81.25%	45.25%	0.6575	(0.55,0.76)	Good

Table 2: **Inter-rater reliability and accuracy** (% correct) measured using the Fleiss' Kappa coefficient statistic

Comparison	Outcome	Av % Agreement	Fleiss' Kappa coefficient	95% CI	Evaluation	Av % Correct
Comparing raters at time1	1		0.74		Good	79.55%
	2		0.71		Good	83.52%
	3		0.71		Good	84.09%
	combined	84.47%	0.73	(0.65,0.80)	Good	82.39%
Comparing raters at time2	1		0.65		Good	81.82%
	2		0.64		Good	80.11%
	3		0.61		Good	76.14%
	combined	79.35%	0.64	(0.56,0.72)	Good	79.36%
Overall comparison of raters	1		0.70		Good	80.68%
	2		0.68		Good	81.82%
	3		0.66		Good	80.11%
	combined	81.91%	0.68	(0.63,0.74)	Good	80.87%

Key Results / Significance

- Initiated due to medication errors and demonstrated evidence that observation and direct formative feedback showed a statistically significant improvement to medication error rate (K. Davies, Mitchell, & Coombes, 2015; K. Davies, Norris, et al., 2015)
- There was also a lack of a suitable and validated tool to assess nurses' medication administration practice
- Fleiss' Kappa coefficient for intra-rater and inter-rater reliability of observers using the MAEFT measured as good
- Significance - the designed MAEFT demonstrates it is reliable when used by multiple observers to observe different nurse/ patient scenarios where there is a fixed simulated environment

Limitations

- No extensive training with practical demonstration
- No required minimum level of intra-rater reliability achievement prior to observation using the MAEFT
- This could have potentially lowered the achieved intra-rater reliability results
- Tested in a simulated environment only
- Results may not translate to a practical / clinical environment

Conclusions

- Intra-rater and inter-rater reliability and accuracy of the designed MAEFT shows that it is reliable and accurate in a simulated environment
- Further studies have been conducted in a clinical environment to determine generalisability of the MAEFT and results are undergoing analysis
- A pre / post pilot study to assess whether using the MAEFT makes a difference in medication administration practice is planned
- Further research developing an education plan for maximum benefit using the MAEFT is required

References

- Davies, K., Mitchell, C., & Coombes, I. (2015). The Role Of Observation And Feedback In Enhancing Performance With Medication Administration. Journal of law and medicine, 23(2), 316.
- Davies, K., Norris, L., O'Brien, E., Buksh, F., Rogers, K., Kubler, P., Coombes, I. (2015). Internal Medicine Services (IMS) Medication Administration Observation & Feedback. Paper presented at the Royal Brisbane and Women's Hospital Symposium, Brisbane.
- Davies, K. M., Coombes, I. D., Keogh, S., & Whitfield, K. M. (2018). Medication administration evaluation tool design: An expert panel review. Collegian. doi:10.1016/j.colegn.2018.05.001
- Davies M, a. F. J. (1982). Measuring agreement for multinomial data. Biometrics., 38(4), 1047–1051.
- Fleiss, J. L., Levin, B., & Paik, M. C. (2004). Statistical Methods for Rates and Proportions (3rd ed.. ed.). Hoboken: Hoboken : Wiley.
- Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2012). irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84. Retrieved from: https://CRAN.R-project.org/package=irr
- Paul A. Harris, R. T., Robert Thielke, Jonathon Payne, Nathaniel Gonzalez, Jose G. Conde. (2009). Research electronic data capture (REDCap) - A metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform, Apr;42(2), 377-381.
- Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? critique and recommendations. Research in Nursing & Health, 29(5), 489-497. doi:10.1002/nur.20147