Intra-Rater and Inter-Rater Reliability of Manually Quantifying Neurological Outcome of Stroke Patients to be Used for Machine Learning Models

Catherine Chong[1]; Jack Pohlmann, MS[2]; Samad Amini, MSc[3]; Ivy So Yeon Kim, MPH[2]; Charlene Ong, MD, MPHS[2, 4]

Richard Montgomery High School, 250 Richard Montgomery Drive, Rockville, MD 20852[1]; Department of Neurology, Boston Medical Center, 1 Boston Medical Center PI, Boston, MA 02118[2]; Center for Computing & Data Sciences, 665 Commonwealth Ave, Boston, MA 02215[3]; Boston University School of Medicine, 72 E. Concord St, Boston, MA 02218[4]

INTRODUCTION

The **modified Rankin Scale (mRS)** is a six point global outcome rating scale used to evaluate **stroke patient outcomes** by assessing **levels of functional independence.**

The purpose of this report is to:

- **manually quantify functional outcomes** from patient discharge summaries
- analyze **intra- and inter-rater reliability** of manual mRS assessments
- establish a **gold standard label for natural language processing models**

The modified Rankin Scale (mRS)	
Score	Symptoms
0	None
1	No significant disability despite symptoms: able to carry out all usual duties and activities
2	Slight disability: unable to carry out all previous activities, but able to look after own affairs without assistance
3	Moderate disability: requiring some help, but able to walk without assistance
4	Moderately severe disability: unable to walk without assistance, unable to attend to needs without assistance
5	Severe disability: bed-ridden, incontinent, and requiring constant nursing care and attention
6	Dead

Table 1: modified Rankin Scale scoring guidelines adapted from Zihni, et al[1]

METHODS

- 305 BMC stroke patient discharge summaries were graded using the mRS by a trained investigator (Investigator A) to determine if mRS can be reliably extracted from these reports
- 30 of the 305 discharge summaries were randomly selected for a second assessment. Percentage agreement and kappa values were calculated

- Intra-rater reliability
  - **Investigator A's two mRS assessments** blindly compared; reliability determined using Cohen's kappa
- Inter-rater reliability
  - **Investigator A's** first assessment compared to a second investigator's assessment (**Investigator B**); reliability determined using Cohen's kappa
  - **Investigator A's** first assessment compared to mRS originally included in **discharge summary**; reliability determined using Fleiss' kappa
  - Comparing **Investigator A's** first assessment, **Investigator B**, and mRS originally included in **discharge summary**; reliability determined using Fleiss' kappa
  - Reliability when considering **favorable** (mRS 0-2) **vs. unfavorable** (mRS 3-6) **outcomes**

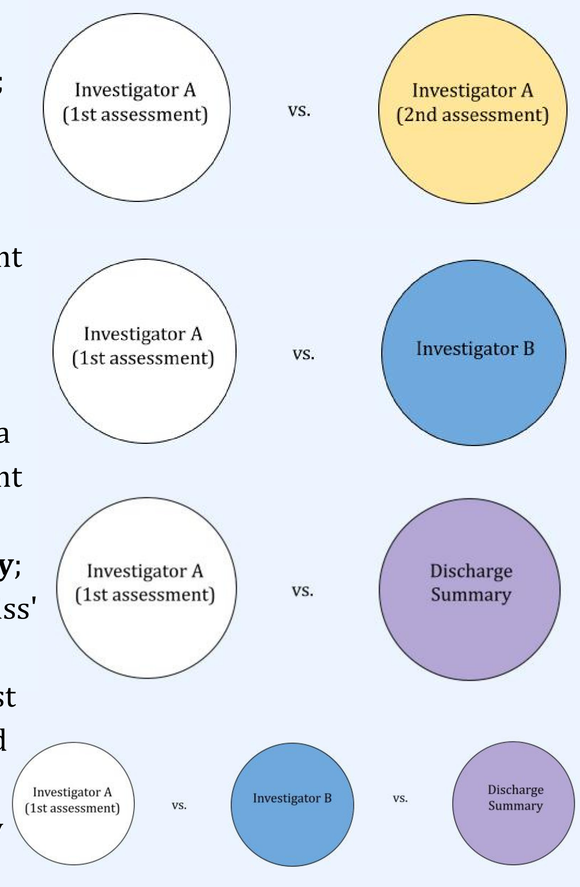

Figure 1: Comparisons for Intra- and Inter- Reliability

RESULTS

Of the reviewed discharge summaries, 278 **(91.2%) were easily assessable on first pass by Investigator A.**

Table 2: Intra- and Inter-Rater Agreement and Kappa Statistic

Raters	N	Full mRS Agreement		Dichotomous mRS* Agreement	
		%	k	%	k
Intra-Rater (Investigator A)	30	76.7	0.89	93.3	0.86
Investigators A v. B	30	55.6	0.77	85.2	0.71
Investigators A v. DS	27	83.3	0.92	100	1.00
Group Agreement	27	44.4	0.61	85.2	0.80

*Dichotomous mRS split into functional outcomes of favorable (mRS ≤2) or unfavorable (mRS >2)

mRS: modified Rankin Scale; N: number of reports reviewed; %: percentage agreement; k: kappa statistic

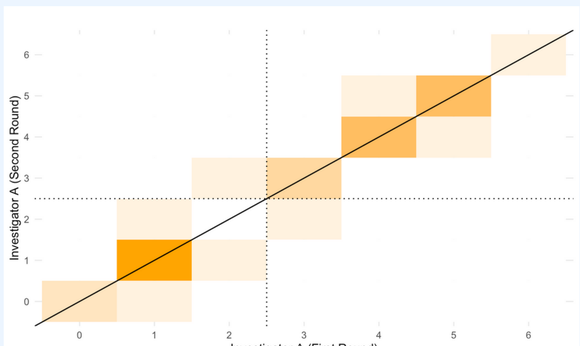

Figure 2: Intra-Rater mRS Assessment Agreement

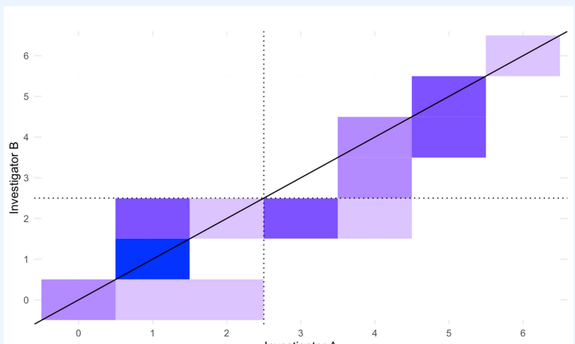

Figure 3: Inter-Rater mRS Assessment Agreement between Investigator A (1st Assessment) and Investigator B

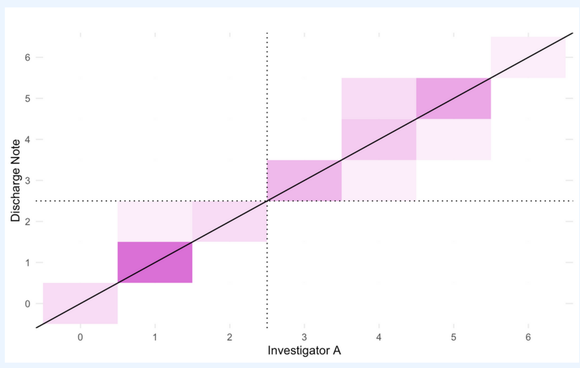

Figure 4: Inter-Rater mRS Assessment Agreement between Investigator A (1st Assessment) and Discharge Summary

**Darker boxes indicate greater instances of agreement between raters

Most reviewed reports follow the trend line of perfect agreement, indicating good intra- and inter-rater agreement

DISCUSSION/CONCLUSION

This study demonstrated that information extracted from **stroke patient discharge summaries** are **sufficient** to **reliably extract an mRS**. A **relatively low percentage agreement between Investigators A and B** may be due to the fact that Investigator A reviewed the patients' full discharge summaries while Investigator B only reviewed their discharge exams. This discrepancy is important to note when considering what information to train the machine learning model on to generate a mRS. However, the kappa value (0.71) demonstrates that inter-rater reliability is still substantial. Additionally, allowing a tolerance of 1 point shows great percentage agreement between the two investigators (85.2%). Overall, the results, most notably kappa values signifying **moderate to near perfect agreement,** indicate high intra-rater and inter-rater reliability, giving confidence that these mRS ratings can be **employed as a gold standard label** to train a natural language machine model. Ultimately, this would expedite evaluating functional outcomes at discharge, helping to better assess post-stroke condition and to determine the course of treatment intervention.

References

[1] Linwood, Simon L, et al. "Chapter 6 Moving Toward Explainable Decisions of Artificial Intelligence Models for the Prediction of Functional Outcomes of Ischemic Stroke Patients." Digital Health, Exon Publications, Australia, 2022, https://www.ncbi.nlm.nih.gov/books/NBK580624/. Accessed 7 Aug. 2023.

2 Chang, Winston. "Inter-Rater Reliability." R Graphics Cookbook, O'Reilly, Beijing Etc., 2013, http://www.cookbook-r.com/Statistical_analysis/Inter-rater_reliability/. Accessed 7 Aug. 2023.

Acknowledgments

Special thanks to Jack Pohlmann and Dr. Charlene Ong for all their guidance and mentorship throughout the research process. I would also like to thank the rest of the Ong lab for their continued help and support. Finally, I would like to thank the RISE program for this amazing opportunity.