BMJ Health & Care Informatics

# Development of a scoring system to quantify errors from semantic characteristics in incident reports

Haruhiro Uematsu ,[1] Masakazu Uemura,[2] Masaru Kurihara ,[2] Hiroo Yamamoto,[2] Tomomi Umemura,[2] Fumimasa Kitano,[2] Mariko Hiramatsu,[2] Yoshimasa Nagao[1,2]

**To cite:** Uematsu H, Uemura M, Kurihara M, *et al*. Development of a scoring system to quantify errors from semantic characteristics in incident reports. *BMJ Health Care Inform* 2024;**31**:e100935. doi:10.1136/bmjhci-2023-100935

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjhci-2023-100935).

Received 11 October 2023
Accepted 02 April 2024

[1]Department of Quality and Patient Safety, Nagoya University Graduate School of Medicine, Nagoya, Aichi, Japan
[2]Department of Patient Safety, Nagoya University Hospital, Nagoya, Aichi, Japan

**Correspondence to**
Dr Haruhiro Uematsu; hiro_uematsu@hotmail.com

## ABSTRACT

**Objectives** Incident reporting systems are widely used to identify risks and enable organisational learning. Free-text descriptions contain important information about factors associated with incidents. This study aimed to develop error scores by extracting information about the presence of error factors in incidents using an original decision-making model that partly relies on natural language processing techniques.

**Methods** We retrospectively analysed free-text data from reports of incidents between January 2012 and December 2022 from Nagoya University Hospital, Japan. The sample data were randomly allocated to equal-sized training and validation datasets. We conducted morphological analysis on free text to segment terms from sentences in the training dataset. We calculated error scores for terms, individual reports and reports from staff groups according to report volume size and compared these with conventional classifications by patient safety experts. We also calculated accuracy, recall, precision and F-score values from the proposed 'report error score'.

**Results** Overall, 114 013 reports were included. We calculated 36 131 'term error scores' from the 57 006 reports in the training dataset. There was a significant difference in error scores between reports of incidents categorised by experts as arising from errors ($p<0.001$, $d$=0.73 (large)) and other incidents. The accuracy, recall, precision and F-score values were 0.8, 0.82, 0.85 and 0.84, respectively. Group error scores were positively associated with expert ratings (correlation coefficient, 0.66; 95% CI 0.54 to 0.75, $p<0.001$) for all departments.

**Conclusion** Our error scoring system could provide insights to improve patient safety using aggregated incident report data.

### WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Incident reporting systems have an important role in patient safety. Incidents caused by errors have a particularly significant influence on patient safety. There are various methods to analyse errors in healthcare settings, but to our knowledge, no studies have explored methods to quantify errors or analyse organisational trends by using the scores developed from free text in incident reports.

### WHAT THIS STUDY ADDS

⇒ We developed error scores that partly rely on natural language processing techniques to obtain quantitative information about the presence of error factors in incident reports. Group error scores, representing averaged error scores in a certain group, were positively associated with manual ratings of patient safety experts. Our error scoring system could provide insights to improve patient safety using aggregated incident report data.

### HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The proposed model will be used to monitor chronological trends in errors in groups and increase the awareness of workers and general risk managers. This system will also potentially be helpful for preventing future incidents by providing a warning (score changes), as well as for educational purposes. In future, a useful tool to improve patient safety may be developed by combining and balancing multiple factors to produce scores using the same methodology applied herein.

## INTRODUCTION

Many healthcare organisations have endorsed patient safety measures over the years.[1] However, the rates of medical errors and adverse events continue to be of serious concern.[2] Measures of quality are relatively well established, but the measurement and monitoring of safety continue to be problematic.[3]

Incident reporting systems allow healthcare workers to voluntarily disclose adverse events and 'near misses'.[4 5] These systems function as barometers of risk in the healthcare setting and provide a foundation for organisational learning and improvement.[6] In addition, voluntary confidential submission is thought to deepen our understanding of events and promote a safe environment.[5 7] The use of incident reports is strongly recommended by the Institute of Medicine.[8] However, their varying quality is considered suboptimal for organisational learning.[9] Reports submitted by frontline workers can provide valuable insights,[4] especially the free-text sections used to describe incidents in greater detail,[10] but interpretation of incidents is challenging for various reasons, including inadequate use of evolving health information technology.[11]

The integration of artificial intelligence into patient safety measures has gained greater attention in recent years.[12 13] Studies have explored how to obtain better value from incident reports using health information technologies. One recent study proposed an original decision-making model that partly relies on natural language processing (NLP) techniques to quantify the severity of incidents from aggregated big data and measure organisational trends using the central limit theorem.[14] This model was novel in two ways. First, it used an original vectorisation approach to weigh terms from a bag of words. This enabled conversion of narrative free-text data into quantitative measures. Second, the model aimed to investigate organisational patterns and trends using a computer-assisted decision-making model. Generally, techniques using NLP help to answer binary questions or classify incident types for individual reporting.[15] However, the WHO recommends collecting systemic insights from aggregated incident data,[6] and this model was helpful in investigating particular factors in incident reports at the organisational level. However, it is not clear whether it could also be used to measure other factors in incident reports.

Incidents caused by errors potentially have a significant influence on patient safety. The occurrence of errors could lead to malpractice suits, which have an impact on healthcare costs.[16] Errors also create a serious public health problem[17] and are associated with stress for healthcare professionals.[18] We therefore attempted to extract information about errors from incident reports using the model from a previous study. To date, various methods have been applied in healthcare settings to analyse errors.[19 20] However, to our knowledge, no studies have used models to quantify errors or analyse organisational trends in incident reports.

This study aimed to develop error scores to quantify errors in incidents using semantic characteristics in incident reports, and to confirm the criterion-related validity of these error scores by comparing them with manual ratings of patient safety experts.

## METHODS

### Data sources

#### Incident reporting systems

All incident reports were collected at Nagoya University Hospital (NUH), Japan. NUH is a 1080-bed hospital that contributes to advanced medical care, education and research. NUH is the only national university hospital in Japan accredited by the Joint Commission International, an accreditation body for healthcare quality and safety. NUH has used an incident reporting system since 2000 and a reporting culture is well established.[21] Every employee can report incidents anonymously through the electronic health record system. The system collects background data about incidents using a structured format and free-text descriptions. Collected reports are reviewed by trained general risk managers (GRMs), a multidisciplinary group including physicians, nurses, a pharmacist and lawyers. Our hospital has been making considerable efforts to eliminate severe error-containing events, and GRMs sort incident reports according to information such as the severity and nature of errors. Severity is classified into five categories using the grading system developed at NUH: 'Near Miss', 'No Harm', 'Low Harm', 'Severe Harm' and 'Catastrophic/Fatal Event'. It has similarities with other grading systems such as those of the WHO and National University Hospital Council of Japan.[14]

#### Manual data labelling (definition of error)

The term 'error' has various meanings depending on the context; its precise meaning is actively debated,[22] and no universal grading scale is used in healthcare. The WHO defines an error as a failure to carry out a planned action as intended or the application of an incorrect plan.[23] Errors are divided into active and latent. Active errors are caused by unsafe acts committed by personnel resulting from slips, lapses, mistakes or violations.[24] Latent errors may provoke further errors or create inherent weaknesses in the system. In NUH, GRMs review all incident reports, and incidents are labelled as containing errors if they are associated with any types of system, process or human errors, regardless of whether a patient was harmed. Reports are classified as error free when the incident is considered very unlikely to have occurred as the result of an error.

#### Generating training and validation datasets

We retrospectively extracted free-text data from incident reports dating from January 2012 to December 2022 at NUH. We included all free-text data from the submitted incident reports, regardless of the type of incident or the length of the text. Data were randomly allocated to equal-sized training and validation datasets. The training dataset was used to generate error scores according to the GRM classification, and the validation dataset was used to test the scores.

### Semantic feature representation

We segmented the original free-text data to the smallest unit using morphological analysis to define the semantic characteristics in incident reports. Of the possible parts of speech, we used only nouns in the analysis because they appear most frequently in the text and represent the smallest unit of meaning. We did not preprocess the data because we prioritised applying the method to real-world data. During segmentation, sentences were processed into an unordered collection of words, known as a bag of words. This analysis was performed by an

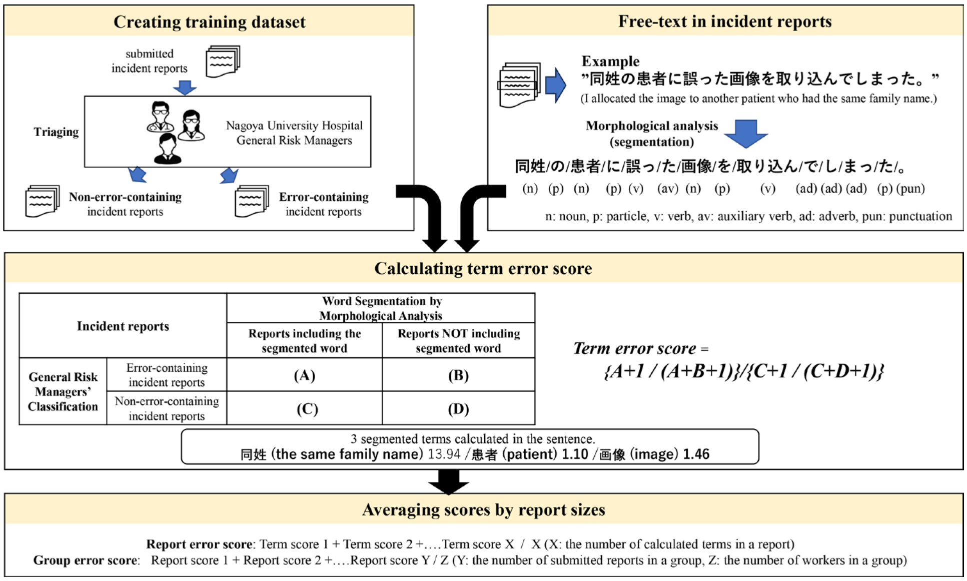

Creating training dataset
submitted incident reports
Triaging
Nagoya University Hospital General Risk Managers
Non-error-containing incident reports
Error-containing incident reports
Free-text in incident reports
Example
”同姓の患者に誤った画像を取り込んでしまった。”
(I allocated the image to another patient who had the same family name.)
Morphological analysis (segmentation)
同姓/の/患者/に/誤っ/た/画像/を/取り込ん/で/し/まっ/た/。
(n) (p) (n) (p) (v) (av) (n) (p) (v) (ad) (ad) (ad) (p) (pun)
n: noun, p: particle, v: verb, av: auxiliary verb, ad: adverb, pun: punctuation
Calculating term error score
Incident reports
Word Segmentation by Morphological Analysis
Reports including the segmented word
Reports NOT including segmented word
General Risk Managers' Classification
Error-containing incident reports
Non-error-containing incident reports
(A)
(B)
(C)
(D)
Term error score = {A+1 / (A+B+1)}/{C+1 / (C+D+1)}
3 segmented terms calculated in the sentence.
同姓 (the same family name) 13.94 /患者 (patient) 1.10 /画像 (image) 1.46
Averaging scores by report sizes
Report error score: Term score 1 + Term score 2 +….Term score X / X (X: the number of calculated terms in a report)
Group error score: Report score 1 + Report score 2 +….Report score Y / Z (Y: the number of submitted reports in a group, Z: the number of workers in a group)

**Figure 1** Process for calculating error scores from original text in incident reports.

open-source engine, MeCab, which was equipped with two commercially available medical dictionaries, MANBYO (MANBYO_201907_Dic-sjis) and Comejisyo (ComeJisyo Sjis-2), for application to medical writing.

## Calculating error scores

Figure 1 provides an overview of how the error scores were developed from incident reports. After segmentation, the bag of words was transformed into a numerical representation using the original vectorisation (*{A+1/(A+B+1)}/{C+1/(C+D+1)}*), inspired by the epidemiological concept of relative risk. All segmented terms were examined using the $\chi^2$ test in terms of the relative frequency of their use in error-containing reports and other reports, as classified by GRMs. We modified the formula by adding one to both the numerator and denominator to pick up more terms from free-text data and avoid zero probability. When a term appeared more frequently in reports of error-free incidents (C+1/(C+D+1)>A+1/(A+B+1)), implying that it is less important in incidents arising from errors, we reversed the numerator and denominator and replaced the plus sign with a minus sign (−[C+1/(C+D+1)]/[A+1/(A+B+1)]). A term score having more impact on the likelihood of errors being associated with the incident therefore becomes greater than 1, and one with the opposite effect becomes less than −1.

Once the error scores for segmented terms had been calculated, we averaged the scores for each report unit. Then, the scores for a certain group (clinical or non-clinical departments/wards) were calculated by averaging the scores for individual reports in that group, adjusted by its number of workers. 'Term error score', 'report error score' and 'group error score' were defined in this manner. We also analysed the score distributions according to report volume size.

## Statistical analysis

We used the Wilcoxon signed-rank test to compare the 'report error score' with the manual GRM ratings. In addition, we calculated accuracy, recall, precision and F-score values from the report error score to evaluate the performance. For this analysis, we set a cut-off value on the basis of the receiver operating characteristic (ROC) curve using the training dataset. To determine the association of group error score level with manual ratings, we used Pearson's product–moment correlation test. A validation dataset not used for generating the training dataset was analysed using R software (V.4.3.0; R Project for Statistical Computing, Vienna, Austria).

# RESULTS

## Sample characteristics

Overall, 116 786 incident reports were collected during the study period. The incident reports by year and reporter occupation are shown in table 1. Reports in all years were made most often by nurses or midwives, accounting for 73.4% of all reports. After 2018, when 'other healthcare professionals' were subdivided by profession, physicians, pharmacists and rehabilitation therapists were the most likely (after nurses) to submit reports.

We included 114 013 reports in the study, with 2773 being excluded because the GRMs made no assessment about whether or not they were associated with errors. Of these reports, 71 038 (62.3%) were determined to contain errors; 57 006 reports were included in the training dataset and 57 007 in the validation dataset.

## Development of error scores

Using morphological analysis, error scores were calculated for 36 131 terms in the incident reports in the training dataset. The median term error score was −1.36 (IQR: −3.27 to 1.22). The highest term error score (45.25)

**Table 1** Incident reports by year and occupation of reporter

| | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Total submitted incident reports** | 9183 | 9752 | 10 082 | 11 443 | 11 109 | 10 333 | 10 086 | 10 676 | 10 939 | 11 883 | 11 300 | 116 786 |
| **By reporters** | | | | | | | | | | | | |
| Physician | 635 | 674 | 768 | 958 | 822 | 700 | 746 | 987 | 1218 | 1033 | 971 | 9512 |
| Nurse/midwife | 7603 | 7899 | 7734 | 8457 | 7975 | 7379 | 7074 | 7357 | 7491 | 8517 | 8268 | 85 755 |
| Other healthcare professionals | 914 | 1142 | 1542 | 1985 | 2263 | 2212 | | | | | | 10 058 |
| Pharmacist | | | | | | | 458 | 803 | 853 | 915 | 636 | 3665 |
| Radiological technologist | | | | | | | 221 | 172 | 170 | 222 | 266 | 1051 |
| Medical technologist | | | | | | | 310 | 259 | 156 | 145 | 188 | 1058 |
| Rehabilitation therapist | | | | | | | 314 | 387 | 395 | 413 | 381 | 1890 |
| Orthoptist | | | | | | | 51 | 45 | 55 | 56 | 33 | 240 |
| Biomedical equipment technician | | | | | | | 222 | 201 | 194 | 197 | 145 | 959 |
| Nursing assistant | | | | | | | 1 | 2 | 1 | 1 | 2 | 7 |
| Dental hygienist | | | | | | | 0 | 0 | 0 | 3 | 5 | 8 |
| Nutritionist | | | | | | | 219 | 226 | 178 | 143 | 144 | 910 |
| Administrative assistant | | | | | | | 447 | 186 | 191 | 200 | 238 | 1262 |
| Security guard | | | | | | | 0 | 0 | 0 | 2 | 2 | 4 |
| Others | 31 | 37 | 38 | 43 | 49 | 42 | 23 | 51 | 36 | 36 | 21 | 407 |

The category of ‘other healthcare professionals’ was subdivided after 2018.

was for ‘temoto joho’ (patient identifiers that healthcare workers can access to prevent misidentification), followed by ‘jikanme’ (hour(s) passed since an action was taken; 37.91), ‘kansasha’ (a person who inspects compounded medications; 30.58), ‘shokusatsu’ (a diet card with a patient identifier; 27.82) and ‘kanjagonin’ (patient misidentification; 23.85). The terms with high error scores are shown in online supplemental appendix 1.

The median report error score was 0.50 (IQR: −1.26 to 1.46). The median group error score, which is the total report error score of a group divided by the number of workers therein, was 0.06 (IQR: −0.46 to 0.61). Group error scores were high for the clinical nutrition (2.86), administration (2.73) and hospital pharmacy (2.20) departments, and low for the geriatrics ward (−2.38) and rehabilitation department (−2.09).

The SDs of these scores steadily decreased by level (3.45 for the term error score, 2.31 for the report error score and 0.85 for the group error score) (online supplemental appendix 2).

### Validation and performance of the report error score

The median report error score was −1.66 (IQR: −3.66 to −0.05) for error-free reports and 1.11 (IQR: 0.35–1.80) for reports of incidents that GRMs labelled as error containing; the difference was significant (p<0.001, *d*=0.73 (large)) (figure 2). Regarding the performance metrics, accuracy, which indicates the model’s ability to correctly predict the outcome (error containing or non-error containing) on the basis of all reports, was 0.8. Recall, that is, the probability of identifying error-containing reports among the GRM-classified error-containing reports, was 0.82. Precision, that is, the concordance between GRM-categorised error-containing reports and error-containing reports as determined by the model, was 0.85. Finally, the F-score, which reflects the balance between precision and recall, was 0.84. These results were obtained using an optimal cut-off score of ≥0.037 for error-containing incident reports derived from the ROC analysis (figure 3).

### Correlation of group error scores with manual classifications

A total of 119 organisational units were eligible to submit incident reports, including all departments and wards in NUH. Incident reports that GRMs rated as error containing were plotted against the group error scores for these 119 organisational units, and the correlation coefficient of 0.66 was highly significant (95% CI 0.54 to 0.75, p<0.001)(figure 4A). In the subgroup analysis focusing on the clinical departments in which incident reports were only submitted by physicians (n=58), the correlation coefficient was 0.71 (95% CI 0.55 to 0.82, p<0.001) (figure 4B).

## DISCUSSION

### Main findings and implications

The writing quality of individual reports widely varied because the reporters sometimes used the same term in different contexts or used different expressions to describe the same event. In addition, in medical dictionaries, ‘error’ appeared in terms such as ‘Human error’ and ‘Error message’, which may have affected the scores. However, the results became more reliable as the volume of reports increased, in line with the central limit theorem.

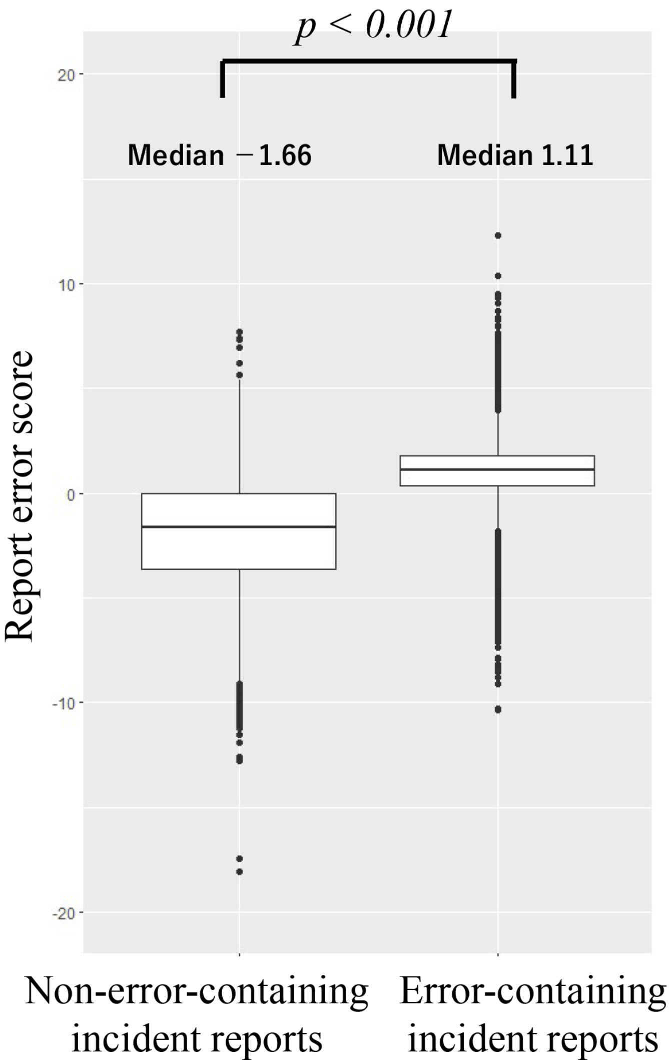
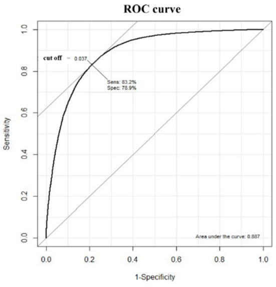


Figure 2 Box plot of report error scores among incident reports manually categorised by general risk managers (GRMs).

Notably, the report error score demonstrated that the model could more effectively identify reports of incidents arising from errors compared with manual categorisation by GRMs; the model’s performance metrics were good. These findings suggest that our model could be useful to analyse errors documented in individual reports, but we emphasise that it was designed to evaluate organisational trends in aggregated reports.

More importantly, higher error scores for departments were associated with a higher submission rate of error-containing incident reports. This phenomenon was also observed for group severity scores which indicate the severity of incidents using this model.[14] Severe events rarely occur, but events associated with errors are



Sens, sensitivity; Spec, specificity

| | | Report categorization based on the report error score, n (cut-off ≧ 0.037) | |
|---|---|---|---|
| | | Error-containing | Non-error-containing |
| **Report Categorization by GRMs, n** | Error-containing | **29,257<br>(a)** | **6,408<br>(b)** |
| | Non-error-containing | **4,993<br>(c)** | **16,273<br>(d)** |

GRMs, general risk managers

| **Performance metrics** | |
|---|---|
| Accuracy (a + d/a + b + c + d) | 0.80 |
| Recall (a/a + b) | 0.82 |
| Precision (a/a + c) | 0.85 |
| F-score ($\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$) | 0.84 |

**Figure 3** Performance metrics of report error score. ROC, receiver operating characteristic.

relatively common. The results suggest that our model is able to analyse factors involved in incidents regardless of their frequency of occurrence.

Departments with higher error scores, such as the clinical nutrition, administration and hospital pharmacy departments, tended to submit more reports. However, their reports included many near-miss and less severe events. The error score simply indicates the existence of error in association with an incident, not the severity of the error or its consequences. Each department provides their own services, and scores therefore cannot be compared directly among departments. The scores are also influenced by whether departments are correctly submitting reports of all incidents, including those arising from errors and other reasons. Although we are aware that the outcomes would have been more accurate had outliers been removed, the results are nevertheless considered robust given the sufficient data volume.

### Comparison with previous related work

When artificial intelligence-enabled decision support systems are implemented correctly, they can improve patient safety.[13] Researchers have explored the potential of applying NLP techniques to incident reports, often in conjunction with machine learning.[15] Most studies used a binary classification, but research aiming to identify multiclass classifications is emerging gradually.[25] These studies were designed to answer questions about individual incident reports. However, the writing quality (ie, complexity and length) of incident reports varies greatly.[26] Our model is unique in that we aimed to analyse groups of reports to understand organisational patterns and trends. We performed statistical analysis to compare the results between groups, but we could not find adequate classifiers to evaluate groups in the context of machine learning and NLP. We therefore adopted rank-based tests, which are sometimes used in NLP.[27 28] The drawback of rank-based tests is their relatively weak statistical power, but our sample size was large enough to overcome this limitation.

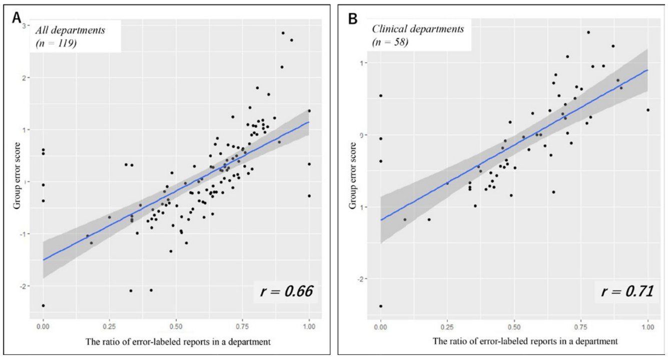


**Figure 4** Scatter plot of the ratio of general risk manager (GRM)-labelled error-containing incident reports against averaged group error scores in each organisational unit at Nagoya University Hospital (NUH). In A, all departments were plotted. Only the clinical departments were included in B.

Various vectorisation methods, such as binary, term frequency, thresholding and term frequency-inverse document frequency methods, are generally used to transform segmented terms into numerical representations.[29] We adopted the same vectorisation method to weigh semantic characteristics as was applied to the severity score, which is used to quantify event severity on the basis of training data and GRM classifications.[14] The severity score can also be used to predict organisational trends. A study on severity scores highlighted that many terms used in reports of severe incidents did not appear in reports of non-severe incidents. However, that study had a huge number of non-severe incident reports and far fewer severe reports.[14] To alleviate this imbalance, the formula was updated in this study by adding one. This method reduced the number of words with a zero probability and has been used in other vectorisations, such as term frequency-inverse document frequency[30] and Bayesian vectorisation.[31] However, direct comparison with other vectorisation models was outside the scope of this research.

### Limitations

This study had several limitations. First, it used data from a single facility in Japan. All incident reports were written in Japanese, and the results may vary by language. Moreover, we applied a consensus method to triage reports using our institutional definition of 'error'. Unfortunately, the inter-rater reliability of the GRMs in terms of error scores was not confirmed, although we consider the quality of our safety department to be high. In addition, the judgements of multiple trained GRMs were considered, including legal experts. The number of incident reports may vary among hospitals depending on the reporting culture. As incident reports share similarities, we believe that this model is widely applicable, although additional research is required to confirm its applicability to other languages or institutions.

Second, we did not perform any qualitative analysis of the segmented terms generated by the morphological analysis, and the narrative descriptions in the reports were not included in the analysis. Although these factors would have influenced the quality of the scores, we nevertheless consider the study useful because it included a large sample of real-world data, including incomplete reports and ones with inaccurate event descriptions. However, some measures, such as maintenance of dictionaries for morphological analysis and preprocessing of raw free-text data to correct typing errors, could improve the results.

### Challenges for future work

In future, our scoring model could be used to monitor chronological trends in errors at the group level, as well as to increase the awareness of workers and GRMs. It might therefore provide data that could help prevent future incidents. We also expect this system to be useful for educating new GRMs.

We will continue to try to improve the performance of the model. We modified the vectorisation formula to increase calculable terms in free-text data; other possible measures include data preprocessing, updating dictionaries and identifying the optimal number of incident reports to assess group error scores.

In addition to severity and error, other factors are involved in incidents; we will aim to quantify these factors using the same methodology applied herein. In future, a useful tool could be developed to enhance organisational patient safety by combining multiple scores, including severity and error scores, in a balanced manner. This study represents a useful step towards that goal.

## CONCLUSIONS

We developed a decision-making model to quantify errors by analysing the semantic characteristics of free-text data in incident reports. Analysing scores by organisational unit revealed strong correlations with expert ratings. By expanding the scope of this model, an incident reporting system promoting patient safety could be obtained.

**Acknowledgements** We thank the staff of the Department of Patient Safety at Nagoya University Hospital and members of ASUISHI and CQSO. We also thank Atsushi Okawa, Nobuyuki Toyama, Yasuyuki Nasuhara, Toshihiro Kaneko, Masashi Uramatsu, Yumi Arai, Kouichi Tanabe and Tatsuya Fukami for their help in this research project. We thank Melissa Leffler, MBA, and Michael Irvine, PhD, from Edanz (https://jp.edanz.com/ac) for editing a draft of this manuscript.

**Contributors** YN and MU conceived the idea for the study. HU led the writing of this paper. MU, MK, HY, TU, MH, FK and YN analysed and interpreted the data. MU, MK, HY and YN contributed to the writing of the paper as well as participated in revising this manuscript. All authors contributed substantially to the writing of the paper, and all reviewed and approved the final draft. HU and YN are the guarantors of this study.

**Funding** This research was supported by the Ministry of Health, Labour and Welfare Policy Research Grants, Research on Region Medical (20IA2001).

**Competing interests** None declared.

**Patient consent for publication** Not applicable.

**Ethics approval** This study was approved by the Nagoya University Hospital Research Ethics Committee (2020-0181).

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Data are available upon reasonable request. The data that support the findings of this study are available from the corresponding author upon reasonable request.

**Supplemental material** This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

**ORCID iDs**
Haruhiro Uematsu http://orcid.org/0000-0003-2800-6802
Masaru Kurihara http://orcid.org/0000-0001-9195-4202

## REFERENCES

1 Dzau VJ, Shine KI. Two decades since to err is human: progress, but still a "chasm". *JAMA* 2020;324:2489–90.
2 Makary MA, Daniel M. Medical error-the third leading cause of death in the US. *BMJ* 2016;353:i2139.
3 Vincent C, Burnett S, Carthey J. Safety measurement and monitoring in Healthcare: a framework to guide clinical teams and healthcare organisations in maintaining safety. *BMJ Qual Saf* 2014;23:670–7.
4 Pham JC, Girard T, Pronovost PJ. What to do with healthcare incident reporting systems. *J Public Health Res* 2013;2:e27.
5 Evans SM, Smith BJ, Esterman A, *et al*. Evaluation of an intervention aimed at improving voluntary incident reporting in hospitals. *Qual Saf Health Care* 2007;16:169–75.
6 World Health Organization. *Patient safety incident reporting and learning systems: technical report and guidance*. Geneva: World Health Organization, 2020.
7 Stavropoulou C, Doherty C, Tosey P. How effective are incident-reporting systems for improving patient safety? A systematic literature review. *Milbank Q* 2015;93:826–66.
8 Kohn KT, Corrigan JM, Donaldson MS. *To err is human: building a safer health system*. Washington, DC, 1999.
9 Scott J, Dawson P, Heavey E, *et al*. Content analysis of patient safety incident reports for older adult patient transfers, handovers, and discharges: do they serve organizations, staff, or patients *J Patient Saf* 2021;17:e1744–58.
10 Howell A-M, Burns EM, Bouras G, *et al*. Can patient safety incident reports be used to compare hospital safety? Results from a quantitative analysis of the English national reporting and learning system data. *PLoS One* 2015;10:e0144107.
11 Mitchell I, Schuster A, Smith K, *et al*. Patient safety incident reporting: a qualitative study of thoughts and perceptions of experts 15 years after 'to err is human'. *BMJ Qual Saf* 2016;25:92–9.
12 Bates DW, Levine D, Syrowatka A, *et al*. The potential of artificial intelligence to improve patient safety: a scoping review. *NPJ Digit Med* 2021;4:54.
13 Choudhury A, Asan O. Role of artificial intelligence in patient safety outcomes: systematic literature review. *JMIR Med Inform* 2020;8:e18599.
14 Uematsu H, Uemura M, Kurihara M, *et al*. Development of a novel scoring system to quantify the severity of incident reports: an exploratory research study. *J Med Syst* 2022;46:106.
15 Young IJB, Luz S, Lone N. A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. *Int J Med Inform* 2019;132:103971.
16 Hoshi T, Nagao Y, Sawai N, *et al*. Assessment of medical malpractice cost at a Japanese national University hospital. *Nagoya J Med Sci* 2021;83:397–405.
17 Rodziewicz TL, Houseman B, Hipskind JE. Medical error reduction and prevention. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing Copyright © 2023, StatPearls Publishing LLC, 2023.
18 Higham H, Vincent C. Human error and patient safety. In: Donaldson L, Ricciardi W, Sheridan S, et al, eds. *Textbook of patient safety and clinical risk management*. Cham (CH): Springer Copyright, 2021: 29–44.
19 Thomas EJ, Petersen LA. Measuring errors and adverse events in health care. *J Gen Intern Med* 2003;18:61–7.
20 Benn J, Koutantji M, Wallace L, *et al*. Feedback from incident reporting: information and action to improve patient safety. *Qual Saf Health Care* 2009;18:11–21.
21 Fukami T, Uemura M, Nagao Y. Significance of incident reports by medical doctors for organizational transparency and driving forces for patient safety. *Patient Saf Surg* 2020;14:13.

22 Fondahn E, Lane M, Vannucci A. *The Washington manual of patient safety and quality improvement*. Philadelphia, Pennsylvania: Wolters Kluwer, 2016.

23 World Health Organization. Patient safety curriculum guide: multi-professional edition. 2011. Available: https://apps.who.int/iris/bitstream/handle/10665/44641/9789241501958_eng.pdf;jsessionid=E3E8BA7049BED778EACF49D665F9FCD4?sequence=1

24 Reason J. *Human error*. Cambridge: Cambridge University Press, 1990.

25 Wang Y, Coiera E, Magrabi F. Using convolutional neural networks to identify patient safety incident reports by type and severity. *J Am Med Inform Assoc* 2019;26:1600–8.

26 Fong A, Hettinger AZ, Ratwani RM. Exploring methods for identifying related patient safety events using structured and unstructured data. *J Biomed Inform* 2015;58:89–95.

27 Rousseau JF, Ip IK, Raja AS, *et al*. Can automated retrieval of data from emergency department physician notes enhance the imaging order entry process? *Appl Clin Inform* 2019;10:189–98.

28 Donnelly LF, Grzeszczuk R, Guimaraes CV, *et al*. Using a natural language processing and machine learning algorithm program to analyze inter-Radiologist report style variation and compare variation between radiologists when using highly structured versus more free text reporting. *Curr Probl Diagn Radiol* 2019;48:524–30.

29 Ong MS, Magrabi F, Coiera E. Automated identification of extreme-risk events in clinical incident reports. *J Am Med Inform Assoc* 2012;19:e110–8.

30 van Zaanen M, Kanters P. *Automatic mood classification using TF*IDF based on lyrics*. ISMIR, 2010.

31 Sueno HT, Gerardo BD, Medina RP. Converting text to numerical representation using modified Bayesian vectorization technique for multi-class classification. *International Journal* 2020;9:10.30534.